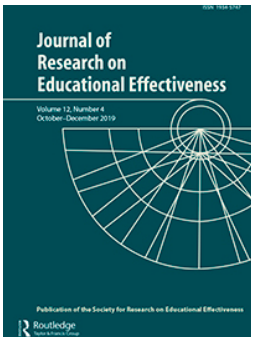

Journal of Research on Educational Effectiveness

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

Combining Archival Data and Program-Generated Electronic Records to Improve the Usefulness of Efficacy Trials in Education: General Considerations and an Empirical Example

Mark C. White, Brian Rowan, Ben Hansen & Timothy Lycurgus

To cite this article: Mark C. White, Brian Rowan, Ben Hansen & Timothy Lycurgus (2019) Combining Archival Data and Program-Generated Electronic Records to Improve the Usefulness of Efficacy Trials in Education: General Considerations and an Empirical Example, Journal of Research on Educational Effectiveness, 12:4, 659-684

To link to this article: https://doi.org/10.1080/19345747.2019.1636438

Published online: 06 Dec 2019.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2019, VOL. 12, NO. 4, 659–684
https://doi.org/10.1080/19345747.2019.1636438

METHODOLOGICAL STUDIES

Combining Archival Data and Program-Generated Electronic Records to Improve the Usefulness of Efficacy Trials in Education: General Considerations and an Empirical Example

Mark C. White[a], Brian Rowan[a], Ben Hansen[a] and Timothy Lycurgus[a]

ABSTRACT
There is growing pressure to make efficacy experiments more useful. This requires attending to the twin goals of generalizing experimental results to those schools that will use the results and testing the intervention's theory of action. We show how electronic records, created naturally during the daily operation of technology-based interventions, contain the information needed to attend to these twin goals. These records allow researchers to define the population of schools considering adoption of an intervention and to plan an experiment to generalize to these schools. They also allow researchers to identify schools likely to fully implement the intervention, such that the theory of action can be properly tested. Designing experiments to address these goals involves many tradeoffs and prioritizing the different purposes of the planned experiment. We discuss these challenges, linking experimental purposes with design decisions.

KEYWORDS
efficacy trial
generalization
program evaluation
experimental design

A growing literature is raising concerns about the external validity of large-scale education experiments. At issue is the extent to which the impact estimates from a given experiment actually generalize to the specific, important populations of interest. We know that impact estimates can vary markedly across schools and that the samples of schools and districts enrolled in experiments often differ in important ways from the populations of schools and districts where the program under study will ultimately be adopted (Bell, Olsen, Orr, & Stuart, 2016; Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017; Tipton et al., 2016). These findings are leading to the development of new approaches to experimental design and analysis involving prospective sampling strategies that help researchers recruit experimental samples that more closely resemble important user populations (Tipton et al., 2014) and/or that use post hoc adjustments to experimental results to target impact estimates for different user populations (e.g., Stuart, Bradshaw, & Leaf, 2015). At the core of these efforts is the desire to estimate average causal effects for specific and clearly defined target populations, thereby attending to the problems of external validity present in experimental research in education.

CONTACT Mark C. White mrkwht@umich.edu Institute of Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48104, USA.
[a]Institute of Social Research, University of Michigan, Ann Arbor, Michigan, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree.

This article extends the work of Tipton et al. (2014) on the use of propensity scores to recruit an experimental sample that is broadly representative of a specific target population in the context of an effectiveness/scale-up trial. We focus on the unique challenges of efficacy studies, showing in particular how to use data routinely generated by education intervention programs to address these challenges. Our central argument is that efficacy trials have two goals (only one of which has been addressed in previous work on the external validity of education experiments). First, like scale-up (or effectiveness) trials, efficacy studies have the practical goal of estimating the causal impact of a program in schools likely to adopt that program, a set of schools we refer to as the potential user base (PUB). In addition, however, efficacy trials also have the scientific goal of testing a program's theory of action (Flay, 1986). The first goal comes from the desire to create effect estimates useful to either policy makers or practitioners making a program adoption decision. The second goal addresses the scientific question of understanding different pathways to achieve specific outcomes. At base, this latter goal requires that schools implement the program under study faithfully, for one cannot test the program's theory of action against a counterfactual if the theory of action was never implemented in the experimental group. Building on the literature on improving experimental design through purposeful sampling, we show how the electronic records already generated by many programs can be used to design efficacy studies that both study the program in samples closely resembling the PUB and test a program's theory of action.

