Toward a System of Evidence for All: Current Practices and Future Opportunities in 37 Randomized Trials

Elizabeth Tipton[1], Jessaca Spybrook[2], Kaitlyn G. Fitzgerald[1], Qian Wang[2], and Caryn Davidson[2]

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 20 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 article, we report on a study of 37 cluster randomized trials funded by the Institute of Education Sciences between 2011 and 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 were analyzed to determine the general demographics of schools included in funded research, as well as how these samples compare to important policy relevant populations. We 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 overrepresented, whereas schools from small school districts in rural areas and towns are underrepresented. The article concludes with a discussion of how recruitment practices might be improved in order to meet the goals of the evidence-based decision-making movement.

Keywords: decision making; descriptive analysis; educational policy; evaluation; mixed methods; program evaluation; research methodology; research utilization; survey research

Over the past 20 years, the evidence-based (EB) decision-making movement in education has dramatically increased the breadth and strength of evidence regarding the efficacy of both core curricula and supplementary programs used in schools. This movement has focused on the use of strong research designs—for example, randomized experiments and strong quasi-experiments—to determine if interventions actually *cause* improvements in student outcomes. Much of the focus of this movement has been on the importance of having a strong research design, since without one, it is unclear if changes in student outcomes are a result of the intervention itself or are actually the result of a myriad of other factors: for example, natural student growth over time; other programs or curricula in use; the types of schools implementing the program; the backgrounds and abilities of teachers implementing the program (or not); or the backgrounds, experiences, and abilities of students in these classrooms (Shadish et al., 2002).

Since its debut in 2002, the Institute of Education Sciences (IES) has provided the backbone of this EB decision-making movement in education. Over this period, 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 and, to date, has funded over 300 evaluations of Pre-K–16 curricular and other programs (Chhin et al., 2018; Spybrook, Shi, et al., 2016). Furthermore, IES expects both grant- and contract-funded evaluations to design studies with high internal validity, with priority given to randomized trials. These trials

[1]Northwestern University, Evanston, IL
[2]Western Michigan University, Kalamazoo, MI

Educational Researcher, Vol. 50 No. 3, pp. 145–156
DOI: 10.3102/0013189X20960686

typically include between 40 and 60 schools each, with roughly half receiving a new program and the other half continuing with business as usual (Spybrook, Shi, et al., 2016).

The fact that over 300 such causal impact studies have been conducted in a period of less than 20 years shows great promise in the quest to build a strong evidence base on the efficacy of educational programs. This effort has resulted in a robust database of interventions with strong evidence—found in the What Works Clearinghouse (WWC)—providing practitioners and policy-makers with information for sound decision making regarding program and curricular choices in schools. The field has learned that the effects of many educational interventions are smaller than expected but that even small effects can provide practically meaningful improvements in student outcomes (Lortie-Forgues & Inglis, 2019). However, these 300 interventions tested are only a fraction of those developed, marketed, and sold to schools each year.

Over this same period, first No Child Left Behind (NCLB) and more recently the Every Student Succeeds Act (ESSA) also began requiring school officials using federal funds to purchase curricula that had been evaluated previously and, ideally, evaluated using a strong causal design. The ESSA legislation, in fact, asks decision makers—for example, school district superintendents, principals, curriculum specialists—to seek evidence not only that a program works, but also that it has been shown to work in *populations like theirs* (e.g., U.S. Department of Education, 2016b). These questions regarding program efficacy in "populations like theirs" points to a broader question regarding the target populations for randomized trials and the capacity for studies to generalize to these target populations. As we will show in this article, information on the population and sample characteristics of the schools taking part in these randomized trials is not often available, which makes it challenging for policy-makers and school leaders to assess how well a program may work in their context given the unique features of their own locales, schools, teacher, and students.

