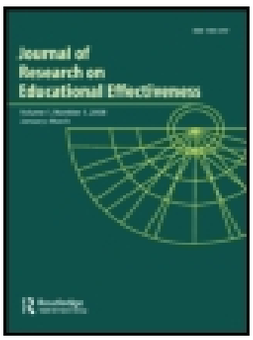

Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: http://www.tandfonline.com/loi/uree20

Site Selection in Experiments: An Assessment of Site Recruitment and Generalizability in Two Scale-up Studies

Elizabeth Tipton, Lauren Fellers, Sarah Caverly, Michael Vaden-Kiernan, Geoffrey Borman, Kate Sullivan & Veronica Ruiz de Castilla

To cite this article: Elizabeth Tipton, Lauren Fellers, Sarah Caverly, Michael Vaden-Kiernan, Geoffrey Borman, Kate Sullivan & Veronica Ruiz de Castilla (2016): Site Selection in Experiments: An Assessment of Site Recruitment and Generalizability in Two Scale-up Studies, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2015.1105895

To link to this article: http://dx.doi.org/10.1080/19345747.2015.1105895

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
http://dx.doi.org/10.1080/19345747.2015.1105895

METHODOLOGICAL STUDIES

Site Selection in Experiments: An Assessment of Site Recruitment and Generalizability in Two Scale-up Studies

Elizabeth Tipton[a], Lauren Fellers[a], Sarah Caverly[b], Michael Vaden-Kiernan[b], Geoffrey Borman[c], Kate Sullivan[b], and Veronica Ruiz de Castilla[b]

ABSTRACT

Recently, statisticians have begun developing methods to improve the generalizability of results from large-scale experiments in education. This work has included the development of methods for improved site selection when random sampling is infeasible, including the use of stratification and targeted recruitment strategies. This article provides the next step in this literature—a template for assessing generalizability after a study is completed. In this template, first records from the recruitment process are analyzed, comparing differences between those who agreed to be in the study and those who did not. Second, the final sample is compared to the original inference population and different possible subsets, with the goal of determining where the results best generalize (and where they do not). Throughout, these methods are situated in the post hoc analysis of results from two scale-up studies. The article ends with a discussion of the use of these methods more generally when reporting results from randomized trials.

KEYWORDS

generalization
external validity
recruitment

Over the past five years, statisticians and educational researchers have begun developing and implementing new methods aimed at improving generalizations from large-scale experiments (see Schochet, Puma, & Deke, 2014). This literature begins by requiring researchers to better identify and understand features of possible inference populations for whom the results of a study might be relevant, and then provides methods for site selection (Tipton, 2014b; Tipton et al., 2014), methods for assessing the similarity between the sites in a study and these possible inference populations (e.g., Olsen, Orr, Bell, & Stuart, 2013; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2014a), and methods for estimating average treatment impacts for these populations (e.g., O'Muircheartaigh & Hedges, 2014; Tipton, 2013). To date, this work has focused on generalizations over units (not time, outcomes, or settings) and on the development of new sample selection designs, new statistics, and new estimators, as well as investigations of their properties (e.g., see Kern, Stuart, Hill, & Green, 2016; Tipton, Hallberg, Hedges & Chan, 2015).

In comparison, to date very little work on generalization has focused on the *implementation challenges* that researchers face when attempting to design for and assess generalizability in

CONTACT Elizabeth Tipton tipton@tc.columbia.edu Teachers College, Columbia University, Human Development, 525 W. 120th Street, New York, NY 10027, USA.
[a]Columbia University, New York, New York, USA
[b]American Institutes for Research, Austin, Texas, USA
[c]University of Wisconsin–Madison, Madison, Wisconsin, USA

practice. For example, there is scant published information regarding strategies and incentives for recruitment (for an exception, see Roschelle et al., 2014) or for understanding biases that arise due to differences between the kinds of sites that agree to take part in experiments and those that refuse. Similarly, although methods for assessing generalizability have been developed—offering researchers tools for drawing boundaries around where the results of a study might apply and might not—little guidance on their use in "real" studies has been provided.

This article contributes to this literature on implementation in generalization by providing a post hoc analysis of generalizability in two IES scale-up studies (Everyday Math; Open Court Reading) conducted by SEDL[1] between 2011 and 2015. These studies are unique in many regards, including the fact that intended inference populations were defined at the outset and plans were made to recruit samples representative of them (see Tipton et al., 2014). Despite these intentions, however, recruitment in these studies was much more difficult than anticipated, and the resulting samples differed in many and large ways from the intended inference populations.

In this article, we begin by recounting the story of recruitment, including issues encountered along the way, and then tackle two important questions. First, we ask: what can we learn from these studies about the kinds of school districts that are willing to take part in large-scale experiments? This section of the article compares features of those districts that agreed to be in the study to those that did not, and summarizes feedback from the districts. Second, we ask: given the difficulties of recruitment and large differences between the final study samples and stated inference populations, are there any population subsets where the results may apply? This section of the article investigates a novel way to apply statistical methods for assessing generalizability in order to help researchers determine where the results may be most useful.

In the final section of the article, we discuss in detail lessons learned from these two scale-up studies regarding generalization. This section highlights issues uncovered in recruitment and in the generalizability analyses, as well as areas that need future work. Our hope is that this article will begin a conversation on practical concerns with generalizability and that it will motivate others to collect better data, to analyze this data, and to share the results broadly as an area of scholarly pursuit.

Two Scale-up Studies

This article is situated around two scale-up studies, one evaluating Open Court Reading (OCR) and the other evaluating Everyday Math (EM). Both the OCR and EM programs are aimed at elementary school students and are used widely in schools throughout the country (see Borman, Dowling, & Schneck, 2008; Slavin & Lake, 2007). Instead of conducting separate evaluations, SEDL decided to use a novel design in which the separately IES-funded scale-up studies would be combined. In addition to offering cost savings, combining the studies was conceived of as a recruitment incentive for districts because every school in the study would receive a program.

The fact that the two studies were combined meant that a single sample of school districts would need to be recruited for both evaluations. Within each of these recruited districts, four schools would be randomly selected, and of these, two schools would be randomly

[1] SEDL merged with AIR in May 2015.

assigned to receive the EM program while the other two would be assigned to receive the OCR program. The EM schools would then act as the control for the OCR schools, and vice versa. Based on a power analysis, the original study design aimed to recruit 15 districts, and within each district, four schools; this led to an anticipated total of 60 schools.

As scale-up studies, the goal for each was to determine what the program's effect would be in a broad, well-defined inference population. Based on discussions with the study team and technical working group, it was determined that the most relevant inference population for each study would be the population of schools and districts *currently using* the program. This focus was chosen because it indicated the types of schools and districts the programs were currently marketing to, and it was believed that without some exogenous change, future users were likely to be similar to those already using the program. With this in mind, the goal of the combined study was to estimate two population average treatment effects, one for the population of schools that currently use OCR, and the other for the current users of EM.

