Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-19-27

The Convenience of Large Urban School Districts: A Study of Recruitment Practices in 37 Randomized Trials

Elizabeth Tipton
Associate Professor of Statistics and IPR Fellow
Northwestern University

Jessaca Spybrook
Professor of Evaluation, Measurement and Research
Western Michigan University

Katie Fitzgerald
Department of Statistics
Northwestern University

Qian Wang
Department of Evaluation, Measurement and Research
Western Michigan University

Caryn Davidson
Department of Evaluation, Measurement and Research
Western Michigan University

Version: November 19, 2019

DRAFT

Please do not quote or distribute without permission.

ABSTRACT

As a result of the evidence-based decision-making movement, the number of randomized trials evaluating educational programs and curricula has increased dramatically over the past twenty years. Policy makers and practitioners are encouraged to use the results of these trials to inform their decision making in schools and school districts. At the same time, however, little is known about the schools taking part in these randomized trials, both regarding how and why they were recruited and how they compare to populations in need of research. In this paper, the researchers report on a study of 37 cluster randomized trials funded by the Institute of Education Sciences between 2011 - 2015. Principal Investigators of these grants were interviewed regarding the recruitment process and practices. Additionally, data on the schools included in 34 of these studies was analyzed to determine the general demographics of schools included in funded research, as well as how these samples compare to important policy relevant populations. The authors show that the types of schools included in research differ in a variety of ways from these populations. Large schools from large school districts in urban areas were over-represented, while schools from small school districts in rural areas and towns are under-represented. The paper concludes with a discussion of how recruitment practices might be improved in order to meet the goals of the evidence-based decision-making movement.

The Convenience of Large Urban School Districts: A Study of Recruitment Practices in 37 Randomized Trials

The evidence-based (EB) decision-making movement in education encourages practitioners and policy-makers to consider findings from rigorous evaluations when adopting both core curricula and supplementary programs. For example, the *Every Student Succeeds Act* (ESSA) requires school officials using federal funds to purchase curricula that has evidence that the program has been evaluated previously, and ideally, using a strong causal design (i.e., “Tier 1 evidence”). To help facilitate such decision-making, reviews of research are made available via clearinghouses and websites, including the *What Works Clearinghouse* (WWC), the *Educational Endowment Foundation* (EEF), *Evidence for ESSA*, and *Blueprints for Healthy Youth*.

Since its debut in 2002, the *Institute of Education Sciences* (IES) has provided the backbone of this EB decision-making movement in education. In addition to establishing the WWC, much of IES has been devoted to increasing the number and quality of evaluations of education curricula and programs over a wide range of topics and student ages. To date, IES has funded over 300 evaluations of Pre-K - 16 curricular and other programs (Chhin, Taylor, & Wei, 2018; J Spybrook, Shi, & Kelcey, 2016). Furthermore, IES expects both grant- and contract-funded evaluations to follow quality standards developed by the WWC, with priority given to randomized trials. These trials typically include between 40-60 schools each, with roughly half receiving a new program and the other half continuing with business as usual (J Spybrook, Shi, et al., 2016).

As the EB movement has grown, however, it has become clear that decision makers - e.g., school district superintendents, principals, curriculum specialists - seek programs and curricula based not only on if there is evidence that the program works, but also if it works in *populations like theirs* (e.g., Department of Education, 2016). To this end, the WWC now provides a query option including questions regarding characteristics of schools in studies such as urbanicity, student grades, and demographics. Unfortunately, the information available for these queries is limited in scope, as primary research articles rarely report clearly or consistently features of the sample of schools or students included in their studies (Fellers, 2017).

Given this dearth of information, how should practitioners and policymakers seeking population information make these curricular decisions? This is the exactly the question that statisticians developing methods for generalizing the results of randomized trials have begun asking more broadly, from education to social welfare to medicine (for an overview, see Tipton & Olsen, 2018). At its core, this research highlights that population-specific questions only matter if treatment effects vary: if the effects of interventions are constant, then where or with whom a trial takes place does not matter (Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013). However, to date, there have been a variety of examples that suggest that such an *a priori* assumption of constant treatment effects is likely unwarranted (e.g., Schanzenbach, 2006; Weiss et al., 2017; Yeager et al., 2019). Altogether, this highlights that the "effect" reported in a randomized trial is an *average*, and that this average might differ from one population or sample to another.

Read another way, this research implies that the results from a single trial can be used to inform multiple, different population average treatment effects. To do so, in addition to information on the sample of schools and students found in a trial, information on the target population for decision making is required; this information can often be gleaned from national databases of schools and school districts (e.g., Common Core of Data [CCD]). Importantly, a target population can be defined broadly (e.g., all schools in the US) or narrowly (e.g., all low-income, urban schools serving second graders in Maine). For a given study, these methods can be used to provide information on the populations in which generalizations are warranted versus where they require strong extrapolations (e.g., Tipton, 2014), as well as methods for estimating average treatment effects for these populations (Hartman, Grieve, Ramsahai, & Sekhon, 2015; Stuart et al., 2011; Tipton, 2013).

While these new methods for estimating population average treatment effects are promising, theoretical studies show that their effectiveness is limited when the sample and population differ strongly. When they are very different – e.g., when the average proportion of students in poverty is small in the sample but large in the population – the precision of the reweighted estimate can be dramatically worse than in the original study (Tipton, 2013). In the extreme, this results in an under-coverage problem – wherein a portion of the population is not represented at all in the sample, making it impossible to estimate the population average treatment effect without extrapolations. This problem is not simply theoretical. Tipton and colleagues (2016) compared the schools taking part in two IES funded Goal 4 scale-up studies with different target populations and found that the samples in these trials represented only a small subset (< 1/3) of the populations that used these programs. At a broader scale, Stuart and colleagues (2017) conducted a review of 19 randomized trials and regression discontinuity designs funded by the *National Center for Education Evaluation and Regional Assistance* (NCEE) contracts within IES before 2011. For 11 of these studies, PIs were able to share data regarding the school districts – but not schools – that took part. They found that compared to school districts in national target populations (where programs might be implemented), the school districts taking part in these studies were larger (e.g., more schools and students), had higher proportions of students on free- or reduced- priced lunch, larger percentages of non-white students, and were in more urban and less rural areas.

To date, these two studies provide the only evidence regarding the broader match between the types of schools found in education randomized trials and the populations in need of such research for decision-making. Questions about this match (or mismatch) are important at multiple levels. Funders may wonder if their portfolio of research collectively represents well the population of schools in need, whether those serving low-income students or those in low-achieving school districts. At the other extreme, those interpreting the findings of individual studies may want to know the extent to which the results from such a study represent well their specific population. At either level, if differences between these samples and populations arise, information on where and why the differences exist is important, including the constraints and processes found in recruitment.

In this paper, we seek to shed light on these questions regarding the generalizability of results from education randomized trials. To do so, we report on the findings from a study of recruitment practices in 37 cluster-randomized Goal 3 (Efficacy) and Goal 4 (Effectiveness) studies funded by IES between 2011 – 2015[1]. We ask three questions in this paper:

1. **What is the recruitment process in randomized trials in education?** This question includes understanding the strategies, constraints, and processes that researchers use and face in recruitment and the locations of schools that are ultimately recruited into studies.
2. **Overall, how similar are the schools taking part in randomized trials to different policy-relevant target populations?** We expect this question to be particularly relevant

[1] Beginning in 2019, RFAs no longer refer to "goals", but instead identify studies as "exploration", "development", "efficacy" or "replication". The latter two are similar to the Goal 3 and Goal 4 studies defined here.

to IES and to funders more broadly, who may wish to understand how well their entire portfolio of studies represents schools in need of research.