To date there has been very little research on the target populations and types of schools participating in large randomized trials in education. In an evaluation of two IES-funded Goal 4 scale-up studies, Tipton and colleagues (2016) compared the schools taking part in the studies to different target populations and found that the samples in these trials represented only a small subset (less than one-third) 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, principal investigators (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 (FRL), 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 potential populations of schools in the United States. Questions about the match (or mismatch) between study samples and potential populations of interest 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 article, 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)[1] studies funded by IES between 2011 and 2015.[2] We ask three questions in this article:

1. **How—and under what constraints—are schools recruited into 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 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 potential target populations of schools, including narrower populations and broader policy-relevant populations?** 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 might apply to schools and contexts like their own. In practice, however, it can be difficult to define how broad or narrow such a population might be. Given the goals of IES, the broadest possible population of interest would include all public schools in the United States.

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	32,324	55,133
	(67%)	(65%)	(65%)
Very high poverty (>80% FRL)	13,134	9,695	17,722
	(24%)	(19%)	(21%)
Low-achieving districts (bottom 25th percentile of achievement)	15,246	14,136	22,487
	(28%)	(28%)	(27%)

Note. Elementary is defined as any school serving K–5 students and middle/high is any school serving Grade 6–12 students. In some cases, these categories overlap, for example, in K–12 schools. All schools and high-poverty schools are based on data in the Common Core of Data (CCD). Low-achieving districts are based on data from Stanford Education Data Archive (SEDA).

Additionally, given general concerns with equity in education research, we define three other potential populations of interest: high-poverty schools (>40% FRL students; this roughly corresponds to Title I schools), very-high-poverty schools (>80% FRL students), and schools in low-achieving districts (bottom 25th percentile). 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 potential target populations. We defined these populations using the 2015 to 2016 Common Core of Data (CCD) (U.S. Department of Education, 2016a). In order to define "all schools," we focus only on "regular" public schools found in the continental United States; further information on these inclusion criteria are found in Supplemental Figure 1 (Supplemental Figures 1 through 4 are available on the journal website). This population is further restricted to high-poverty and very-high-poverty schools based upon the percentage of students on FRL found in the CCD. Note that high-poverty schools represent about 65% of public schools in the United States, whereas 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.

Sample Data

We focus on a sample of IES Goal 3 (Efficacy) and Goal 4 (Effectiveness) studies funded between 2011 and 2015. This range of dates was selected in order to ensure (a) that 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 PIs.

Beginning in the fall of 2017, we contacted each PI and requested an interview regarding recruitment in their studies.[3] 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 article authors and typically lasted 30 to 40 minutes. Overall, we conducted interviews regarding recruitment in 37 of the studies (three 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 9 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%).[4] In each study, schools were located within the CCD in the year prior to recruitment;[5] when data were missing, data from the year of recruitment were used (e.g., if the school was new). In total, the final sample includes 34 studies, 449 school 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), and 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.

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 potential target populations (all schools, high poverty, very high poverty, low-achieving districts). As the literature on the generalization of causal effects indicates, these comparisons would ideally include all potential *moderators* of treatment effects (Stuart et al., 2011; Tipton, 2013). In the CCD, the possible moderators available include those related to locale and demographics. Of these we focus on the following set:

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)
- % FRL
- % English language learners (ELL)

Note that these variables include those related to concerns with educational equity—for example, student race, ethnicity, socioeconomic status, and gender—as well as contextual variables often of importance to policy-makers and practitioners (e.g., urbanicity). Certainly, there may be other variables of importance that are not included in this list—for example, baseline student achievement in schools, teacher turnover, student absenteeism—which is a limitation of this paper. Across the schools in these 34 studies, we calculated means and standard deviations for 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 potential target populations. We then compared each study sample to two "local" target populations, defined as the relevant state(s) and school district(s). This is particularly important given that the intent of Goal 3 studies is to establish the efficacy of a program under ideal conditions and with a homogeneous set of schools. As such, a narrower target population such as district or state is very reasonable. For each of these target populations, we made 34 separate comparisons to each of the study samples. For each school, we predicted the probability that a school would be in a study given the potential moderators under study.[6]