This article introduces approaches to sample recruitment leading to impact estimates that can be generalized both to the population of schools likely to adopt the program under study (i.e., the PUB) and to the population of schools that are likely to implement the program with fidelity. The fundamental challenge to overcome is that these populations are unknown, at least in the general case. However, many programs (and especially technology-based programs) routinely, in the course of daily operation, generate data that would allow researchers to estimate whether schools belong to these populations. In this article, we refer to such data as "electronic records" and we discuss the variety of forms these records might take and how they might be used for recruiting an experimental sample. As we show in the article, electronic records allow us to estimate which schools are likely in the PUB and which are likely to implement the program with fidelity, allowing researchers to target these schools in the sampling phase of an experiment.

We organize our presentation as follows. Section 1 presents the organizing framework, describing ways to use electronic records to define target populations and recruit experimental samples representative of these populations. Section 2 describes a particular program (known as Burst©:Reading) that serves as the focus of our empirical case study of these methods. Section 3 follows the structure of Section 1, providing an empirical case study with a simulation to explore the impact of different experimental design choices. The article concludes with a set of comments about the challenges and possibilities of the approaches discussed here for designing more useful experiments.

Organizing Framework

This section presents the organizing framework of the article, shown in Figure 1. In Figure 1, parallelograms are used to represent data sets and rectangles are used to

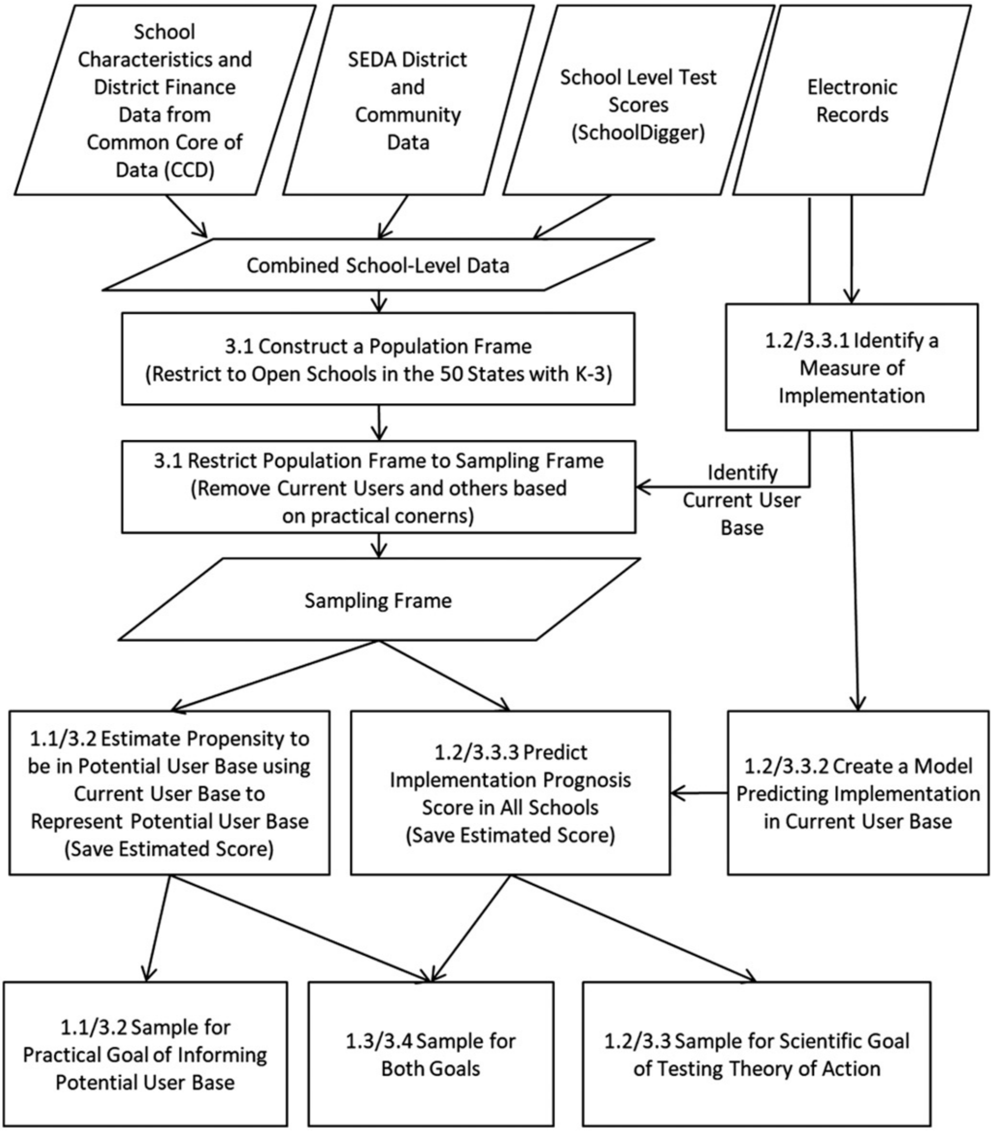
School Characteristics and District Finance Data from Common Core of Data (CCD)
SEDA District and Community Data
School Level Test Scores (SchoolDigger)
Electronic Records
Combined School-Level Data
3.1 Construct a Population Frame (Restrict to Open Schools in the 50 States with K-3)
1.2/3.3.1 Identify a Measure of Implementation
3.1 Restrict Population Frame to Sampling Frame (Remove Current Users and others based on practical conerns)
Identify Current User Base
Sampling Frame
1.1/3.2 Estimate Propensity to be in Potential User Base using Current User Base to Represent Potential User Base (Save Estimated Score)
1.2/3.3.3 Predict Implementation Prognosis Score in All Schools (Save Estimated Score)
1.2/3.3.2 Create a Model Predicting Implementation in Current User Base
1.1/3.2 Sample for Practical Goal of Informing Potential User Base
1.3/3.4 Sample for Both Goals
1.2/3.3 Sample for Scientific Goal of Testing Theory of Action