3. **For each individual study, how similar is the sample of schools to these various target populations of schools?** We expect this question to be particularly relevant to individual researchers and to decision-makers interpreting evidence from individual trials.

Overall, these questions are descriptive in nature, and answering them requires both quantitative and qualitative data and analyses. In the next section, we introduce the data collected and methods used. We then provide results for each of these three questions, followed by a discussion of the findings, suggestions for improving practice, and a short conclusion.

Data and Methods

Population Data

Practitioners and policy-makers often wish to understand the extent to which research findings apply to schools and contexts like their own. Exactly how broadly or narrowly such a population is defined depends on the degree to which the effect of an intervention is expected to vary across students and contexts and the level at which decision-making occurs. Given the goals of IES and education research more generally, in this paper we focus on four broadly defined target populations of schools: all public schools, high poverty schools (> 40% free-or-reduced lunch students), very high poverty schools (> 80% free-or-reduced lunch students), and schools in low-achieving districts (bottom 25%ile). For each, we provide separate analyses for elementary schools (K - 5) and middle/high schools (6 - 12); note that these populations overlap in some cases.

In Table 1, we provide the total count of schools in each of these target populations. We defined these populations using the 2015 - 2016 Common Core of Data (U.S. Department of Education, 2016). In order to define 'all schools', we focus only on 'regular' public schools found in the continental U.S.; further information on these inclusion criteria are found in Supplemental Figure 1. This population is further restricted to high poverty and very high poverty schools based upon the percent of students on free-or-reduced lunch (FRL) found in the CCD. Note that high poverty schools represent about 65% of public schools in the U.S., while very high poverty schools represent about 21% of schools. In order to identify schools in low achieving school districts, we merged the achievement data at the geographic district level from the Stanford Education Data Archive (SEDA) with the CCD. The metric is pooled across years, grades, and subjects, and is on a Cohort Scale. We defined low-achieving districts to be those that had mean achievement scores at or below the 25th percentile on this metric.

Table 1. Population information

	Elementary Schools	**Middle/ High Schools**	**Total Schools**
All public, regular schools	54,898	50,055	84,252
High poverty (> 40% FRL)	37,021 (67% ES)	32,324 (65% MHS)	55,133 (65% TS)
Very high poverty (> 80% FRL)	13,134 (24% ES)	9,695 (19% MHS)	17,722 (21% TS)
Low-achieving districts (bottom 25%ile of achievement)	15,246 (28% ES)	14,136 (28% MHS)	22,487 (27% TS)

Note: Elementary is defined as any school serving K - 5 students and Middle/High is any school serving 6 - 12 students. In some cases, these overlap, for example in K - 12 schools. All schools and high poverty schools are based on data in the CCD. Low-achieving districts are based on data from SEDA.

Sample Data

We focus on a sample of IES Goal 3 (Efficacy) and Goal 4 (Effectiveness) studies funded between 2011 – 2015. This range of dates was selected in order to ensure that (a) recruitment for the studies was complete at the time of interviews and (b) that once contacted, researchers would be likely to have records regarding recruitment. Studies were selected if they recruited K – 12 schools or districts (and excluded if they were partnership studies) and randomized schools, teachers, or classrooms to an intervention. These inclusion criteria resulted in 40 intervention studies distributed across 36 unique Principal Investigators (PIs).

Beginning in the fall of 2017, we contacted each PI and requested an interview regarding recruitment in their studies[2]. These interviews took place over an 18-month period, with some taking place in person (e.g., at conferences) and others on the phone. Interviews were conducted by the two lead paper authors and typically lasted 30-40 minutes. Overall, we conducted interviews regarding recruitment in 37 of the studies (3 study PIs did not respond to repeated requests), which resulted in 33 total interviews. Twenty-four of the interviews were recorded and later transcribed and in nine interviews notes were taken instead.

In addition to participating in the interview, we requested that PIs share the names of the schools in their sample. Out of the 37 studies we obtained interview data from, we were able to obtain school data from 34 studies (92%)[3]. In each study, schools were located within the CCD in the year prior to recruitment[4]; when data was missing, data from the year of recruitment was used (e.g., if the school was new). In total, the final sample includes 34 studies, 449 school

[2] In 3 studies, instead of the PI, we interviewed the person in the study team that oversaw recruitment.
[3] In 3 studies, the PIs were unable to provide any data regarding schools; this was either because of the study IRB, or because the PI no longer had access to records.
[4] In 10 (29%) of these 34 studies, the PIs were not able to directly provide us with a list of schools. Instead, they pulled the data from the CCD for us and provided us with a list of demographics which we then used in our analyses.

districts, and 1,479 schools, which in total served 971,263 students. The specific details for the data collection process can be found in Supplemental Figure 2.

Finally, we also sought to determine the intended target population for each study. To do so, we looked for criteria listed in the grant abstract, published papers, or mentioned in the interviews. We additionally coded other more “local” target populations, based upon the state and school districts where each study took place.

Methods

For Question 1, regarding the recruitment process and the types of schools in the studies, we began our analyses with data collected in the interviews. A mixed-methods approach was taken to analyze the interview data. During the first read of interviews, passages relevant to answering research questions were marked and emerging categories were noted. A spreadsheet was then developed and key variables, such as ease of recruitment, type of intervention, level of intervention (district, school, teacher, student), level of recruitment (district, school, teacher) were noted for each study. Relevant text regarding the recruitment strategy was transcribed into the spreadsheet and descriptive coding was then used to categorize chunks of text in a separate document. Descriptive coding “summarizes in a word or short phrase – most often a noun – the basic topic of a passage of data” (Saldana, 2016). Study ID numbers were retained along with text so the other variables could be considered. Once text was organized by code, the text in each code grouping was more deeply analyzed, codes were refined, and each grouping was summarized. Code weaving was used to group related codes and look for more meaningful patterns across the data.

Finally, based on findings from the interviews, we conducted an additional analysis regarding the location of schools in the studies. Data on the location of schools was found in the CCD, and we mapped these school locations in R using ggplot2 (Wickham, 2016). Additionally, we coded the number of states in which a study was conducted, the location of the study PI and co-PIs, and if the PI was a research firm or university. We compared differences in these trends across institution type.

In Question 2, we compared the schools across the 34 study samples to the four previously defined target populations (all schools, high poverty, very high poverty, low-achieving districts). For each comparison, we focused on the following features:

School features:

- District size (i.e., number of schools) and school size (i.e., number of students)
- Urbanicity: Urban, Suburban, Town, Rural
- Student-Teacher Ratio

Student demographics:

- Race/Ethnicity (% white, % Black, % Hispanic),
- Gender (% female),
- % Free-or-Reduced Lunch (FRL)
- % English Language Learners (ELL)

Across the schools in these 34 studies, we calculated means and standard deviations of these variables in the sample and population. Based upon these results, we conducted additional analyses regarding district size across samples and the target populations.

In Question 3, we conducted separate analyses for each of the 34 studies. We began by comparing each study sample to each of the four target populations. We then compared each study sample to two "local" target populations, defined as the relevant state(s) and school district(s). For each of these target populations, we individually compared each of the 34 study samples[5] and summarized the degree of overall similarity using the generalizability index (Tipton, 2014). This index ranges from 0 to 1, with higher values indicating greater similarity; for example, a value of 0.80 indicates that the sample and population are 80% similar. Values of the index greater than about 0.90 indicate that the sample and population are as similar on these variables as a random sample of the same size, while values less than about 0.50 indicate that the sample and population are so different that generalizations are unwarranted (Tipton, 2014; Tipton, Hallberg, Hedges, & Chan, 2017). Values in between these extremes indicate that reweighting would be needed in order to estimate the population average treatment effect; this reweighting results in reduced precision. In order to maintain anonymity, we report these index values in aggregate for each of the 6 possible target populations across the 34 studies using boxplots and summary statistics.