In Tipton et al. (2014), these populations of current users were carefully defined using the Common Core of Data. Each population frame included the school districts in the 48 contiguous states and Washington, DC, that at the time (Fall 2011) purchased any amount of the OCR or EM programs, respectively. The publisher, McGraw-Hill Education (MHE), provided Tipton and colleagues with user data from 2008–2010; overall, of the districts using either program, 30% purchased only the OCR program, 56% purchased only the EM program, and 14% purchased both programs. The OCR population therefore consisted of the 44% of districts purchasing the OCR program, while the EM population consisted of the 70% of districts purchasing the EM program.[2]

A unique feature of these evaluations was that the districts in the two inference populations were *not* actually eligible to be in the studies. Eligible schools were thus defined as those that had *not* purchased either the OCR or EM programs in the previous three years (2008–2010), as well as having at least four elementary schools with at least 44 students in each grade. This resulted in a total of 675 eligible school districts across the country. The aim therefore was to develop a method for selecting 15 school districts out of the 675 eligible districts that were compositionally similar to the OCR and EM user populations.

In order to achieve this goal, Tipton and colleagues divided the populations into strata using a propensity score approach. This method required that the researchers select a set of covariates that could explain heterogeneity in district-average-treatment effects (i.e., a *sampling ignorability condition*). The research team selected 11 factors (resulting in 21 variables) that they believed could potentially moderate the effectiveness of the OCR and EM programs based upon previous research and the curriculums' theories of change: these included district resources (e.g. revenue, location), family and community resources and context (e.g. labor force participation, location), and student characteristics (e.g. % of ELL students). Tipton and colleagues argued that if this list is comprehensive—in the sense that these covariates fully explain variation in treatment impacts—and the sample of districts in the experiment was compositionally similar to the intended inference populations, then the average treatment effects estimated in the evaluations would be unbiased for their respective population impacts.[3]

[2] Although for proprietary reasons we do not include the actual numbers of districts purchasing these programs from MGH, in the actual analyses these numbers were used.

[3] With two inference populations and a single sample, Tipton and colleagues recognized that compositional similarity would not be possible with a single sample for both populations. Instead the goal was to get a sample similar to the *average* of these two populations and then use post hoc adjustments for each.

The 675 eligible districts were then compared to each of these inference populations on this set of covariates and, using propensity score methods, divided into a total of nine strata. The study recruitment team was then given a list of eligible schools in each of the nine strata, as well as recruitment objectives for each (based upon proportional allocation of the 15 required districts). These strata would allow both a more diverse sample of districts to be recruited, and, when implemented as intended, for the resulting samples to be compositionally similar to the intended inference populations.

Recruitment and Refusals

Recruitment Process

Although the original study plan described above included plans to recruit 15 districts (and 60 schools) over the course of a single year, the recruitment process and final resulting sample differed in several important ways. In Figure 1 we provide an overview of this process using a flowchart; similar charts are often included in CONSORT statements in clinical trials (Egger, Juni, & Bartlett, 2001) and PRISMA statements in meta-analysis (Moher, Liberati, Tetzlaff, Altman, & the PRISMA Group, 2009). This figure highlights the amount of time needed to achieve the full sample, as well as changes in eligibility criteria over time and the total volume of districts contacted.

In comparison, although Figure 1 focuses on those contacted, in Table 1 we detail the number of districts and schools that actually entered the OCR and EM studies by year. As this table illustrates, while the final sample did implement the cross-over design in cohorts 1 and 2, none of the districts in cohort 3 took part in the design: five districts were involved in both studies as intended with the cross-over design, while two districts were only involved in the EM study, and two were only involved in the OCR study. Additionally, instead of 15 districts (and 60 schools), the final experiment included nine districts (and 74 schools), with seven districts taking part in either the EM or OCR studies. In the remainder of this section, we provide further details on this recruitment process.

As seen in Figure 1, the recruitment process included three waves of eligibility. Additional waves were needed when, after the first year of recruitment, only four districts had agreed to take part in the studies. In Year 1, all targeted districts were contacted via e-mail blasts, postcards, and cold calls to superintendents or other district officials. Recruiters also offered conference calls (nine in Year 1) with the study leads.[4] The initial study designs had included plans to work closely with McGraw-Hill during recruitment, in particular with sales representatives. However, in practice the research team found that the representatives were concerned that the experiment would compete with their sales quotas, and turnover at the publisher's office made developing contacts difficult. Without entrée provided by the publisher—or through a letter of support from the funder (IES), as is typical in contract research but not grant research—the research team found it difficult to inspire district offices to want to take part in the study.

In addition to these issues, in Year 1, the research team also realized that the cross-over design was less palatable to the districts than intended. The research team had combined the

[4] An additional eight conference calls were conducted in Year 2 and seven in Year 3. Additionally, in Year 2, one agreed to an on-site visit with the study team (San Diego).