Figure 1. Overview of the process to move from electronic records to an efficacy trial sampling plan.

represent processing steps. Processing steps are labeled with section numbers to connect to text. Figure 1 starts by constructing a sampling frame by combining administrative data on all schools to create a population frame and restricting this population frame based on the nature of the program under study and practical concerns. This results in a sampling frame, which is the set of all schools from which the experimental sample could be recruited. There are two pathways for considering recruitment based on the practical and scientific goals of efficacy trials. The first (the left branch under sampling frame; Section 1.1) involves estimating schools likely to be in the PUB and sampling so that the experimental sample will broadly represent the PUB. The second (the right branch under sampling frame; Section 1.2) involves predicting which schools are likely to conduct the program with fidelity and recruiting an experimental sample of high-fidelity implementers. We discuss each of these in turn before discussing considerations for combining these goals (Section 1.3).

Sampling for the PUB

Defining the PUB

Given the cost of experiments and relative dearth of effectiveness trials (Institute of Education Sciences, 2016), education efficacy trials are expected to take up the practical goal of estimating the effect of a program for schools likely to adopt the program (i.e., the PUB). In some cases, this is a very specific population, such as when policy makers are considering legislation, which might encourage or require a clearly defined population to use a program. For example, Reading First funding was targeted to school systems with more than 15% (or at least 6,500) students living below the poverty line, creating a clear PUB for programs that might be supported by Reading First funding. In cases such as this, the PUB is clearly defined and a researcher can use the approaches discussed in this article with this prespecified PUB.

In most cases, however, the PUB is not well defined since programs diffuse through the educational system through market-driven processes. In fact, because requiring schools to adopt specific programs is very rare, even when a policy is enacted (e.g., funding sources such as Reading First or Title I), the policy's impact occurs through market forces, as multiple programs compete to take advantage of the new policies and are differentially successful within specific subpopulations of the broader population targeted by the new policy. Fundamentally, schools have many choices, which both makes the PUB uncertain and raises the importance of generating useful evidence on program effectiveness to support schools' choices (Slavin, 2017).