We summarized the degree of similarity between these estimated probabilities for schools in the study sample versus those in the population using the generalizability index (Tipton, 2014). The generalizability index is the geometric mean[7] of the two probability distributions (in the sample and population) and takes values between 0 and 1, with higher values indicating greater similarity; for example, a value of .80 indicates that the sample and population are 80% similar on the covariates under study. Values of the index greater than about .90 indicate that the sample and population are as similar on these variables as a random sample of the same size, whereas values less than about .50 indicate that the sample and population are so different that generalizations are unwarranted (Tipton, 2014; Tipton et al., 2017). Typically, low generalizability index values indicate that there are parts of the population that are simply not represented *at all* in the study—what is referred to as *undercoverage* (Tipton, 2013).

Note that this similarity index is conditional on the covariates included in the analysis. If there are important moderators that are not in this model, then the true degree of similarity might be *smaller* than that stated by the index. Finally, in order to maintain anonymity, we report these index values in aggregate for each of the six 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 nearly half of the studies (48%; $n = 18$), researchers relied on some type of connection to the districts whether direct or indirect. In 24% ($n = 9$) of these studies, the researchers had longer-standing relationships with districts either through providing prior professional development or conducting previous studies. These relationships 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. Although 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 (14%; $n = 5$). 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

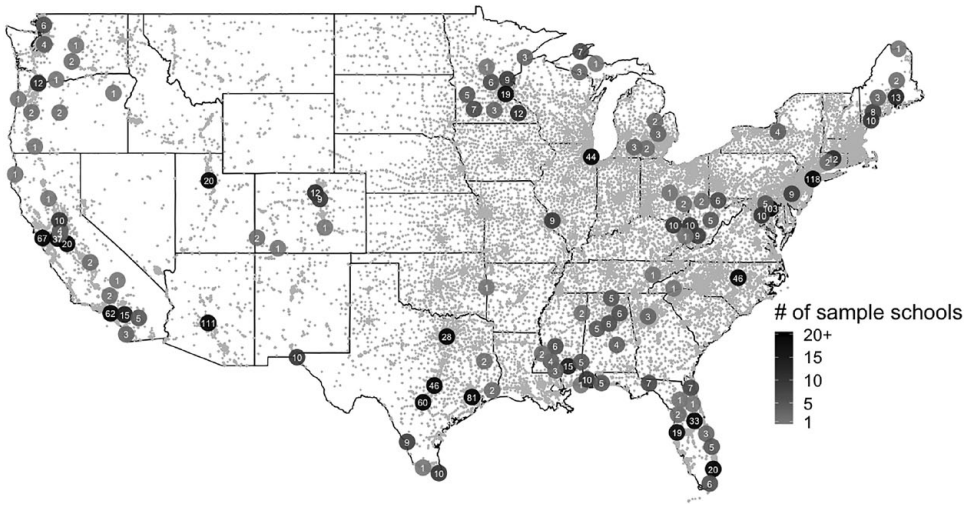

FIGURE 1. *Locations of study and population schools.*

intervention itself. Geography, in particular, played an important role (38%; $n = 14$). 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, although no clear preference was reported, district size played a role in recruitment (27%; $n = 10$). 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 (22%; $n = 8$). For example, interventions which were intended for a very specific population, such as ELL students in middle school, made recruiting challenging because the criteria were 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 1; 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 United States. Overall, the map suggests that states along the coasts are overrepresented in studies, with one-third 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 two to eight. 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 percentages were conducted both entirely within and out of state (36% within, 43% outside).