Results

Question 1: Recruitment

Recruitment process

The qualitative analyses of interviews showed that the ease of recruitment was heavily dependent on relationships between researchers and schools. In the majority of the studies (n = 18), researchers relied on some type of connection to the districts whether direct or indirect. In 9 of these studies, the researchers had longer-standing relationships with districts either through providing prior professional development or conducting previous studies. These relationships

[5] Each sample was compared to each population using a propensity score model including the number of students, urbanicity, race/ethnicity, % of students with free- or reduced-priced lunch, and number of schools in the district. We were not able to include the student-teacher ratio, % female, or % English language learners (ELL) because of missing data and/or model convergence problems.

ensured that there was already trust on both sides which made it easier for them to navigate those districts and obtain approval for a study. While prior relationships helped to some degree, they were not always available. In these cases, researchers sometimes had to cold call schools, making it very difficult to recruit. In these situations, researchers reported spending a lot of time designing recruitment materials, advertising materials, attending conferences, purchasing mailing lists, and making cold calls, often for little return (n = 5). Overall, the majority of researchers preferred to utilize a top down recruitment strategy where they sought to garner support from district leadership as a first step.

Researchers also reported a variety of constraints that affected recruitment, including geography, district size, and the intervention itself. Geography, in particular, played an important role (n = 14). Most researchers chose districts that were close to the PI or co-PI locations because it made it easier and less expensive for them to visit the schools. Additionally, while no clear preference was reported, district size played a role in recruitment (n = 10). For example, many researchers talked about the challenges associated with 'red tape' that comes along with extremely large districts. Having a strict set of criteria for the intervention also impacted recruitment (n = 8). For example, interventions which were intended for a very specific population, such as ELL students in middle school, made recruiting challenging because the criteria was so narrow. As a result of these constraints, researchers were sometimes limited in their ability to leverage prior working relationships and/or connections.

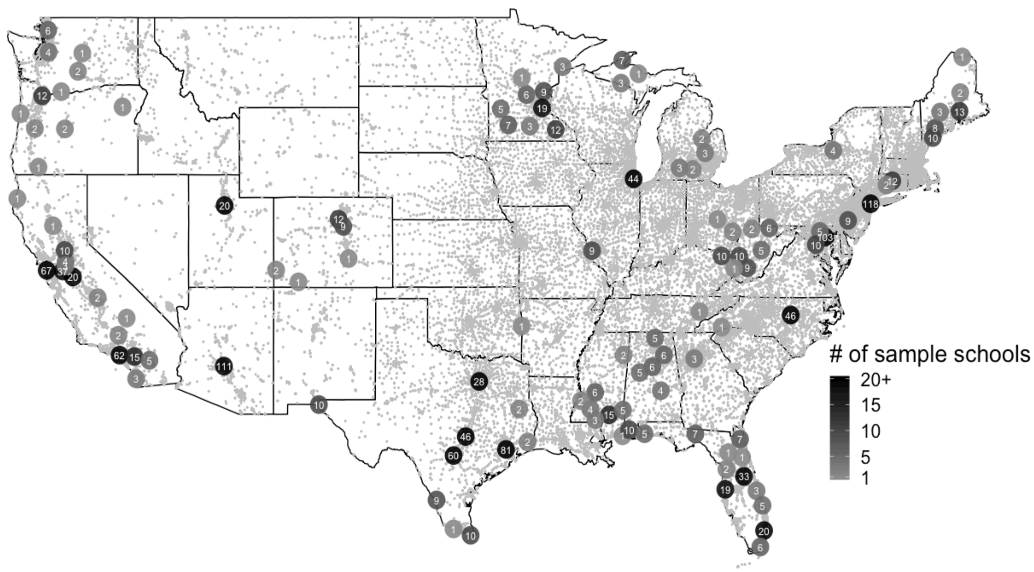
Location analysis

The approximate location of the schools across the 34 studies is presented in Figure 2; separate figures for elementary and middle/high schools are provided in Supplementary Figures

3 and 4. Note that we do not include Hawaii or Alaska as there are no study schools in those states, and that we have combined geographic areas here so as to not identify the exact location or identity of any of the specific schools or PIs associated them. In addition to indicating via circles the school locations in studies, the map also indicates in light gray the locations of public schools throughout the US. Overall, the map suggests that states along the coasts are over-represented in studies, with ⅓ of schools in studies found in Texas and California. In comparison, surprisingly few studies took place in the Midwest relative to its number of schools.

Additionally, we compared location trends for studies conducted by PIs at research firms (42%) to those conducted by PIs at universities (58%). For each study, we coded the number of states that were included in a study and if the study was conducted nearby to a firm office or the PI or co-PI's university. In general, the number of states that a study was conducted in did not differ across firms and universities, with about 75% of studies in both conducted in a single state. In the other 25% of studies, the total number of states included ranged from 2 – 8. In addition, there were differences by PI affiliation regarding the proximity of the schools included in the study to the PIs and co-PIs. On average, PI teams at universities were more likely to recruit in their state (74%) compared to those at research firms (57%). Furthermore, of those studies conducted by PIs at universities, over twice as many were conducted entirely in their state (53%) as were conducted entirely out of state (26%); in comparison, at research firms, roughly the same percentage were conducted both entirely within and out of state (36% within, 43% outside).

Figure 2. Locations of study and population schools



Question 2: Representation of target populations broadly

Descriptive comparisons

Comparisons across the 34 study samples and each of the four target populations, divided by grade (elementary, middle/high) can be found in Table 3. Note that at the bottom of the table, the sample sizes that meet the inclusion criteria for each target population are indicated, as well as the proportion of the total population, and the generalizability index. In what follows, we discuss each target population in order.

All schools. As the generalizability index indicates, for both elementary schools and middle/high schools the sample of schools included in studies differ highly from those in the target population (index = .59, .57 respectively). In particular, schools included in studies came from larger school districts and had larger numbers of students per school, were more likely to be urban and less likely to be in towns or rural areas, had higher percentages of students in poverty, and included more minority students than those in the population.

High poverty schools. 79% of study elementary schools and 72% of study middle/high schools were in high poverty schools (> 40% FRL), compared to 67% and 65% of schools in the population, respectively. However, the high-poverty study schools were less similar to these high-poverty population schools than in the overall comparison (index = .45, .41 compared to .59, .57). Like the comparison of all schools, high-poverty study schools were in larger school districts and in schools with larger numbers of students than in the population. These study schools were also more often in urban and suburban areas and less often in town and rural areas, and included larger percentages of minority students. Even within this definition of high-poverty, study schools included larger percentages of high-poverty students than in the populations.

Very high poverty schools. 42% of study elementary schools and 29% of study middle/high schools were in very high poverty schools (> 80% FRL), compared to 24% and 19% of schools in the population, respectively. These very high-poverty study schools were the least similar to their respective schools in the population overall (index = .21, .12 compared to .59, .57). While the percentages of students on FRL were very similar in the study sample and populations on average, again study schools were found in larger school districts and schools with larger numbers of students than those in the population. Similarly, larger percentages of study schools came from urban areas and smaller percentages from town and rural areas than in the populations, and study schools served larger percentages of minority students than in the populations.

Schools in low-achieving districts. 44% of study elementary and 36% of study middle/high schools were found in low-achieving school districts, compared to 28% of schools in the population for each. Overall, study schools were more similar to the population schools than those found in the high- and very high-poverty analyses (index = .58, .51 compared to

.45,.41 [high poverty] and .21,.12 [very high poverty]). As with the other analyses, elementary study schools were found in larger school districts and serve larger numbers of students compared to the population; for middle/high schools, study schools served larger numbers of students, but were not found in larger school districts. Urban schools were again over-represented relative to the population, and for elementary schools, town and rural schools were under-represented.