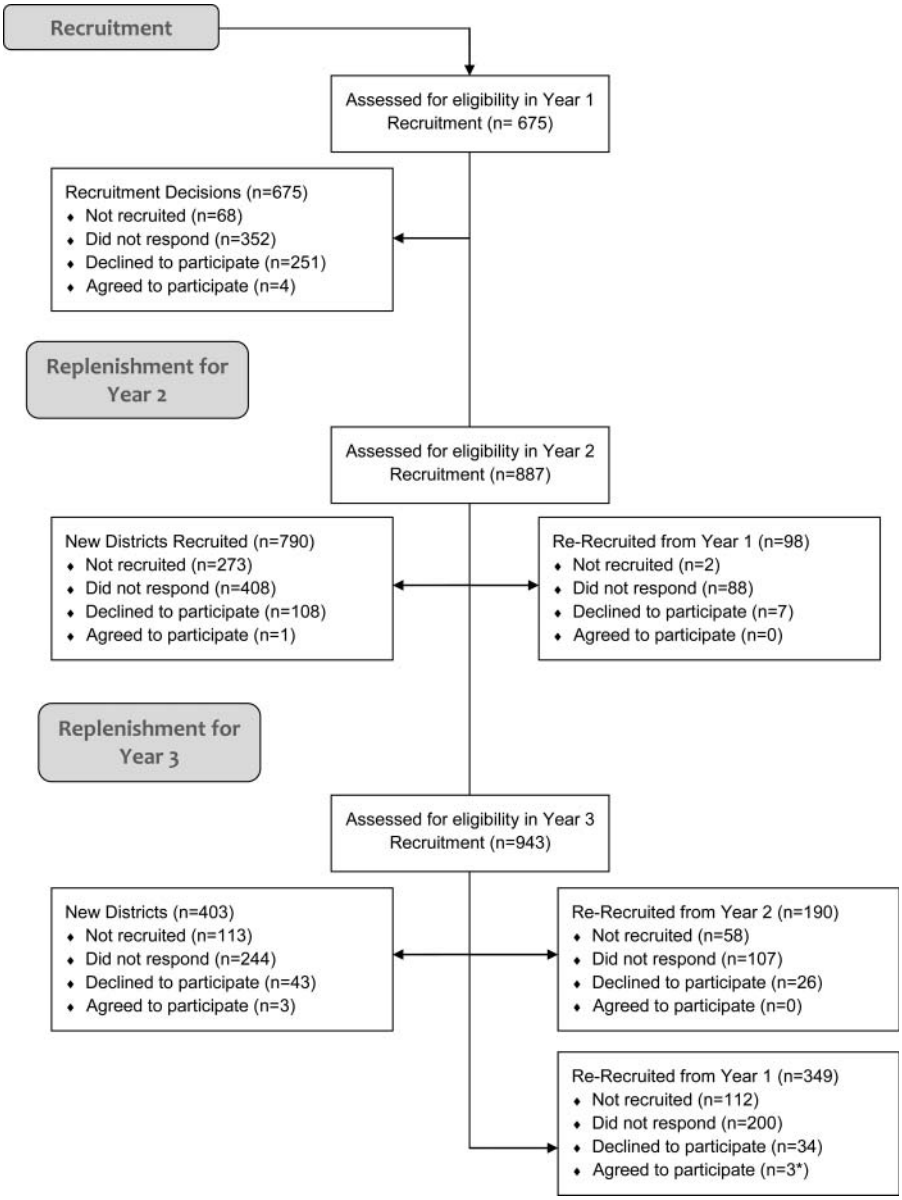

Figure 1. Flowchart of school district recruitment process.

studies with the intention of being able to provide every school in the study with a program (OCR or EM), which was conceived of as an incentive for participation. However, in practice, the recruiters found that often this design was not that desirable, possibly due to the fact that both experiments involved core programs. Thus, in order for a district or school to be involved, they would need to be open to implementing new core math or reading programs. Recruiters found that schools were in the market for either a core math *or* a core reading program, but not both. Additionally, many districts did not like the idea of having

Table 1. Final sample by wave of recruitment.

		Schools		
Recruitment year	District	Both OCR & EM	OCR only	EM only
Year 1	1	2	—	—
	2	4	—	—
	3	7	—	—
	4	7	—	—
Year 2	5	4	—	—
Year 3	6	—	9	—
	7	—	—	20
	8	—	—	4
	9	—	17	—
Total	9	24	26	24

multiple reading and math programs occurring across schools because it would make it challenging for students if they changed schools.

By the end of Year 1, with only four districts in the study, the research team decided to make three important changes. First, they decided to replenish the list of eligible schools in the hopes of identifying new districts (and a larger potential pool) to recruit. To do so, they shifted the eligibility criteria to include districts that hadn't used the program from 2009–2011 (versus 2008–2010 for Year 1). Additionally, based on better information from the publisher, a larger set of districts was considered eligible based on the definition of "purchased." In Year 1, "purchased any" had included districts that had spent any amount on the programs, which sometimes meant they had purchased an evaluation set but not the actual program. In Year 2 (and later Year 3), this rule was relaxed, and "purchased any" included only those that spent enough to indicate that a school had implemented the program. In total, these changes meant that in Year 2 and then Year 3, one and four additional districts, respectively, were added to the experiment.[5] Over the three years, a total of 1,917 unique districts were thus contacted regarding enrollment in the study.

Additionally, in Years 2 and 3, the requirement for the cross-over design was relaxed, allowing districts and schools to decide if they wanted to take part in either the OCR, EM, or both studies. Given the difficulties recruiting districts, the decision was also made to focus on the required sample of 60 schools, with the goal of getting as many districts as possible (with 60 schools), but without requiring the final number to be 15; this decision had been made in the first year, when two of the districts requested larger implementations (both with seven schools each). This meant that recruiters were able to open up the possibility to districts of having more than four schools involved in the experiment, which in some cases provided an incentive to participate. Finally, in reaction to pressures from the funding agency (IES) to ensure that the timeline for the project was met—and that an adequate sample was achieved—the focus on the original recruitment plan outlined by Tipton and colleagues was set aside as the primary guide for recruitment. Although informally the focus remained on getting as diverse a sample as possible, the focus on getting *any* sample, but particularly a large enough sample of schools (but not necessarily districts) to meet the statistical power needed for the study, became the primary goal.

[5] Year 3 recruitment actually resulted in six total districts agreeing to participate in the study, but for district circumstantial reasons two districts had to drop out of the study before randomization.

Refusal Analysis

Although one concern with generalization is that the final sample of sites is similar to the intended population, another concern is the extent to which sites that agree to be in an experiment (i.e., volunteers) differ from those that were recruited.[6] Because the goal of both the OCR and EM evaluations was to understand the impact of the programs in broad inference populations and recruitment was based on comprehensive lists of eligible school districts, the recruitment team was encouraged to keep track of feedback from the districts. For each eligible district that was contacted, recruiters tracked four possible exit points, found in Figure 1: never contacted (NC); never responded (NR); yes (Y); and no (N). Each year, the recruitment team first excluded some districts ("never contacted") based on knowledge they had regarding the school district; for example, they may have known that a district was undergoing a change in governance, or that it had a long process for research access (e.g., research applications with reviews spanning up to four months for approval). Once a district gave a definitive "yes" or "no," they were removed from further contact in subsequent waves. If they responded "no," in many cases further information was collected from the recruiters, including any reasons given for not participating and from whom. Those that "never responded," however, continued to be contacted until the study ended. This meant that districts were contacted for 1–3 years.

The fact that information on recruitment was collected provides an opportunity to better understand how districts that ultimately agree to be in a large-scale experiment differ from those that decline participation. In Table 2, we compare the nine districts that ultimately agreed to participate across the three waves of collection with those in the other three categories ("no," "never responded," "never contacted") on 12 factors; these include the 11 factors used to create the strata (see Tipton et al., 2014), as well as an additional variable indicating the number of schools in the district (which emerged as an important variable during the recruitment process because larger districts were desirable). These values given here are standardized in relation to the population standard deviation for each of the four outcomes. In the table, bold values indicate when the absolute standardized mean difference is significantly different from zero ($p < 0.05$); this bolding is meant only as a guide, not a hard rule, given the small sample sizes.

As Table 2 indicates, districts that were a priori not recruited at all differed more from the total set of eligible districts than either the nonresponders or refusers. Districts that agreed to be in the study ("yes") varied from the average eligible district on 12 out of the 24 variables. For example, the districts taking part in the experiment were, on average, poorer on multiple measures (i.e., all had lower education levels, higher poverty, lower employment, more free and reduced-price lunch qualifiers, and lower median incomes), especially in the EM study.

The quantitative differences between those districts that agreed to be in the experiment and those that did not were also echoed by the recruitment team's assessment of the process. In many instances there was no reason given ($n = 155$ districts). Of those giving a reason for declining participation ($n = 317$), the most common reasons for declining were that districts were currently satisfied with their math/reading programs (20%), had recently adopted new programs (20%), were adjusting and focusing on the goals of alignment with Common Core State Standards (CCSS) (6%), were in the middle of administrative changes (8%), or other

[6] The inference population and those recruited might differ if there are eligibility requirements.