Generally speaking, programs diffuse through homogeneous networks with infrequent jumps between such networks for at least three reasons (Rogers, 2003). First, practitioners prefer programs that have evidence of effectiveness (either research-based or anecdotal) in contexts similar to their own (Nelson, Leffler, & Hansen, 2009; Slavin, 2017). Second, similar schools have similar needs and so adopt similar programs. Third, marketing strategies initiated by program providers will tend to target (and be successful with) specific types of schools because of the specific features or benefits emphasized. Each of these forces creates the tendency for the PUB to look like the current user base. All of this suggests an important point: that the set of current users of a program should broadly reflect the PUB, and as a result, we argue that a program's current user base can be used as a synthetic population to represent the PUB.[1]

Electronic records, because they store information generated during the usage of a program, provide a source to identify current users of a program. Note that current users may be a subset of the schools that have recently purchased the program, as

[1] Our target population is the unknown potential users, not the current users as Tipton et al. (2014), a subtle difference which motivates some analytic choices as we discuss. When the target population is current users, we argue for quasi-experimental approaches due to (a) their greater external validity, (b) their cheaper cost, and (c) changes to program implementation that may occur over time.

program implementation is often low (Fixsen et al., 2005) and some evidence suggests that most purchased student licenses never get activated (CLEVER.com, 2018). Electronic records should also be at the school level, whereas purchasing data are often at the district level. When possible, researchers should compare purchasing data and electronic records to identify subscriber schools that fail to use the purchased program. When this occurs, researchers face a choice of how to address these schools. In some cases, a careful analysis of the electronic records (and comparison of these records to purchasing data) might suggest that a program is not used consistently enough across schools to justify an efficacy trial.

There are a few predictable cases where the PUB differs from the current user base such that the current user base will not represent the PUB. First, the earliest adopters of a program (often termed innovators) tend to be unique and unlike later adopters (e.g., they are more willing to adopt untested programs; Rogers, 2003). Thus, until a program has had some time to spread, the current user base might not reflect the PUB. Second, there can be exogenous shocks to the diffusion process. For example, a shift in marketing strategy, the availability of new funding sources, or prominent news stories on a program may lead new populations of schools, who are unlike the current user base, to begin adopting the program.[2]

Electronic records provide information needed to explore this concern, as they often contain information on when schools began using a program (i.e., as indexed by when they first appear in the data). Using such information, researchers can empirically test whether programs are being adopted by homogeneous populations by comparing the set of schools adopting the program in the last year or two to schools that adopted the program longer ago. When these groups of schools are similar, there is evidence to support the assumption that the PUB will reflect the current user base, suggesting benefit in using the current user base to define the PUB. When these groups of schools are not similar across time, an exogenous shock may have affected the diffusion process. In consultation with program developers, researchers can try to identify this shock. If the shock is identifiable (e.g., a marketing push) and unlikely to shift the PUB again, researchers can focus on a subset of current users as the PUB. Otherwise, researchers might decide to use a theoretically or practically defined population as the PUB (e.g., all Title I schools if Title I monies are often used to fund the program).

Sampling From a Defined PUB

Sampling schools to participate in an experiment can be a challenge. Research generally suggests that probability sampling is nearly impossible in experiments due to the challenge of recruiting districts (Stuart et al., 2017; Tipton et al., 2014). Recommendations, then, focus on stratifying schools in a sampling frame and sampling from within each stratum such that the percentage of schools in the experimental sample in each strata and the percentage of schools in the PUB in each strata are equal (Tipton et al., 2014). This ensures that the experimental sample broadly reflects the PUB. In this section, we focus on the case where the PUB is unclear, as we believe this case to be the most broadly relevant. When the PUB is

[2] The very results of the experiment may thus disrupt the diffusion of a program, but given the weak role that experimental evidence plays in adoption decisions (Nelson et al., 2009), this seems unlikely in most cases.

prespecified (e.g., all Title I schools), sampling can proceed as described in Tipton (2013). However, even when the PUB is prespecified, the approach described here can be used by first reducing the sampling frame to the prespecified PUB and then following the recommendations below. This leads to an experiment that is targeted to schools likely to adopt the program within the broader population of interest.

This section follows directly the recommendations of Tipton et al. (2014) with the PUB as the inference (or target) population. Due to space, we only outline the process here focusing on complexities arising from estimating the PUB, referring interested readers to Tipton et al. (2014) for more details.