The role of district size

In the analyses presented in Table 3, it is clear that the number of schools per district ("district size") is considerably larger across the study schools than in every target population. In Figure 2, we investigate this further, presenting the distribution of the logged district size in study schools and the total target population ("all schools"). Notice that on the x-axis, different cut-points for non-logged values are given, ranging from 1 school in a district to over 1000 schools. As depicted in the figure, there is considerable mismatch between the district size in the population and studies. First, fully 24% of school districts in the U.S. consist of a *single* school and another 49% consists of school districts with between 2 and 5 schools; in our sample of schools in studies, these are represented in only 26% in total. Second, school districts composed of more than 125 schools (e.g., L.A. Unified) account for only 0.25% of districts in the U.S., but 6% in our sample.

Given these differences in the distribution of district size between the population and study samples, we conducted additional analyses of the samples and populations in relation to district size. In Table 4, we present these findings. In this table, the population of school districts in the U.S. is divided into five categories (columns) based upon the number of schools in the total population. The first four columns break down the largest 10% of districts in terms of size, and the last column includes the smallest 90% of districts, which include 11 or fewer schools. The rows correspond to the districts, schools, and students in the sample and different populations. As the table indicates, these 90% smallest districts account for nearly half (49%) of schools in the country and nearly 40% of students. However, these smallest 90% of districts are represented in only 18% of schools and 12% of students in studies. In contrast, while the 0.25% of largest school districts (> 125 schools) account for 11% of schools in the population, they account for 30% of study schools. Importantly, trends for high poverty and very high poverty schools, and low achieving school districts are similar to these overall trends.

While not indicated in this table, additional analyses indicate that compared to the largest 10% of school districts, these 90% smallest districts are more likely to be in rural areas (48% versus 18%) and in towns (17% versus 10%) and are comprised of smaller schools (402.9 versus 662.0 students), while the average percent of students on free or reduced lunch are nearly equivalent (47% versus 52%).

Figure 2. **Distribution of log-district size in studies versus total population**

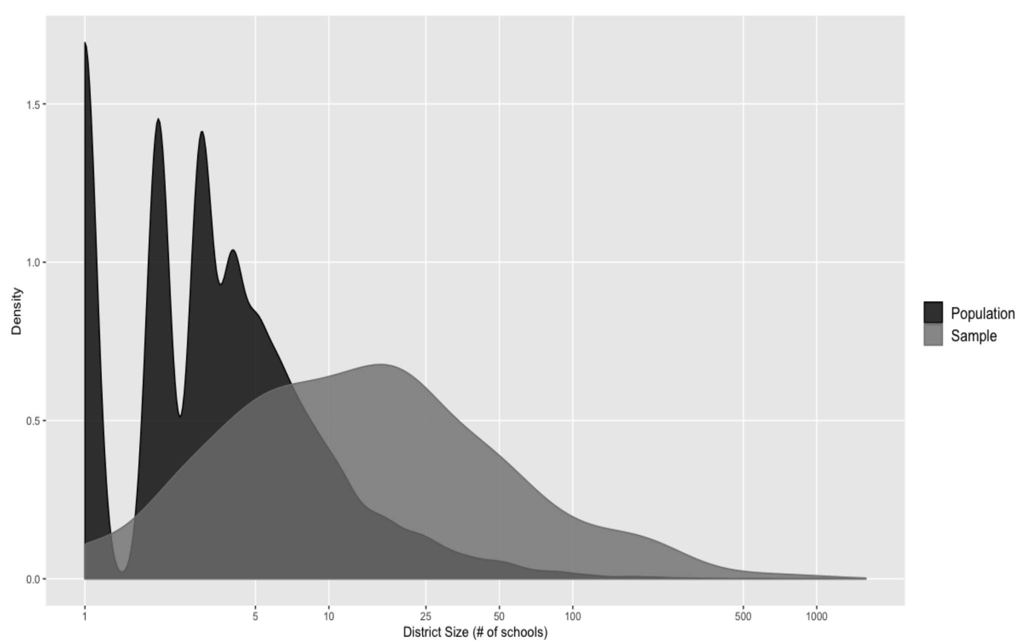


Table 3. Comparison of study schools to each target population by grade-level

Elementary Schools								
	All Schools		**High Poverty**		**Very High Poverty**		**Low Achieving Districts**	
	Sample	Population	Sample	Population	Sample	Population	Sample	Population
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Students per school	571.05 (234.17)	481.97 (237.87)	557.95 (233.91)	476.49 (234.59)	570.81 (225.73)	503.32 (234.74)	577.98 (228.25)	488.5 (230.25)
% Urban	50% (0.50)	31% (0.46)	54% (0.50)	35% (0.48)	70% (0.46)	53% (0.50)	59% (0.49)	45% (0.50)
% Suburban	31% (0.46)	34% (0.47)	26% (0.44)	26% (0.44)	22% (0.41)	24% (0.43)	23% (0.42)	24% (0.42)
% Town	6% (0.24)	13% (0.34)	6% (0.24)	16% (0.37)	3% (0.17)	10% (0.30)	8% (0.27)	14% (0.35)
% Rural	13% (0.34)	22% (0.41)	14% (0.34)	23% (0.42)	5% (0.22)	13% (0.33)	11% (0.31)	17% (0.38)
% Female	49% (0.03)	49% (0.03)	49% (0.03)	49% (0.03)	49% (0.03)	49% (0.04)	49% (0.03)	49% (0.03)
% White	30% (0.31)	51% (0.33)	25% (0.31)	43% (0.34)	7% (0.14)	17% (0.24)	17% (0.26)	29% (0.29)
% Black	21% (0.29)	15% (0.24)	25% (0.31)	19% (0.27)	32% (0.35)	31% (0.34)	26% (0.34)	26% (0.31)
% Hispanic	42% (0.32)	24% (0.27)	45% (0.33)	29% (0.30)	57% (0.34)	43% (0.36)	51% (0.35)	37% (0.34)
% FRL	66% (0.28)	54% (0.29)	77% (0.17)	71% (0.18)	91% (0.05)	92% (0.06)	76% (0.24)	75% (0.22)
Stud/Tchr Ratio	18.13 (4.81)	16.42 (4.15)	17.27 (4.26)	16.41 (4.15)	17.41 (4.10)	17.01 (4.50)	19.66 (5.07)	17.85 (4.73)
% ELL	15% (0.15)	10% (0.11)	15% (0.16)	11% (0.12)	18% (0.11)	16% (0.14)	19% (0.12)	15% (0.13)
# District Schools	303.85 (512.36)	80.72 (234.85)	359.88 (549.26)	97.68 (265.83)	484.62 (598.86)	157.56 (329.39)	205.55 (303.65)	96.9 (215.66)
N (% of Total)	921 (100%)	54,898 (100%)	730 (79%)	37,021 (67%)	386 (42%)	13,134 (24%)	404 (44%)	15,246 (28%)
Generalizability Index		0.59		0.45		0.21		0.58
Middle/High Schools								
	All Schools		High Poverty		Very High Poverty		Low Achieving Districts	
	Sample	Population	Sample	Population	Sample	Population	Sample	Population
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Students per school	798.08 (534.59)	640.47 (523.66)	812.2 (523.92)	598.26 (491.46)	770.04 (428.62)	574.11 (435.97)	780.63 (498.86)	601.06 (467.77)
% Urban	39% (0.49)	28% (0.45)	43% (0.50)	33% (0.47)	65% (0.48)	53% (0.50)	55% (0.50)	41% (0.49)
% Suburban	32% (0.47)	30% (0.46)	32% (0.47)	24% (0.42)	24% (0.43)	21% (0.42)	16% (0.36)	21% (0.41)
% Town	9% (0.29)	14% (0.34)	7% (0.26)	15% (.36)	3% (0.17)	9% (0.29)	9% (0.28)	14% (0.35)