Table 2. Comparison of means by recruitment decision.

		Total recruitment ($n = 1{,}917$)					
		All eligible $n = 2{,}421$		NR $n = 584$	DR $n = 1{,}368$	N $n = 459$	Y $n = 10$
Category	Covariates	Mean	*SD*	SMD	SMD	SMD	SMD
Student	% students ELL	49.9	23.7	0.05	−0.02	−0.06	**0.95**
	% students F/RL	10.6	13.8	0.23	0.01	−0.08	**0.67**
	Race/ethnicity of district						
	% White	54.9	29.1	−0.13	−0.05	0.06	**−0.56**
	% Hispanic	23.5	25.8	0.18	0.04	−0.04	0.20
	% Black/African American	12.9	18.3	−0.10	0.01	−0.04	**0.55**
	% other	7.4	10.7	0.16	0.01	0.01	0.23
Community	Educational attainment						
	% Grade 8 or lower	6.6	6.0	0.09	0.01	−0.13	**0.44**
	% <HS grad	8.7	4.2	0.03	−0.02	−0.14	**0.87**
	% HS grad	18.9	3.4	−0.04	−0.04	0.01	0.31
	% Postsecondary	12.9	7.1	−0.02	0.03	0.09	**−0.67**
	% 5- to 17-year-olds in poverty	17.3	11.1	−0.03	0.00	−0.09	**0.79**
	% labor force	58.4	7.2	−0.10	0.04	0.14	**−0.76**
Census area financials	Median income (overall)	$ 57,989	$ 21,195	0.06	0.01	0.09	**−0.66**
District	Urbanicity of districts						
	% Urban	23.3	37.2	0.00	0.05	0.03	**0.43**
	% Suburban	39.7	41.3	−0.08	0.06	0.05	**−0.60**
	% Town or rural	36.9	41.6	0.09	−0.10	−0.08	0.22
	Geographic location						
	% Northeast	15.9	36.6	−0.24	0.04	0.04	**−0.44**
	% Midwest	20.0	40.0	−0.15	−0.02	0.13	0.08
	% South	30.9	46.2	−0.06	0.00	−0.10	0.16
	% West	33.2	47.1	**0.38**	−0.01	−0.04	0.11
	District revenue (thousands)	$ 161,470	$ 303,277	−0.02	0.01	−0.01	−0.18
	Number of students in district*	7,099	11,871	0.02	0.01	−0.02	−0.06
	Number of schools in district*	13.17	19.39	0.00	0.00	0.01	0.06

Note: NR = not recruited; DR = did not respond; N = no; Y = yes.
*These variables were not used for stratification.

initiatives (8%). Some other reasons for declining participation were that districts were not interested specifically in these MHE curricula (6%), or that they were not interested in participating in an RCT (8%; either because there was not enough principal and teacher buy-in or because as a district they wanted all schools to implement the same programs). Some districts noted the time and budget constraints, as well as constraints on teachers who would have to adopt these programs in a short turnaround time.

Importantly, a significant portion of districts never responded to the recruitment materials, including postcards, e-mail blasts, and phone calls. In some cases, recruiters found it difficult to determine who the key decision makers regarding a program were in the district; for example, in some districts the decision maker was the director of curriculum and instruction while in others it was the superintendent and/or school board. In some instances, districts became nonresponsive once they had received additional information on the study. In other cases, district leadership did not return calls or respond to initial e-mails.

Overall, this analysis suggests two important findings. First, the kinds of districts that agree to be in large-scale experiments do in fact differ from those that are eligible. In this experiment in particular, districts agreeing to take part tended to be poorer. The fact that

one of the main incentives to participate was receiving the OCR and/or EM programs for free could be driving this result, though it may be that it is generally more difficult to motivate middle- and upper-middle-class school districts to be in an evaluation. Second, of those actively declining participation, the main reason given was that district resources were tied up with other changes, be they competing programs, new standards or requirements, or other administrative changes. In the "Discussion" section we address these findings in more detail, with a focus on implications for future studies.

Assessment of Generalization

As the previous section showed, the sample selection process that *actually* occurred differed markedly from the proposed method. Given the general difficulty in recruiting districts and schools into the study, an important question is whether the final samples in each evaluation are all that similar to their intended inference populations. If there are differences—and not surprisingly, we will show that there are—a second question, which we explore in great detail here, is whether it is at all possible to find some subpopulation for whom they are similar. Importantly, here we focus only on the inference populations (and their subsets) defined at the outset of the study—the OCR and EM populations of current users—and on the covariates specified for stratification in an effort to reduce bias. In the "Discussion" section we discuss the importance of this focus in greater detail.

In order to make these comparisons and assessments of similarity, unlike our previous focus on district comparisons, in this section we concentrate instead on comparisons at the *school* level. We do so because each of the two evaluations included only seven school districts (not the intended 15). Given this small number of districts, the outcomes analysis (reported elsewhere) treats the districts as fixed effects, with schools treated as random. Furthermore, as results from Tipton et al. (2015) indicate, it is difficult to make sound statistical comparisons when the number of covariates studied (here 17) is greater than the number of sites in the sample (here 7).

A difficulty with making school-level generalizations in this study is that the populations of current users provided from McGraw-Hill only included district information, not school information. It is not possible from the available information to identify which schools within the districts were actually using OCR and/or EM. In order to conduct this analysis, we began by defining the population of current users by including every non-magnet, non-charter elementary school in the districts identified by the publisher. We also included in this data set the 48 schools in the EM and the 50 schools in the OCR studies; here 24 schools were in EM only, 26 were in OCR only, and 24 were in both studies. We then pulled data from the Common Core of Data on school-level versions of the covariates included in the original stratification plan; for example, instead of the proportion of the district students eligible for free or reduced-price lunch programs, we now included the proportion of the students in each school eligible for free or reduced-price lunch. A subset of these variables remained at the district level (e.g., the ELL proportion, and all census-oriented measures). Finally, we excluded from the analysis the variables defining geographic regions. These had initially been included in the stratification plan for face validity purposes; once the other covariates were taken into account, the evaluation team did not believe that treatment effects would vary in relation to geography alone.