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 2. 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. Seventy-nine percent 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,

Table 2
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)
Student-teacher 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	921	54,898	730	37,021	386	13,134	404	15,246
(% of total)	(100%)	(100%)	(79%)	(67%)	(42%)	(24%)	(44%)	(28%)
Generalizability index	.59		.45		.21		.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)
Student-teacher 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	558	50,055	403	32,324	160	9,695	199	14,136
(% of total)	(100%)	(100%)	(72%)	(65%)	(29%)	(19%)	(36%)	(28%)
Generalizability Index	.57		.41		.12		.51	

Note. FRL = free or reduced-price lunch; ELL = English language learners.

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. Forty-two percent 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). Although 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 of very-high-poverty schools nationwide, and study schools served larger percentages of minority students than in the populations.

Schools in low-achieving districts. Forty-four percent 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 overrepresented relative to the population; and for elementary schools, town and rural schools were underrepresented.

The role of district size. In the analyses presented in Table 2, 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 nonlogged values are given, ranging from 1 school in a district to over 1,000 schools. As depicted in the figure, there is considerable mismatch between the district size in the U.S. population of public schools compared to those included in studies. First, fully 24% of school districts in the United States consist of a *single* school and another 49% consist of school districts with between two and five 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 United States but 6% in our sample.

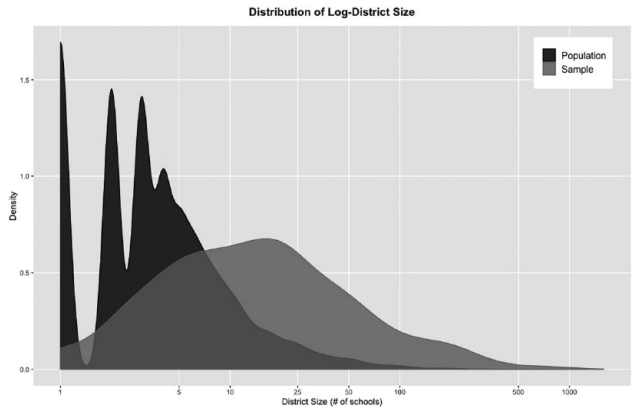

FIGURE 2. *Distribution of log-district size in studies versus total population.*

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 3, we present these findings. In this table, the population of school districts in the United States 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, although 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.

Although 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 composed of smaller schools (402.9 versus 662.0 students), whereas the average percentages of students on FRL are nearly equivalent (47% versus 52%).

Question 3: Study-Specific Target Populations

While Question 2 focuses on potential national target populations, it would be reasonable for individual researchers and studies, particularly Goal 3 studies, to be focused 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 (22%; $n = 8$) or said they did not plan for generalizability ahead of time (38%; $n = 14$). For those that did discuss generalizability, two explicitly

Table 3
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 district	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 district	2,076,571 (16%)	2,882,214 (23%)	2,131,871 (17%)	1,926,224 (15%)	3,651,808 (28%)

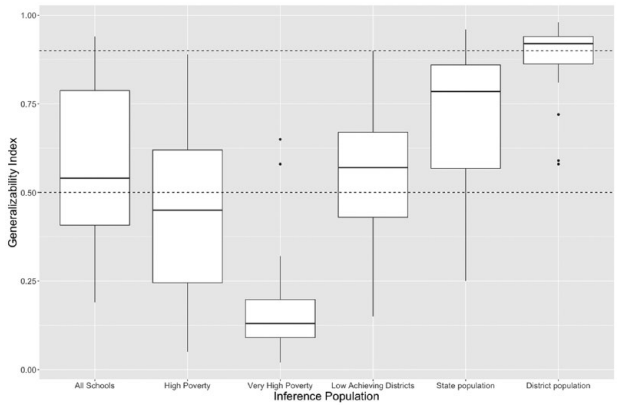

FIGURE 3. *Comparison of similarity between each study and different populations.*
Note. Horizontal dashed lines indicate two rules of thumb: Values of the generalizability index less than .50 are considered too different to generalize, whereas values greater than .90 are considered as similar as found 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.

planned for generalization (using a stratified sampling plan to specific populations) and two others mentioned planning for generalization less formally. Additionally, some researchers (14%; $n = 5$) 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 study sample and population. We therefore attempted to define possible study target 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 potential 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 potential national populations and the next two boxplots correspond to the more "local" target populations.