% Rural	20% (0.40)	29% (0.45)	18% (0.38)	30% (0.46)	8% (0.27)	16% (0.37)	21% (0.41)	22% (0.42)
% Female	48% (0.07)	49% (0.05)	48% (0.08)	49% (0.05)	48% (0.06)	49% (0.06)	47% (0.09)	49% (0.05)
% White	44% (0.36)	54% (0.34)	31% (0.32)	45% (0.35)	11% (0.18)	17% (0.25)	28% (0.29)	31% (0.30)
% Black	21% (0.29)	15% (0.24)	28% (0.31)	19% (0.28)	38% (0.38)	32% (0.35)	39% (0.37)	26% (0.31)
% Hispanic	29% (0.31)	22% (0.26)	35% (0.33)	28% (0.30)	47% (0.38)	42% (0.37)	29% (0.33)	34% (0.33)
% FRL	58% (0.27)	52% (0.28)	71% (0.18)	68% (0.18)	89% (0.06)	92% (0.07)	73% (0.21)	73% (0.22)
Stud/Tchr Ratio	16.29 (3.71)	16.31 (4.58)	16.3 (3.66)	16.36 (4.60)	16.46 (3.83)	17.07 (4.93)	16.39 (3.27)	17.52 (5.03)
% ELL	11% (0.12)	9% (0.11)	14% (0.12)	11% (0.12)	17% (0.13)	17% (0.14)	12% (0.12)	15% (0.13)
# District Schools	108.4 (205.03)	77.11 (241.81)	134.01 (231.93)	100.32 (286.62)	183.65 (270.29)	173.51 (363.75)	82.43 (104.01)	92.11 (210.683)
N (% of Total)	558 (100%)	50,055 (100%)	403 (72%)	32,324 (65%)	160 (29%)	9,695 (19%)	199 (36%)	14,136 (28%)
Generalizability Index		0.57		0.41		0.12		0.51

Table 4. Comparison of sample and population sizes by district size

		District Size				
		0.25% Largest	**0.25 - 2% Largest**	**2 - 5% Largest**	**5 - 10% Largest**	**90% Smallest**
		> 125 Schools	39 - 125 Schools	21 - 38 Schools	12 - 20 Schools	< 12 Schools
		Total (%)	Total (%)	Total (%)	Total (%)	Total (%)
Districts	Sample	29 (6%)	66 (15%)	70 (16%)	79 (18%)	206 (46%)
	Population	38 (0.25%)	253 (1.68%)	434 (3%)	771 (5%)	13,528 (90%)
Schools	Sample	454 (30%)	375 (25%)	207 (14%)	183 (12%)	260 (18%)
	Population	9,370 (11%)	13,405 (16%)	10,260 (12%)	9,930 (12%)	41,287 (49%)
	Population, High Poverty	7,691 (14%)	9,115 (17%)	6,900 (13%)	6,688 (12%)	24,739 (45%)
	Population, Very High Poverty	4,345 (25%)	3,795 (21%)	2,381 (13%)	1,935 (11%)	5,266 (30%)
	Population, Low-achieving	3,315 (15%)	4,618 (20%)	3,236 (14%)	3,230 (14%)	8,441 (37%)
Students	Sample	323,714 (33%)	283,762 (29%)	135,115 (14%)	113,385 (12%)	115,287 (12%)

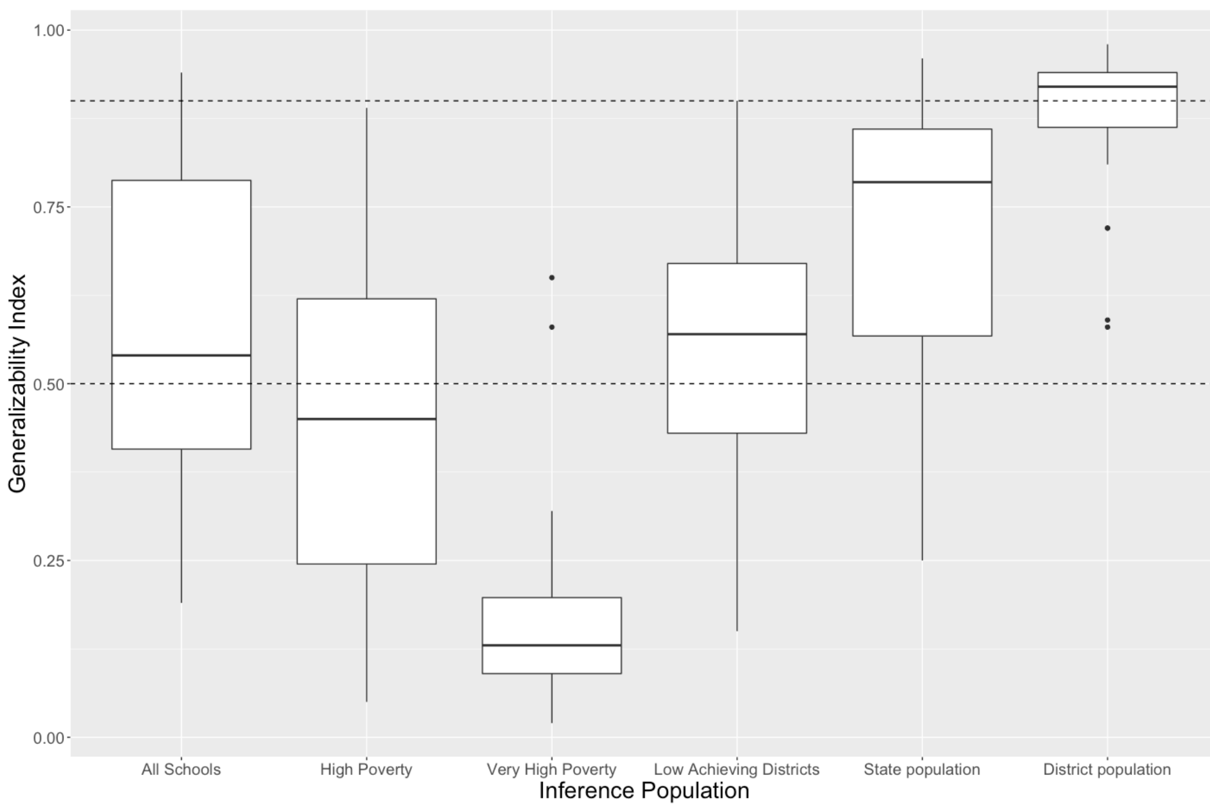
	Population	6,658,798 (14%)	9,822,199 (20%)	6,976,373 (14%)	6,255,638 (13%)	18,768,557 (39%)
	Population, High Poverty	5,267,514 (17%)	6,233,701 (21%)	4,391,634 (15%)	3,960,993 (13%)	10,420,461 (34%)
	Population, Very High Poverty	2,641,745 (28%)	2,271,626 (24%)	1,421,341 (15%)	1,038,298 (11%)	2,223,993 (23%)
	Population, Low-achieving	2,076,571 (16%)	2,882,214 (23%)	2,131,871 (17%)	1,926,224 (15%)	3,651,808 (28%)

Question 3: Study-specific target populations

While Question 2 focuses on national target populations, it is possible that individual researchers and studies were not focused on these populations, but instead on more narrowly defined target populations. We therefore conducted additional analyses of the interviews, seeking information on how the PIs conceived of their study target populations. Analyses of the interviews indicated that the majority of researchers either did not talk about generalizability goals at all (n = 8) or said they did not plan for generalizability ahead of time (n = 14). For those that did discuss generalizability, two explicitly planned for generalization (using a stratified sampling plan to specific populations) and two others mentioned planning for generalization less formally. Additionally, some researchers (n = 5) noted that their goal was to get enough diversity in the sample to generalize, but that it was not the main focus of their recruitment. Another two researchers argued that if their intervention was found to work in low-income, high poverty schools, then this would imply that it should work anywhere.