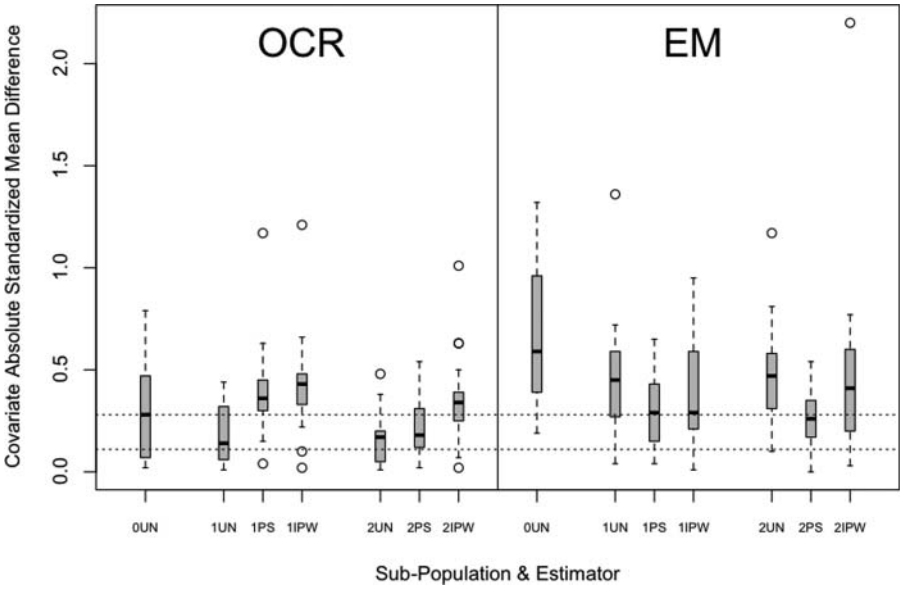

Figure 2. Covariate ASMD values by study, sub-population, and estimator.

Method for Assessing Generalizability

In order to summarize the degree of similarity between the sample of schools found in each study and the intended inference populations on the set of 17 covariates (based on 11 factors), we calculated a *generalizability index* (Tipton, 2014a). This index is calculated using three steps. First, schools in the sample are compared to those in the inference population using a logistic regression model including covariates previously specified. Second, based on this model, histograms of the estimated propensity score logits are compared. These histograms include $j = 1 \ldots k$ bins, each containing the proportion w_{pj} of the inference population and w_{sj} of the sample. Third, the generalizability index is calculated as

$$B = \sum_{j=1}^{k} \sqrt{w_{p_j} w_{s_j}}$$

which in general takes values between 0 and 1. Tipton (2014a) shows that values greater than about 0.90 indicate that the sample is as similar to the inference population as a comparable random sample of the same size, indicating "very high" generalizability. When below this cutoff, the generalizability index values indicate how well post hoc adjustments may perform in estimation of the population average treatment impact. These estimators are most accurate when $B > 0.80$ ("high" generalizability), while values of $B < 0.50$ indicate that a less biased estimate of the population average treatment effect is not possible ("low" generalizability). When $0.50 < B < 0.80$, less biased post hoc estimators of the population average treatment impact may be possible, but often come with a large variance inflation penalty.

For the OCR study, this index was estimated to be 0.61, while for the EM study it was estimated to be 0.57, both considered "low" to "medium" generalizability. This limited generalizability is further evidenced by the large absolute standardized mean

Table 3. Comparison of subpopulations of samples of schools in OCR and EM studies.

		OCR populations							EM populations					
	OCR expt	Original pop		Sub-pop 1		Sub-pop 2		EM expt	Original pop		Sub-pop 1		Sub-pop 2	
	Mean	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	Mean	*SD*	Mean	*SD*	Mean	*SD*
F/RL %	60.1	58.1	29.0	59.9	22.4	61.2	20.8	75.7	51.2	29.4	65.7	21.7	64.0	20.3
Hispanic %	14.0	28.3	30.3	23.9	27.4	13.2	15.5	4.8	19.5	25.5	24.9	27.9	17.2	21.8
Black %	12.6	20.3	27.6	15.7	23.4	19.4	26.6	51.3	17.7	25.8	17.8	24.7	20.1	26.7
White %	65.7	42.1	32.8	52.2	32.2	60.0	30.4	41.0	54.1	33.6	50.2	34.6	58.1	33.4
Other %	7.7	7.9	14.6	7.1	14.4	6.3	15.4	2.9	7.1	12.7	6.0	14.4	3.8	9.1
ELL (district ave %)	9.0	12.7	13.2	10.7	14.6	7.1	11.0	2.7	9.0	10.8	11.8	14.2	5.0	4.8
Urban %	38.0	39.2	48.8	21.5	41.1	22.2	41.6	43.8	34.5	47.5	26.4	44.1	20.2	40.1
Suburban %	24.0	31.7	46.5	26.6	44.2	16.4	37.1	6.3	35.0	47.7	21.7	41.2	18.2	38.6
Rural/town %	38.0	29.1	45.4	51.9	50.0	61.4	48.7	50.0	30.4	46.0	51.9	50.0	61.6	48.6
District revenue (millions)	$ 122.55	$ 874.79	$ 1,625.45	$ 104.65	$ 91.29	$ 99.22	$ 94.57	$ 181.86	$ 736.20	$ 1,494.95	$ 119.13	$ 118.69	$ 100.35	$ 101.01
Education < 9th (district, %)	7.4	7.6	6.1	7.6	6.2	6.7	4.0	7.8	5.7	4.8	9.6	5.4	8.5	3.3
Education < HS grad (district, %)	10.4	9.1	3.9	9.8	4.0	10.5	3.7	13.0	8.0	3.8	11.4	3.6	11.7	3.5
Education HS grad (district, %)	20.7	18.2	3.1	19.8	2.8	20.7	2.2	19.4	18.3	3.7	19.6	2.5	20.3	2.1
Education > = College (district, %)	8.9	13.3	6.2	10.6	5.4	9.3	4.1	9.2	14.6	7.0	9.0	4.4	8.4	3.8
Poverty (district, %)	21.6	20.8	10.9	20.1	10.7	22.2	9.6	31.8	18.2	10.8	25.1	10.3	25.0	10.6
Employed labor force (district, %)	56.2	58.6	6.5	56.6	7.2	54.9	6.6	51.7	59.9	6.6	55.4	7.5	54.2	7.6
Median income (district)	$ 43,814	$ 53,637	$ 17,728	$ 50,405	$ 14,847	$ 44,076	$ 8,335	$ 37,111	$ 56,876	$ 20,568	$ 44,197	$ 12,097	$ 42,815	$ 11,685
Total students (district)*	5,957	35,685	61,359	4,652	3,495	4,316	3,462	9,789	29,580	57,250	4,816	3,781	4,131	3,417
Total schools (district)*	16	60	99	10	7	10	8	25	51	92	11	8	9	8
NE %*	0.0	9.3	29.0	10.4	30.6	9.0	29.0	0.0	16.3	36.9	18.3	38.7	20.0	40.0
MW %*	52.0	12.7	33.3	15.3	36.0	18.0	38.0	0.0	28.8	45.3	20.2	40.2	16.6	37.2
S %*	40.0	45.3	49.8	40.0	49.0	49.0	50.0	83.3	32.0	46.7	42.2	49.4	53.2	49.9
W %*	8.0	32.7	46.9	34.2	47.5	25.0	43.0	16.7	22.9	42.0	19.3	39.4	10.2	30.3
Total schools in study (% of orig pop)	50	16,351 (100%)		6,950 (42%)		4,669 (29%)		48	20,807 (100%)		3,811 (18%)		2,841 (14%)	
Generalizability index		0.61		0.81		0.87			0.57		0.76		0.81	

*These variables were not used in the original sampling plan.

differences (ASMD) for the covariates (seen in the "0UN" boxes of Figure 2), which are particularly large in the EM analysis. In Table 3 we also provide descriptive statistics for the populations and samples of schools in the two studies (here focus on the "unweighted" populations columns).