The first step is to create a sampling frame of all schools eligible to be part of the experiment. This generally involves restricting a population frame, such as the Common Core of Data (CCD), based on practical concerns, such as school size, school location, or previous experience with the program. This will almost always result in removal of the current user base from the sampling frame. Next, the researcher uses a propensity score model to estimate the likelihood that a school is in the PUB. Based on the previous recommendations, this model would estimate the likelihood of being in the current user base, treating estimated propensities as the schools' propensity to be in the PUB. Balancing schools in the experimental sample and the current user base on this propensity score will create an experimental sample that broadly represents the PUB (on observables) when potential users look like current users. Balance is created by stratifying the current user base into S equally sized strata (usually five). The cutoffs in the propensity score between strata are used to stratify the sampling frame into the same S strata. An equal number of schools are then sampled from each stratum, starting with schools whose propensity score is closest to the average propensity score in their strata. This will result in an experimental sample that is balanced on the propensity score with the current user base and, since the propensity score is a balancing score, a sample that is broadly balanced on all covariates used to generate the propensity score.

The nature of the estimated propensity score deserves additional considerations. The spread in propensity score between the current user base and the sampling frame (i.e., the difference in distributions of the two groups) is a model-based estimate of the degree to which the PUB is unique from the broader sampling frame. As the spread grows, the propensity model is estimating a PUB that is unique from the typical sampling frame school. In the extreme, the propensity model suggests all PUB schools have exactly the same covariate values as a current user base school. There is a role, then, for substantive knowledge to guide the construction of the propensity score model, especially since the true PUB is unknown. When the characteristics of a program or the history of program adoption in electronic records are such that the program is likely to diffuse through a unique and homogeneous group of schools (e.g., the program is designed for a narrow group of students, such as those with autism), a large spread in the propensity score is reasonable. In this case, the experimental sample will target a narrow subset of schools in the sampling frame (and will be representative only of this subset). If substantively or theoretically justified, this is appropriate, although it may lead to the practical issue of too few sampling frame schools within a stratum to meet the sampling quota (i.e., there are no sampling frame schools representing some portion of the PUB). This practical challenge suggests that the program has fully diffused

Table 1. Difference in observed characteristics between new and old BURST schools.

	Sampling Frame	2010–12 BURST Adopters	2015–17 BURST Adopter	Standardized Difference Old–New Adopter
County socioeconomic status	0.01	−0.42	−0.29	0.143
N students in district	16,540	116,952	41,105	−0.722
Segregation across FRL status	0.01	0.62	0.56	−0.051
District yearly cohort test score growth (in grade units)	0.96	0.96	0.99	0.247
Total district expenditures	$12,788	$13,190	$12,181	−0.236
Magnet school	3%	5%	3%	−0.116
Charter school	7%	0%	12%	0.503
School in city	29%	46%	42%	−0.077
School in rural	26%	28%	25%	−0.063
School in suburb	33%	22%	20%	−0.043
School in town	12%	5%	13%	0.292
N students	472	627	471	−0.59
Percent free-lunch students	56%	68%	61%	−0.238
Percent Hispanic	24%	37%	25%	−0.381
Percent African American	15%	22%	29%	0.244
Achievement index (3rd grade)	0.01	−0.29	−0.14	0.152
Lagged achievement index (3rd grade)	0.01	−0.15	−0.11	0.045

Note. Achievement index is the average of the percentage of students proficient in math in 3rd grade and the percentage of students proficient in English in 3rd grade, after standardizing these variables within state and year. Standardized differences were generated by RItools (Bowers, Fredrickson, & Hansen, 2010).
FRL = free/reduced price lunch.

through some niche of schools or that the restrictions used to form the sampling frame eliminated schools with some set of characteristics.