As Figure 3 shows, individual studies are typically no more similar to each of the four potential 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. As noted previously, the capacity for studies to generalize to the more local or narrow population rather than the four potential national populations is not surprising given the goals of Goal 3 studies.

Discussion

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

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 overrepresented 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 underrepresented. Importantly, these trends held for not only all public schools, but also for the populations of high-poverty and very-high-poverty school and schools in low-achieving districts.

At the study level, given the intent of Goal 3 studies, it is not surprising that it was not possible to generalize well from the schools in most studies to any of the potential target populations defined. As we would expect, 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). Although this trend is not surprising, the downstream implication is that there is very little evidence regarding program and intervention efficacy for the majority of contexts found in U.S. schools.

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.

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 to 60 schools into a study, it is not surprising that large school districts—bringing with them many schools at once—were thus given priority.

Where Do We Go From Here?

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? Although 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.

Requests for applications 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, 2008; Spybrook et al., 2020; Spybrook, Shi, et al., 2016; Spybrook & Raudenbush, 2009). Whereas the 2004 request for applications (RFA) simply requested that "quantitative studies should, where sufficient information is available, include a power analysis to provide assurance that the sample is of sufficient size" (Institute of Education Sciences, 2004, p. 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. (Institute of Education Sciences, 2016, p. 66)

Spybrook and colleagues (2020) 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, whereas 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

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.*

FIGURE 4. *Suggested language for RFA regarding generalizability.* Based on Tipton and Olsen (2018).

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).

Note that Figure 4 indicates that the ideal RFA would both require researchers to theorize about their intended target population *that would benefit* from an intervention (Step 1) and to separately refine this population based upon resource constraints within the study (e.g., geography; Step 3). Importantly, this approach does *not* require that the target population is national in scope or even broad; it could be narrowly defined based in a particular study.

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 et al., 2007; Spybrook, Kelcey, et al., 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 (*The Generalizer*; 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 clear implications for funding. More recently, as more information has become available regarding the (smaller) effect sizes typically observed in studies, these arguments for a smaller number of randomized trials, each with larger samples, has renewed (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"). Although 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 article. 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. Although 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 et al., 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. In the ideal, this might include federal incentives to schools to take part in evaluations, thus increasing the demand for research on the ground. Furthermore, 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.

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 nonresponse that includes experiments manipulating different approaches in order to determine optimal strategies for reducing nonresponse bias (e.g., Curtin et al., 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—that is, 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—for example, 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

We began this article by highlighting the remarkable growth in findings from large, randomized experiments in education over the past 20 years. In order for this to occur, significant resource investments were required to develop methods and guidelines for these studies, training to develop expertise, and communities of researchers to share best practices and approaches for improved learning. Given the increasing development and marketing of new educational programs, curricula, and supplementary materials, we expect that the need for strong evidence will continue to grow in the future.

At the same time, policy-makers and practitioners are increasingly concerned with determining not just what works in a study, but if it might work in a population, context, or school like theirs. This article indicates that the current system for conducting randomized trials does not adequately meet these goals. Information on study target populations is rarely reported well, and studies currently do not, on their own or in combination, represent well the diversity of school contexts found in the United States. In the face of this, we suggested several possible avenues for improvement, including changes that can occur at the researcher and study level, as well as at the level of funders (IES and others).

We conclude by noting that we are optimistic about the ability of the evidence-based community more broadly to meet these demands from practitioners and policy-makers. We now have a system in place to determine *causality* in education. What we need next is a system for *generalizing* this causal evidence and for translating this evidence to schools. Our hope is that by making these current shortcomings clear, we can nudge the field towards a system of science that harmonizes with the needs of schools, teachers, and students.