When the generalizability index is low to medium, Tipton (2014a) shows that because of undercoverage—parts of the inference population without similar schools in the experiment—it is difficult to impossible to get an estimate of the population average treatment effect without huge variance increases (and large biases remaining). This means that the generalization is doubly penalized: first, the sample is not sufficiently similar to the population to warrant generalizations directly (based on the usual sample average treatment effect estimate), and second, there is little statistically that can be done to adjust for these differences (outside of extrapolation). Even if a population average treatment effect estimator is possible (e.g., a reweighting estimator), Tipton shows that in these extreme cases, it typically results in large variance inflation. This suggests that a more fruitful approach may be to ask: is there some subpopulation to whom the results do generalize well?

In the broader propensity score literature found in observational studies, this undercoverage issue is referred to in terms of the distributions of propensity scores (or their logits) not sharing *common support* or *overlap*. In observational studies, one approach used to improve estimation therefore is to remove units outside this region of overlap. In generalization, this approach would mean removing sites from the inference population outside the support of the propensity score distribution in the sample; while doing so could improve estimation, we argue that this approach sacrifices the ability to clearly define the inference population to whom the results apply. Instead, in this section we develop an alternative approach based on covariate inclusion/exclusion criteria that allows for both improved estimation and allows for the inference population to remain clearly defined and interpretable.

Our goal is to search for possible subpopulations within the inference population to whom the sample of sites in the experiment is most similar. To do so, we compared possible reductions based upon the minimums and maximums of the original 17 covariates. We focus here on minimums and maximums of individual covariates since these restricted ranges can impact the overlap of propensity score distributions. For example, if the population includes schools from a wide range of sizes (e.g., 100–10,000 students) but the sample only includes smaller sizes (e.g., 100–1,000), then by reducing the inference population to only include schools with fewer then 1,000 students, measures of overlap—and thus properties of treatment effect estimators—should improve.

In addition to these first 17 variables that were prespecified (and used for the original stratified sampling plan), we also included four regional variables (e.g., NE) and two variables regarding district size. These six additional variables were included because they were available in the CCD, were correlated with these other variables, and offer easy population conceptualizations. For the OCR and EM studies separately and for each of the 23 covariates, we thus defined exclusions based upon the minimum and maximum value observed in the sample, and then we examined all possible subsets with one, two, or three exclusions.[7] For example, if the sample did not include any schools in the Northeast, a one-variable exclusion would result in redefining the original user population to exclude Northeastern schools. For

[7] We focus only on up to three exclusions because three already greatly reduces the population size for generalization.

each of these possible combinations, we then calculated the generalizability index, as well as the number of population cases that remained. We then chose the subset that resulted in the largest index and largest N (population size) combination. Although not included here, this analysis resulted in a clear negative relationship between the generalizability index and population size, with larger index values being associated with smaller and smaller subpopulations.

In Table 3 we report descriptive statistics comparing the original population and experiment, and comparing the two subpopulations and the experiment. Here subpopulation 2 is the one associated with the largest generalizability index value, regardless of the resulting sample size. In contrast, subpopulation 1 is defined as that with the largest number of population units, given a high generalizability index value (defined as 0.70 for EM and 0.80 for OCR). Finally, for each of these two subpopulations, we additionally used poststratification (PS) and inverse-probability weighting (IPW) to attempt to further reduce any differences between the experimental sample and the subpopulation. For poststratification, we divided the resulting subpopulations into three nearly equal population size strata each. In subpopulation 1, the first stratum was often larger than one third of the population, in order to guarantee that there was at least one experimental school in each stratum. We then used the poststratification estimator provided by O'Muircheartaigh and Hedges (2014). These approaches mirror the approaches used to estimate the population average treatment effect for the OCR and EM studies (and for the particular inference populations).

Finally, in order to assess which subpopulation and estimator is most aligned with the sample of schools in each experiment, we calculated the absolute standardized mean differences (ASMD) between the sample and each subpopulation and estimator (a total of five) for each covariate. The distribution of these values is indicated in a boxplot in Figure 2; here the left panel focuses on the OCR study whereas the right panel focuses on the EM study. Given the sample sizes in the experiment, we would expect that on average, the ASMD would be 0.11 if the sample were randomly selected from the given inference population, and that in 95% of random samples, the value of the ASMD would be below 0.28 (see Tipton et al., 2015 for a discussion). In this boxplot, these two values are indicated with horizontal lines as a tool to help assess similarity.

Generalizability Index Results

OCR Study

The school-level generalization analysis performed best for the OCR study, compared to the EM study. When comparing the sample of experimental schools to the originally defined population of schools in districts that use OCR, the generalizability index was calculated to be 0.61, indicating low to medium generalizability. As Figure 2 illustrates (see "0UN"), the covariate ASMDs averaged 0.33, with a maximum of 0.79, and 9 of the 17 were statistically different. In particular the largest differences were with respect to student demographics, as well as measures of educational attainment and poverty in the associated school district. For example, experimental schools were Whiter, and were in districts that had lower education levels, lower median incomes, and lower district revenues. We therefore sought a subpopulation in which the similarity between the experimental schools and population was higher, thus increasing the ability to make generalizations.

The first subpopulation was defined to maximize the population size conditional on the generalizability index being larger than 0.80. This subpopulation definition was based upon two covariates: district size and the proportion of students in a school that qualified for free or reduced-price lunch. As Table 3 highlights, the distribution of district size differed markedly between the experiment and original population, with the experiment including smaller districts (mean number of students = 5,957; max = 12,350) than the original population (mean = 35,590; max = 288,300). Schools in the experiment also typically served poorer students (min proportion free and reduced-price lunch = 16.8%; 1st quartile = 50.1%) versus the population (min = 0%; 1st quartile = 35.1%). Subpopulation 1 therefore excluded schools in the original population from larger school districts (i.e., >12,350 students) or with wealthier students (i.e., <16.8% free or reduced-price lunch). This increased the generalizability index to 0.81, representing about 42% of the schools in the original user population.

The second subpopulation was defined to maximize the generalizability index (B = 0.87), resulting in a subpopulation with high to very high generalizability. This subpopulation definition differed from the original population in relation to three variables: district size, district income, and school racial composition. Again, schools in the experiment were from smaller districts than in the original population. The experimental schools were also from poorer districts (mean median income = $43,810; max = $60,290) versus the original population (mean = $53,610; max = $229,700). Finally, schools in the experiment typically had fewer Hispanic students (mean = 13.9%, max = 59.5%) than the population (mean = 28.2%; max = 100%). This subpopulation was thus defined to only include schools in the original OCR population from smaller districts (i.e., <12,350 students), with smaller proportions of Hispanic students (i.e., <59.5%), and lower median incomes (i.e., <$60,290). This subpopulation was more restrictive and accounted for only 29% of the original population of schools using OCR.