In most cases, though, substantive and practical considerations will lead researchers to prefer a propensity score that has a fairly narrow spread. Substantively, this indicates uncertainty in estimating the PUB and/or the belief that the program will diffuse relatively broadly throughout the sampling frame. Practically, this leads to a more diverse experimental sample that can be generalized to a wider range of target populations using post hoc approaches. To narrow the spread, researchers can regularize the propensity score by, for example, removing less important covariates, removing interactions, and/or imposing more linearity assumptions (see Section 3.2. for an empirical demonstration of how regularization distributes the sampling frame more broadly across the sampling strata). When regularizing, the researcher should retain covariates theorized to predict both the likelihood of adopting the program and treatment effect heterogeneity as balancing the sample and target population on these variables is most important for removing bias (Tipton, 2013). This regularization is justified because the current user base, which is used to fit the propensity score, is not the inference population, but rather the inference population is the PUB, which should “look like” the current user base due to homogeneous diffusion processes. Here regularization is used to relax how similar schools must be to “look like” each other. In simulations below, we show how regularizing (modestly) increases the ability to generalize to broader populations while (modestly) decreasing the ability to generalize to the current user base.

Sampling Schools Likely to Implement the Program With Fidelity

Efficacy trials seek to test the theory of action of specific programs (Flay, 1986). Since programs are generally studied in actual schools, this rarely involves studying programs in truly ideal conditions, but it does mean that careful attention should be paid to ensure high levels of program implementation. Implementation is a multifaceted problem and needs to be supported through multiple pathways (e.g., see Fixsen et al., 2005). One pathway is by recruiting schools that are likely to use the program with fidelity. Data released by Clever.com, which supports schools in using technology-based programs, found that up to 70% of the student licenses in their system were never activated (but were paid for; CLEVER.com, 2018). This suggests that recruiting schools in the PUB will not necessarily recruit schools likely to use a program. Given the long track record of generally observed poor implementation (e.g., Fixsen et al., 2005), focusing experimental recruitment on schools likely to implement a program with fidelity should improve the ability for an efficacy study to test a program's theory of action.

The problem is that it is impossible to know with certainty in advance of an experiment whether any given school will implement a program at high levels of fidelity. We can, however, hypothesize that at least some observable features of schools will be associated with a school's capacity and/or inclination to implement a program. Electronic records provide an opportunity to test this hypothesis. Electronic records capture logins, screen clicks, videos watched, and/or other indicators that capture how practitioners interact with a program. These can be reduced into a measure of program implementation and a prognostic-style score that estimates expected program usage in a school can be created (e.g., Hansen, 2008). We call this estimate a school's *implementation prognosis score* (IPS).

Using Electronic Records to Capture Implementation

Since all electronic records are likely to be different, specific guidance for how to craft a measure of implementation is difficult to provide. Online courses might calculate average course completion rates, while learning management systems might calculate the percentage of teachers regularly logging into the system. The goal here is to use the program's theory of action to identify a measure in the pathway from program adoption to program impacts (i.e., an important mediator). Electronic records can then be used to estimate this implementation metric in the current user base. If the electronic records contain an outcome measure or if previous research on a program exists, researchers can empirically examine whether the implementation metric actually seems to function as a mediator. In any case, schools high on this implementation metric are enacting the theory of action contained within the program. Schools similar to these schools, then, may also be likely to enact the program's theory of action.

Beyond using this implementation metric for sampling, as discussed next, this metric (along with electronic records more broadly) can support the design of the efficacy trial. The electronic records identify which schools are struggling with implementation and which are successfully implementing a program. Researchers can study these schools to try to understand supports that facilitate implementation. This would require more prestudy planning than most efficacy trials currently support, but could allow researchers to build in additional extra supports within the efficacy trial to ensure that schools are able to successfully implement the program, further supporting the goal of testing the program's theory of action.

Estimating the Schools Likely to Implement the Program With Fidelity

After selecting a measure of implementation, the second step involves modeling our chosen implementation measure using observable variables to construct the IPS. Any predictive model could work here, but the goal is predictive accuracy, so machine-learning approaches are recommended. After building the model in the current user base, the model can be used to predict the IPS for all schools in the sampling frame. The usefulness of the IPS is in direct relation to how well it actually predicts future program implementation. This could be estimated from a statistic such as *r*-squared on a held-out test sample in the current user base or by combining the electronic records with previous experimental evidence on a program. Note that the IPS is likely to be most effective when the experimental sample closely resembles the current user base (as this limits extrapolation) and when experimental incentives and/or supports do not change the nature of how schools use a program.