ORCID IDS

Elizabeth Tipton https://orcid.org/0000-0001-5608-1282
Kaitlyn G. Fitzgerald https://orcid.org/0000-0001-6569-4494

NOTES

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: U.S. Department of Education, Institute of Education Sciences (Grant Award R305D170024) and U.S. Department of Education, Institute of Education Sciences, Multidisciplinary Program in Education Sciences (Grant Award R305B140042).

[1]Of these 37 studies, 35 were Goal 3 and 2 were Goal 4 studies. This ratio is similar to the overall ratio of Goal 3 versus 4 studies funded by the Institute of Education Sciences (IES) since 2002.

[2]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.

[3]In three studies, instead of the principal investigator (PI), we interviewed the person in the study team that oversaw recruitment.

[4]In three studies, the PIs were unable to provide any data regarding schools; this was either because of the study institutional review board (IRB), or because the PI no longer had access to records.

[5]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 Common Core of Data (CCD) for us and provided us with a list of demographics which we then used in our analyses.

[6]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.

[7]If $s(x)$ is the distribution of propensity scores in the sample and $p(x)$ is the distribution in the population, then the generalizability index is defined as $\int\sqrt{s(x)p(x)}dx$.

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.

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. Retrieved from https://doi.org/10.1080/19345747.2012.673143.

Institute of Education Sciences. (2004). *Request for applications: Cognition and student learning research grants (NCER-05-07)*.

Institute of Education Sciences. (2016). *Request for applications: Education research grants (CFDA Number: 84.305A).*

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166.

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. Retrieved from https://doi.org/10.3102/0162373707299460

Saldana, J. (2016). *The coding manual for qualitative researchers*. SAGE Publications.

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

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

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., 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., & 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.

Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works, for whom, and under what conditions. *Educational Evaluation and Policy Analysis*, 0162373720929018.

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. (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*(Suppl. 1), 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., & Miller, K. (2015). *The generalizer*. Retrieved from www.thegeneralizer.org

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. (2016). *Non-regulatory guidance: Using evidence to strengthen education investments*. Retrieved from https://ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

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*. Doctoral diss., University of Michigan.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

AUTHORS

ELIZABETH TIPTON, PhD, is an associate professor of statistics and faculty fellow in the Institute for Policy Research at Northwestern University, 2040 Sheridan Road, Evanston, IL 60208; *tipton@northwestern.edu*. Her research focuses on methods for improving the generalizability of results from randomized trials, including methods for recruitment, study design, analysis, and meta-analysis.

JESSACA SPYBROOK, PhD, is a professor of evaluation, measurement and research at Western Michigan University, 1903 West Michigan Avenue, Kalamazoo, MI, 49008; *jessaca.spybrook@wmich.edu*. Her research focuses on improving the design and analysis of large-scale impact studies in education.

KAITLYN G. FITZGERALD, MS, is a PhD candidate in statistics at Northwestern University, 2006 Sheridan Rd., Evanston, IL 60208; *kgfitzgerald@u.northwestern.edu*. Her research focuses on methods for synthesizing and translating statistical evidence, including work in meta-analysis, data visualization, and statistical cognition.

QIAN WANG, MA, is a PhD candidate in evaluation, measurement, and research at Western Michigan University, 1903 Western Michigan Ave, Kalamazoo, MI 49008; *qian.97.wang@wmich.edu*. Her research focuses on quantitative research design and methods, meta-analysis, and research synthesis in education.

CARYN DAVIDSON, MAT, is a PhD candidate in evaluation, measurement and research at Western Michigan University, 1903 Western Michigan Ave, Kalamazoo, MI 49008; *caryn.k.davidson@wmich.edu*. Her research interests focus on teacher working conditions.

Manuscript received November 21, 2019
Revision received February 18, 2020
Accepted August 26, 2020