As Table 3 and Figure 2 illustrate, the experimental sample is much more similar to both subpopulation 1 and 2 than the original user population. Here subpopulation 2 is best, with only two covariates being statistically significant (versus nine in the original population), and with improved average ASMDs (mean = 0.16; max = 0.48). Here the main difference between subpopulation 1 and 2 is with respect to racial composition. Also indicated in Figure 2, poststratification and inverse-propensity weighting do not do a good job here further reducing these biases. We therefore conclude that the best case for generalization in the OCR study is to use the usual sample treatment effect estimate (with district fixed effects), and that this estimates the average treatment effect for subpopulation 2.

EM Study

In comparison to the OCR study, making generalizations from the EM study is more difficult. When comparing the final study sample to the originally defined population of EM users, we find many large differences. Here the generalizability index is lower (B = 0.57), and the covariate ASMDs are larger (mean = 0.70, max = 1.32), with nearly all of them being statistically significant. Like the OCR study, the largest differences here pertain to demographics and measures of inequality. Here, however, schools in the experiment are *less* White and more African American than the original population, and are in districts with higher poverty rates and lower median incomes and education levels.

As in the OCR study, the first subpopulation was defined to maximize the population size for a fixed generalizability index (here B > 0.75). This resulted in two variables: school size

and adult education levels. Like the OCR study, schools in the experiment were from districts that were smaller on average (i.e., mean number of students = 9,789; max = 14,770) than those in the originally defined population of schools using EM (mean = 29,530; max = 288,300). Additionally, schools in the experiment were from districts with fewer students with low education levels (min percent less than 9th grade = 5.4%; mean = 7.7%) versus the original population (min = 0%; mean = 5.7%). Subpopulation 1 was therefore defined to include only schools in districts using EM that were not large (i.e., <14,770 students) and with lower education levels (i.e., >5.4% with less than 9th-grade education). This resulted in a generalizability index of 0.76, which represented only about 18% of the original population.

The second subpopulation was defined so as to maximize the generalizability index; this resulted in a generalizability index of 0.81, which indicates high generalizability. In addition to the previous two variables used to define subpopulation 1, this subpopulation focused on a third difference: the proportion of students that were English Language Learners (ELL). Schools in the experiment typically included fewer ELL students (mean % ELL = 2.7%; max = 17.2%) than those in the population of schools using the EM program (mean = 9.0%; max = 85.1%). Subpopulation 2 was thus defined as those schools in districts using EM that were not large (i.e., <14,770 students), with lower education levels (i.e., >5.4% with less than 9th-grade education), and with fewer ELL students (i.e., <17.2% ELL). This subpopulation was notably smaller, representing only 14% of the original population.

As Table 3 and Figure 2 illustrate, like in the OCR case, the experimental sample is much more similar to both subpopulations than to the original EM user population. Of the two, the experimental sample is most similar to subpopulation 2, and these degrees of similarity are best when a poststratification estimator is used. This is both in terms of ASMDs (mean = 0.26; max = 0.54) and the number that are statistically significant (7 versus 16). We therefore conclude that the best case for generalization here is to use a poststratified sample treatment effect estimate (with district fixed effects), and that this estimates the average treatment effect for subpopulation 2. Importantly, however, while poststratification will improve balance here, it will also likely result in an estimate with standard errors that are up to about 49% larger (i.e., $\sqrt{\text{VIF}} = \sqrt{\Sigma w_{pj}^2/w_{sj}} = 1.49$; see Tipton, 2013).

Conclusion

In this section we provided an inclusion/exclusion criteria-based method that allows for researchers to define subpopulations where results of a study may best apply. The method is based upon an originally defined inference population and can include a large set of potential covariates. Importantly, the process can be partially automated. Readers interested in applying this to their own analyses should contact the lead study author.

Discussion: Lessons for the Future

In the previous two sections, we offered new approaches to evaluating the generalizability of results from large-scale experiments. The first approach focused on better understanding the similarities and differences between the types of school districts that agree to take part in large-scale experiments (and those that do not), while the second section provided a novel approach to assessing and determining where the results of a study might actually apply (when the results do not generalize well to the originally defined inference population). Throughout we have situated this discussion in terms of two IES scale-up studies, each with

a prespecified inference population. In this section, we distill our findings into lessons for future studies and for future research.

Lesson 1: Preregistering Your Inference Population

The analyses conducted here were enabled by the fact that inference populations of interest were defined at the study outset. The original selection of covariates, while not discussed in full here, resulted from a long discussion regarding understandings of how the program worked, features regarding implementation, and prior research on treatment effect heterogeneity. By having this conversation up front, information on the districts that were contacted, including feedback from these districts, could be tracked, and features of the districts compared using publicly available data. Additionally, the assessment of generalizability could be conducted based on this prespecified inference population and covariate set, with a focus on finding possible subsets where the results generalize well.

This focus on prespecified population definitions and covariate specifications is similar to the focus on prespecified analyses in experiments and in meta-analysis (see e.g., the Campbell Collaboration; http://www.campbellcollaboration.org/), with the focus here on reducing bias resulting from poststudy knowledge of recruitment difficulties. For example, this bias could arise if at the study design phase researchers felt that it was important to stratify on school size, but noticing that only large schools took part in the experiment, decided post hoc that school size wasn't important for generalization after all.

Although focusing on the prespecified population reduces this "researcher" bias, it can also limit the ability of the research team to implement knowledge gained in the study itself in the analysis. For example, during the course of the experiment it may become clear that there are particular school features that are necessary for the program to be implemented well, and these features may not have been included in the original inference population definition. The solution here is similar to that of outcome analyses in experiments—to focus on the original inference population as the "confirmatory" (or primary) analysis, and to consider other possible inference populations as "exploratory."

Lesson 2: Development of Incentives

Recruitment was much more difficult in these scale-up studies than initially anticipated. The original study design assumed that a cross-over design (which was economical as well) would incentivize involvement because every school taking part would receive a program. However, in practice this was not always correct because many districts were in need of only one core program, not two. Furthermore, some districts did not like the idea of school-level randomization, which resulted in different programs in different schools within a district (leading to difficulties with teacher development and problems for students switching schools).

An additional unanticipated problem was that simultaneous to these studies, the Common Core State Standards (CCSS) implementation and testing began. This meant that many districts were focusing their limited resources on CCSS and were not interested in adding to this uncertainty by implementing a new curriculum or experiment. Finally, as the implementation analysis shows, districts that actively said "no" tended to be wealthier than those that said "yes"; this trend continued in the school-level generalization analysis, making

it difficult (or impossible) for the results to generalize well to the original target populations. This suggests that to the degree that the original study design did incentivize some districts to take part, it did so differentially, particularly with respect to measures of wealth (i.e., median income, free or reduced-price lunch rates).