In addition to the interviews, we attempted to identify target populations by reviewing the study abstracts provided on the IES website. This approach did not prove fruitful, however, since abstracts rarely made distinctions between sample and population characteristics. For example, none of the abstracts included inclusion or exclusion criteria for the population, the size of the population, or any assessment of similarity between the sample and population. We therefore attempted to define possible study populations in relation to the study locations in two ways: first, as the state population of schools, and second as the population of schools found in the school districts in which each study took place. For each study, we therefore conducted six analyses, comparing the study sample of schools to each of the four national target populations previously defined, as well as to the state and school-district population of schools; for each a generalizability index value was calculated. Results were then aggregated over studies and results are presented as boxplots of the generalizability index values. In Figure 3, the first four boxplots correspond to the four previously defined national populations and the next two boxplots correspond to the more "local" target populations.

Figure 3. Comparison of similarity between each study and different populations



Note: Horizontal dashed lines indicate two rules of thumb: values of the generalizability index less than 0.50 are considered too different to generalize, while values greater than 0.90 are considered as similar as found in a random sample of the same size on the variables studied. Values in between the dashed lines indicate that statistical adjustments to both the estimates and standard errors would be required to generalize.

As Figure 3 shows, individual studies are typically no more similar to each of the four target populations than the overall analyses provided in Question 2. Notably, none of the individual studies were sufficiently similar to the population of very high-poverty schools to warrant generalizations (i.e., all values < .50). This was true, too, for nearly half of individual studies when compared to the population of all schools, high poverty schools, and schools in low-achieving districts. Only a small handful of studies had values high enough to indicate that only small adjustments would be needed to generalize; in most cases, values were such that strong statistical adjustments would be required for generalization, resulting in changes in the average treatment effects estimated and large increases in standard errors (see Tipton, 2014).

As the last two boxes in Figure 3 indicate, however, generalizations from the studies to state and school district populations were stronger. In general, studies could very clearly generalize to the populations of schools in the school districts in which they took place, and nearly all studies could generalize well – albeit with some statistical adjustments required – to the population of schools in the states in which they took place.

Discussion

In this discussion section, we summarize the findings from this study, with a focus both on understanding current practice and on possible avenues for changing this practice.

Current practice

Overall, we found that the schools taking part in these IES funded grants were typically found in large school districts in urban areas, located nearby to one another and study PIs. Importantly, our definition of "large" here focuses on the top 10% of school districts, which in practice means those with more than 11 schools. Even within this subset, the largest 0.25% of school districts were over-represented relative to the population. In comparison, schools in the 90% smallest school districts (< 12 schools), in rural areas and towns, and in states in the middle of the country – distant from research centers – were under-represented. Furthermore, at the study level, it was not possible to generalize well from the schools in most studies to any of the target populations defined. In fact, the strongest claims towards generalizability for individual studies were with respect to the states and school districts in which the studies took place – leading to a collective abundance of evidence in some states (e.g., Florida, Texas, California) and dearth of evidence in others (i.e., no evidence in 46% of states).

These findings largely align with those found by Stuart and colleagues (2017) in their analyses of samples in 11 contract-funded evaluations conducted before 2011. As our interview data make clear, these decisions regarding district size, urbanicity, and geography are driven by the constraints and costs around recruitment, a process in which total sample size is valued more than sample characteristics or representativeness. In this way, the results of this study also mirror those of previous research regarding the sample characteristics of participants in lab studies in psychology, wherein participants are most often college sophomores at elite colleges, since it is easier to recruit from introductory psychology courses than from outside the university (Heinrich, Heine, & Norenzayan, 2010; Peterson, 2001; Sears, 1986).

Furthermore, our interviews and analyses of study-abstracts indicate that when faced with questions about generalizability, PIs have little training or guidance in how to actually measure and address generalizability. For example, study abstracts rarely distinguish well between sample and population characteristics. It is typical, in fact, for the population to be defined *post hoc* based vaguely upon the location of the schools included in the study, not based upon *a priori* goals or problem prevalence. In comparison, PIs spoke of a need for achieving a given sample size – based upon a power analysis – as quickly and inexpensively as possible. Given the need to recruit 40 – 60 schools into a study, it is not surprising that large school districts – bringing with them many schools at once – were thus given priority.

Towards a better future

Given these trends for recruitment, if we take seriously that the results of large-scale randomized trials are meant to provide evidence for making decisions in broad and local target populations, what are we as a field to do? While this problem may seem daunting, we take solace in remembering that less than 20 years ago, the very idea that one day large-scale randomized trials could take place *and often* in education seemed impossible (Cook, 2002). In what follows, we provide three concrete steps that we, as a field, could take to improve practice.

1. RFAs drive change in practice

One simple approach for improving practices around generalizability in randomized trials is to require researchers to speak to this goal in the grant proposal process. Indeed, this has been the means through which statistical power analyses in IES randomized trials has improved over time (Spybrook, Zhang, Kelcey, & Dong, 2019; Spybrook, 2008; Spybrook & Raudenbush, 2009; Spybrook et al., 2016). Whereas the 2004 RFA simply requested that "[q]uantitative studies should, where sufficient information is available, include a power analysis to provide assurance that the sample is of sufficient size (RFA, page 9)," the 2016 RFA used much more prescriptive and statistical language:

> "Detail the procedure used to calculate either the power for detecting the minimum effect or the minimum detectable effect size. Include the following:

- The statistical formula you used;
- The parameters with known values used in the formula (e.g., number of clusters, number of participants within the clusters);
- The parameters whose values are estimated and how those estimates were made (e.g., intraclass correlations, role of covariates);
- Other aspects of the design and how they may affect power (e.g., stratified sampling/blocking, repeated observations); and
- Predicted attrition and how it was addressed in the power analysis (RFA, page 66)."

Spybrook and colleagues (2019) show that similar changes to the RFAs have also been made regarding power analyses for moderator effects (with the first explicit mention in 2012). They compared the structured abstracts of IES funded cluster-randomized trials before this RFA (i.e., 2004 – 2009) and after (i.e., 2013 – 2018) and found that before 2012 only 31% of studies identified moderators and/or described planned moderator analyses, while after the implementation of the language in the RFA, fully 75% of studies did so.

To some degree, the inclusion of generalizability in the RFA has also increased over time. However, even in the most recent RFA, unlike the language for statistical power, the language for generalizability is less statistical, leaving researchers little guidelines regarding *how* to define a target population, the sample recruitment process, or assess generalizability. Given the success of previous, prescriptive language, in Figure 4, we provide suggested RFA language, based upon guidelines established in the field (see Tipton & Olsen, 2018).