Altogether this suggests that there is work to be done in understanding what incentivizes districts and schools to take part in studies, how to increase this participation, and how incentive structures can be developed that are targeted to different subpopulations. What incentivizes a "rich" district to take part in a study may be different than a "poor" one, for example, or an urban versus a rural district.

Lesson 3: Recruitment Planning

The IES grant process, as it currently stands, does not allow time or money to be allocated to recruitment. In many instances letters of support are required at the time of application, indicating that recruitment is complete. However, work by Spybrook (2013) indicates that very often recruitment continues after a grant has been disbursed, and can often continue long past the anticipated due date. As this article shows, recruitment in these cases can easily become ad hoc, based on the single-minded goal of achieving the required sample for power, with less of a focus on generalizability.

One lesson from this study is that considerable work needs to go into developing recruitment plans for a study, including being built into project budgets and timelines. When not written in, researchers often fall back on the easiest-to-implement and least expensive options—including postcards, e-mail blasts, and unsolicited phone calls—and as this article shows, these tools are often not that effective in generating a sample. A better recruitment plan might involve some effort to think, in advance (at the study design phase), about how to best reach and motivate district or school officials making decisions, including help with entrée from a research organization, a publisher partnership, or a letter directly from the funder. Additionally, it may be that more costly but targeted recruitment could be more effective, for example, focusing intensely on the development of relationships with a handful of targeted, representative districts instead of reaching out more broadly.

Lesson 4: Narrow Generalizations

The second section of this article focused on comparisons between the originally defined inference populations and samples in the experiments. As the analyses indicate, the resulting experimental samples differed in large and marked ways from these populations, making inferences to the populations of schools in districts using EM and OCR impossible. The analysis suggests that in each case inferences may be made to particular subpopulations of schools using the programs. In the OCR study, the experimental sample of schools was found to be most similar to subpopulation 2, which included a little less than one third of the population of schools using OCR. In the EM study, however, this inference population was much smaller, with the experimental sample representing an even smaller fraction of schools (14%) in the population of schools using EM. In this case, however, the best estimator also required poststratification, which will likely lead to precision losses.

A difficulty with this approach to generalization is to explain clearly how to conceptualize these subpopulations. Certainly geographic conceptualizations of populations are simplest;

for example, it would be much easier to refer to the results generalizing to a particular state or geographic region than to the subpopulations developed here. However, researchers in both the probability sampling and research synthesis communities commonly use conceptualizations of this type when describing target populations. Although not as simple to visualize, we argue that this inclusion/exclusion criteria approach may be most useful in many generalizations from experiments. For example, here the average treatment effect calculated in the OCR study can be seen as generalizing to the type of schools that typically use OCR but that are in small to medium-sized districts with lower median family incomes and smaller Hispanic student enrollments. Similarly, the poststratified average treatment impact estimate in the EM study generalizes to the type of schools typically using EM but that are in small to medium-sized districts with lower education levels and with only a small number of English Language Learners. When interpreting results, researchers and policymakers must then simply ask if their school (or group of schools) meets these criteria or not.

Lesson 5: District Size

Interestingly, throughout our analyses of both recruitment trends and the assessment of generalization, the role of district size arose again and again. In the recruitment analysis, we realized that over time, recruiters began to focus on larger districts because they brought with them a larger number of schools. In the generalization analyses, unexpectedly this variable became an important way to define subpopulations for generalization. Although this variable was not originally expected to affect average treatment effects, it was correlated with variables that were, and thus by excluding districts based on size the similarity between the sample and population improved. Importantly, the finding here was that the districts that needed to be excluded were *very large* districts, meaning that although recruiters had moved toward recruiting *larger* districts, those recruited still were not the *largest*.

This finding is important and suggests that district size may be thought of as a proxy for other school and district features that affect treatment effects and implementation, and should therefore be included in future stratification plans for generalization. Importantly, it does not mean that researchers should focus *only* on large districts, but instead that an awareness of the *distribution* of district size in the intended population should be known and efforts to recruit broadly across these sizes are important. Alternatively, if such diversity is deemed impossible (for budgetary reasons), then the inference population can be defined clearly in relation to district size from the beginning of the study. Either way, it is a variable that researchers should think about carefully, both in planning for recruitment and for generalization.

Conclusion

Throughout this article, we have argued that evaluating the recruitment process and conducting a generalizability analysis are important processes central to increasing the relevance of evaluations to practitioners and researchers. To our knowledge, this is the first article to focus entirely on these issues, including a full discussion of potential bias resulting from refusals and from differences between the resulting sample and populations. By keeping good recruitment records and "auditing" these records, we've been able to better understand where differential refusals occur and what the implications of these may be for practice.

Finally, by sharing the story of recruitment, including a CONSORT-type figure regarding the recruitment process, as well as a refusal analysis, we hope to start a conversation regarding effective strategies for recruitment, the development of incentives, and the best approaches for improving generalization outcomes in future evaluations. We see this work as being just as central to the evaluation literature as work on attrition bias, noncompliance, and implementation—all areas of research with large and rich literatures. Furthermore, by providing a discussion of generalizability—generated directly by the researchers themselves, based on prespecified covariates—we hope to offer evaluators a direct method for providing feedback to practitioners and policymakers regarding where the causal results of their studies may apply, and where future work needs to be conducted.

Funding

This work was supported by the National Science Foundation Directorate for Education and Human Resources [Grant number 1118978] and the Institute of Education Sciences [Grant numbers R305A1001150 and R305A100116].

ARTICLE HISTORY

Received 3 June 2015
Revised 25 September 2015
Accepted 2 October 2015

EDITORS

This article was reviewed and accepted under the editorship of Carol McDonald Connor and Spyros Konstantopoulos.

References

Borman, G. D., Dowling, N. M., & Schneck, C. (2008). A multi-site cluster randomized field trial of Open Court Reading. *Educational Evaluation and Policy Analysis*, *30*, 389–407.

Egger, M., Juni, P., Bartlett, C. (2001). Value of flow diagrams in reports of randomized controlled trials. *JAMA*, *285*(15), 1996–1999.

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

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., & the PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and Meta-Analyses: The PRISMA statement. *PLoS Med*, *6*(6), e1000097. doi:10.1371/journal.pmed1000097

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*, 107–121.

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

Roschelle, J., Feng, M., Gallagher, H. A., Murphy, R., Harris, C., Kamdar, D., & Trinidad, G. (2014). *Recruiting participants for large-scale random assignment experiments in school settings*. Menlo Park, CA: SRI International.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Slavin, R. E., & Lake, C. (2007). *Effective programs in elementary mathematics: A best- evidence synthesis*. Baltimore, MD: Johns Hopkins University.

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), 396–442.

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 subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*, 239–266.

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

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

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2015). *Implications of small samples for generalization: Adjustments and rules of thumb* (Working Paper). Retrieved from http://blogs.cuit.columbia.edu/let2119/files/2013/09/Tipton-Hallberg-Hedges-Chan-Jan-2016.pdf

Tipton, E., Hedges, L. V., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114–135.