Figure 4: Suggested language for RFA regarding generalizability (based on Tipton & Olsen, 2018)

Detail the procedure that will be used to recruit a sample of schools that represents a target population in need of the proposed intervention. Include the following: • *Using population level data on schools, define and enumerate the target population of schools that would benefit from and possibly use the intervention under study.* • *Hypothesize ways in which the effect of the intervention might vary across schools and identify variables or their proxies in the target population dataset.* • *Define clear inclusion/exclusion criteria that the study uses that further narrows the target population for practical reasons (e.g., geography), and discuss how these constraints affect the ability to generalize to the above target population.* • *Describe the sample recruitment procedure that will be used to ensure similarity between the sample and target population, including statistics calculated, metrics for success, and planned adjustments for any resulting mismatch between the sample and population.*

2. Additional support and guidance is required

We should be mindful that changes to the RFA alone, without additional supports, are unlikely to change practice. In the case of statistical power, these RFA changes were bolstered by the availability of specialized software (e.g., *Optimal Design*, *PowerUp!*, *PowerUp-Moderator!*), tutorial papers regarding the use of this software (Dong & Maynard, 2013; Raudenbush, Martinez, & Spybrook, 2007; Spybrook, Kelcey, & Dong, 2016), and workshops on these methods. Fortunately, in the case of generalizability, these supports are largely already in place. For example, free software for improved population specification and recruitment planning is available (Tipton & Miller, 2015), and several tutorial and review papers have been provided (e.g., Tipton & Olsen, 2018), as well as workshops at conferences.

Software and training, alone, however, were not enough, even for statistical power. Spybrook and colleagues showed that early methodological papers indicated that for the effect sizes expected in trials in education, larger samples would be required, and that, over time, sample sizes did in fact increase (Spybrook & Raudenbush, 2009; Spybrook, Shi, et al., 2016). This need for larger samples has had downstream implications for funding, with recent arguments suggesting that given the smaller effect sizes observed in studies, perhaps the field would be better off with a smaller number of randomized trials, each with larger samples (Lortie-Forgues & Inglis, 2019). There are similar cost concerns for generalizability, as well, since any improvements would likely require researchers to shift to recruiting in places further away from urban research centers, and in more, but smaller, school districts.

Here it is important to highlight that in a system without improved resources, it is possible for researchers to follow the 'letter of the law' but not the 'spirit of the law'. That is, the most expedient path towards improved generalizability without additional cost is to identify the likely sample (e.g., nearby large urban district) and then to define the target population and recruitment plan accordingly (e.g., "schools in large school districts in urban areas"). While on the one hand, this results in greater clarity with respect to where results might apply, it does little to change the trends found in this paper. If greater representation of *all* types of schools in a population is desired, then funders will need to identify supports that meet these goals.

At a minimum, these supports might include additional funding and longer recruitment and grant timelines. Currently, RFAs request that letters of support are included from schools and school districts that would take part in the study. While on the one hand, these letters indicate that the study PI and team are capable of recruiting schools, on the other hand, this likely biases researchers towards recruiting large school districts that are nearby to them. Additionally, these letters suggest to PIs that they do not need to plan for additional recruitment time or resources in the grant itself, even though research indicates that in most studies, many of the schools in these letters of support ultimately do not agree to be in the study (Spybrook, Lininger, & Cullen, 2013). If we take seriously that the goal of these letters is to ensure that PIs have actually spoken with schools and can work with them, then perhaps there are other more effective approaches. For example, letters could be required for only one or two schools, or from some of the most difficult to recruit schools, as well as detailed recruitment plan for how they will meet their goals.

Finally, more broadly these resources might include the development of collective resources regarding partnerships between researchers and schools and their recruitment into evaluations. Intervention researchers are well poised to understand their own intervention and field, but less well versed in understanding population level data regarding the need for research in practice. Information on target populations, their particular needs, and the types of curricula already implemented in these schools is essential to pivoting practice. This requires supplementing intervention research with rigorous descriptive research on the problems, constraints, and opportunities faced in schools.

3. Research on best-practices in recruitment

In our interviews, we often heard researchers reflect on their early experiences in recruitment, noting that they wish they knew then what they know now about this process. This was particularly true for PIs at universities; those in research firms often had access to this craft knowledge developed across the history of the organization. We were struck by how few opportunities there were for PIs to share this knowledge with one another, across PIs and institutions.

What we are calling for here, however, is not simply a space in journals for PIs to reflect on their recruitment stories, but instead for recruitment to be treated as a scientific enterprise. In the field of sample surveys, there is an entire literature on non-response that includes experiments manipulating different approaches in order to determine optimal strategies for reducing non-response bias (e.g., Curtin, Presser, & Singer, 2005; Singer, 2002; Wagner, 2008). In randomized trials, such a literature could be developed, too. As a starting ground, it would require PIs to collect data on recruitment – i.e., which schools were contacted, strategies used and incentives offered for participation, and reasons given for not taking part. Preliminary studies indicate that this is both possible and useful, providing researchers with a sense of the types of schools interested in taking up a program (Tipton et al., 2016). This first stage would itself provide information on current practice in the field – e.g., what types of incentives are offered to schools and which appear to work.

At a more advanced level, this would involve actually developing multiple possible approaches to recruitment and embedding experiments comparing these approaches within evaluations. Current approaches to experiment with might include the mode of request (e.g., fliers, emails, and websites versus in-person meetings), type of recruiter (e.g., former teacher, study PI), framing of the need for research, incentives offered (e.g., monetary, training), and timing of request.

Conclusion

In this study we presented the results of a study of recruitment practices and sample characteristics of large, field-based experiments funded by grants from IES. We show that the current practice in the field leads to over-representation of large schools in large, urban school districts and under-representation of small schools in small, rural and town school districts. These trends hold in both elementary and middle/ high-school studies and for populations of very high poverty schools and those in low achieving school districts. In the discussion we argue that if we take seriously the demands for evidence from practitioners and policymakers, that we need to find ways to improve recruitment practices. We show that previous attempts to improve statistical power of randomized trials offer us a pathway forward for improving the generalizability of this research base. Our hope is that by making these connections, we can nudge the field towards a system of science that harmonizes with the needs of schools, teachers, and students.

References

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a Culture of Replication: An Examination of Education and Special Education Research Grants Funded by the Institute of Education Sciences. *Educational Researcher*, *47*(9), 594–605.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*(1), 87–98.

Department of Education. (2016). *Non-Regulatory Guidance: Using Evidence to Strengthen Education Investments*. Retrieved from https://ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143.

Fellers, L. (2017). *Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences*. Columbia University.

Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(3), 757–778.

Heinrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher*, *48*(3), 158–166.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*(3), 450–461.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. https://doi.org/10.3102/0162373707299460

Saldana, J. (2016). *The Coding Manual for Qualitative Researchers.* Thousand Oaks, CA: SAGE Publications, Inc.

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR? *Brookings Papers on Education Policy*, (9), 205–228.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515.

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. *Survey Nonresponse*, *51*, 163–177.

Spybrook, J, Lininger, M., & Cullen, A. (2013). From planning to implementation: An examination of changes in the research design, sample size, and statistical power of group randomized trials launched by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, *6*(4), 396–420.

Spybrook, J, Zhang, Q., Kelcey, B., & Dong, N. (2019). Learning from Cluster Randomized Trials in Education: An Assessment of the Capacity of Studies to Determine What Works, For Whom, and Under What Conditions. *Working Paper*.

Spybrook, J. (2008). Are power analyses reported with adequate detail? Evidence from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, *1*(3), 215–235.

Spybrook, J, Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two- and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*(6), 605–627.

Spybrook, J, & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–318.

Spybrook, J, Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the US Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*(3), 255–267.

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

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

Tipton, E, & Miller, K. (2015). *The Generalizer*. Retrieved from www.thegeneralizer.org

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266.

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

Tipton, E, 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.

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.

Tipton, E, & Olsen, R. B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher*, 0013189X18781522.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. (2016). *Common Core of Data*.

Wagner, J. R. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias.*

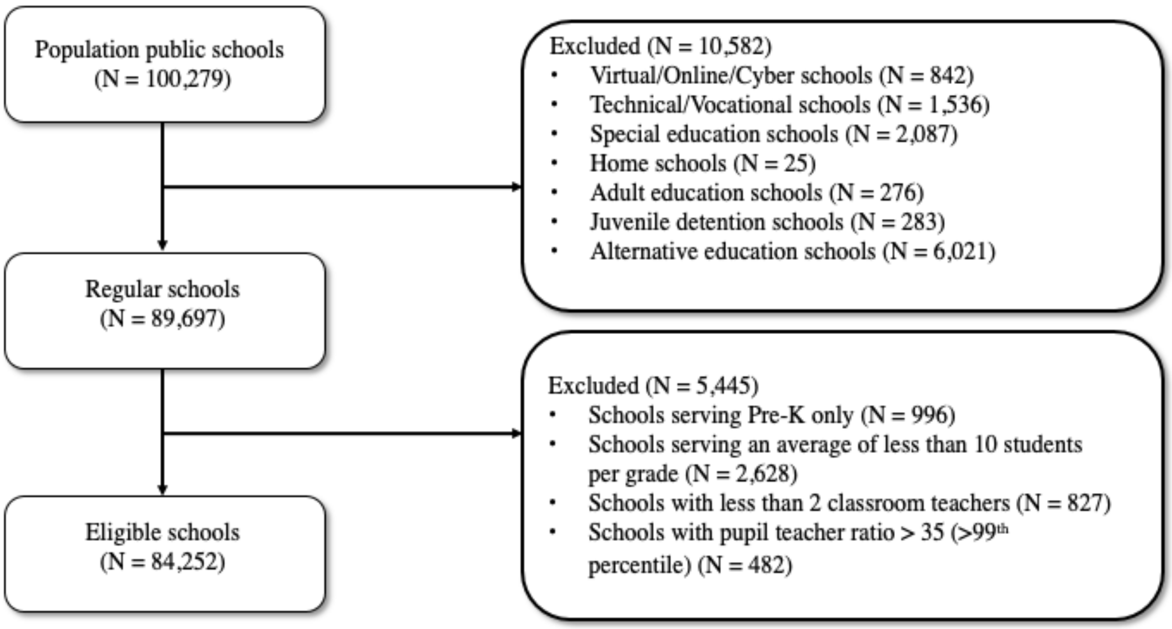
Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag.

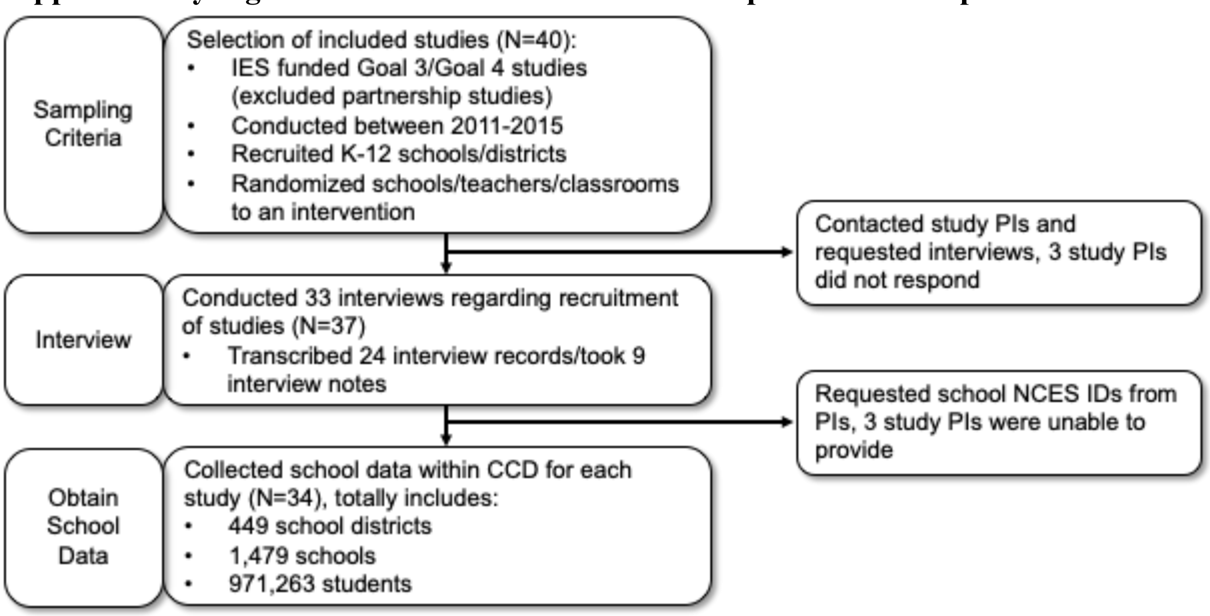
Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., … Dweck, C. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369.

Supplementary Material

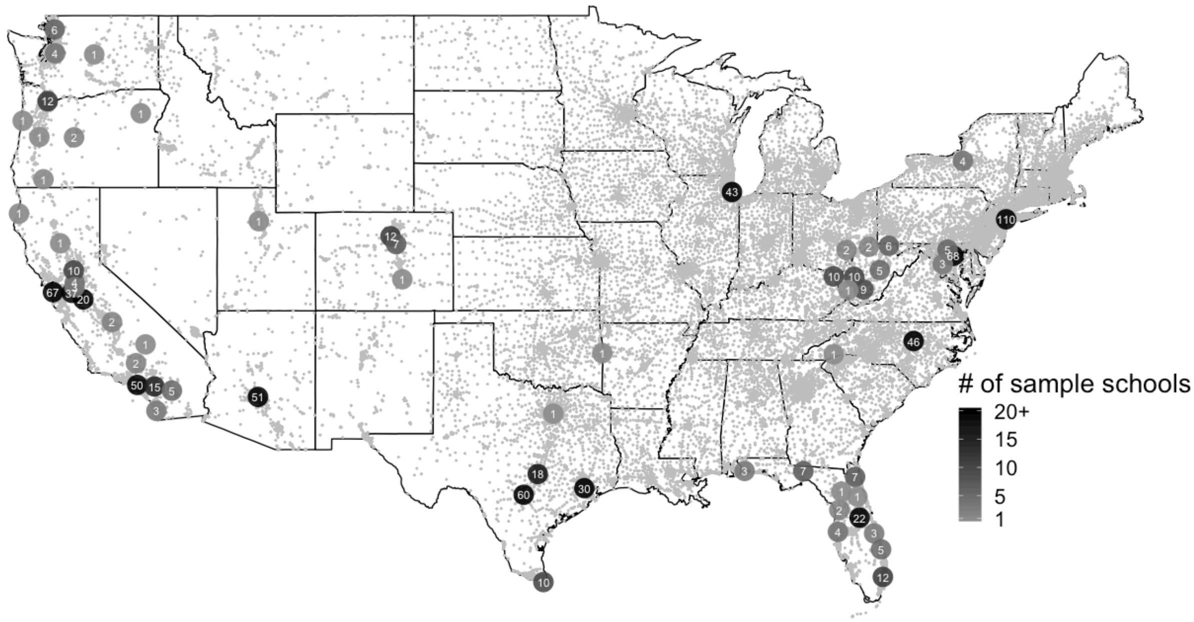
Supplementary Figure 1. Flow-chart of population exclusion criteria



Supplementary Figure 2. Flow-chart of data collection process for sample



Supplementary Figure 3. Map of Elementary Schools in Studies versus Population



Supplementary Figure 4. Map of Middle/High Schools in Studies versus Population