When observed characteristics do not predict program implementation, there may be little benefit in using the IPS for trial design. In this case, researchers might want to focus only on recruiting from the PUB. It is not clear how strong a prediction of future implementation is necessary to make this approach worthwhile and future research will have to explore this point. However, there should be no harm in this approach (especially when combined with a stratification approach), beyond complicating sampling, unless the relationship between the IPS and actual implementation is negative. Note that a side benefit of this approach is that we have generated a moderator (the IPS) and a mediator (the implementation measure) which can be preregistered as key variables in planned analyses before the start of the experiment.

The easiest way to sample for high implementers would be to define the population of high implementers as those schools in the top nth percentile of the IPS (we use the 75th below) and sample from these schools, using any desired sampling strategy. This would ensure that only schools with high predicted levels of implementation were included in the experiment.

Sampling for both the PUB and Likely High Implementers

To this point, we have separately addressed the goals of designing an experiment to generalize to the PUB (i.e., creating results actionable for practitioners) and to generalize to schools likely to fully implement the program (i.e., testing the program's theory of action). While each approach could be used separately, combining them to address both goals is preferred. When combined, the experiment recruits a population likely to generalize to both the PUB and high implementers. If these populations differ, though, this might involve decreasing the ability of the experiment to generalize to either of these target populations individually.

There are two approaches that could be used here, depending on which of the two goals one wishes to prioritize. To prioritize generalizing to the PUB, the stratification approach discussed previously can be used, and within each stratum, schools can be sampled in decreasing order of the IPS. To prioritize generalizing to high implementers, schools below the nth percentile on the IPS can be removed from the sampling frame and then the stratification approach discussed before can be adopted. Importantly, when the IPS and propensity score are independent, these approaches will generally give similar results.

We might, though, expect a strong relationship between the IPS and the propensity score. This would occur if schools understand their needs and capacities, adopting only programs that they will implement well. In practice, a strong relationship creates a conflict between the two goals of efficacy trials because when there is a strong relationship, sampling across the full range of the propensity score necessarily samples across the range of the IPS. Researchers, then, will face trade-offs in prioritizing either goal. These trade-offs may be made more explicit through the using the simulation approach we demonstrate below. Researchers should be aware of this trade-off and take the time to explore the relationship between the IPS and the propensity score.

For example, focusing on both the PUB and high implementers may result in a specific type of school present in the PUB (e.g., Title I schools) not being sampled, removing the experiment's ability to generalize to this subpopulation (e.g., Title I schools). This would occur when a specific type of school is predicted to be poor implementers of the program and there is a positive relationship between the IPS and propensity score. If generalizing experimental effects to this subpopulation is important, researchers can adapt the sampling plan to include these schools, perhaps by adding a stratum specifically for these schools. They should also, however, consider speaking with program developers or specific schools to understand implementation barriers so that specific implementation supports for these schools can be built into the efficacy study. In fact, this sort of pre-efficacy trial inquiry into implementation challenges in schools with weak implementation in the electronic records seems beneficial for any study.

In summary, when both the propensity to adopt a program and the IPSs are used simultaneously to plan experiments, the experiment is designed neither to address the goal of providing information to schools most likely to adopt the program nor to test the theory of action, but instead balances these two potentially conflicting goals. Experimenters will have to make decisions regarding how much of the PUB can be excluded from the experiment to meet the goal of testing the program's theory of action, adapting the sampling approaches discussed to meet sampling goals of specific studies. In our discussion of the simulation below, we demonstrate one approach of how to consider trade-offs between approaches.

Data and Simulation Approach

BURST Reading Program and Electronic Records

The data in our empirical example come from a research project we conducted with Amplify, Inc. to evaluate the efficacy of the BURST©:Reading program (BURST; recently rebranded as MCLASS Intervention). BURST is a personalized beginning reading program that uses a proprietary algorithm to assign students to small-group, supplementary instruction on the basis of their test scores on the Dynamic Indicators of Basic Early Literacy

Skills (DIBELS; Good & Kaminski, 2002). The program functions ideally as follows. All students test on DIBELS at the beginning, middle, and end of the year. DIBELS scores are used to assign students to initial groups at the beginning and middle of the year, and each group is assigned an initial BURST cycle, a set of 10 lesson plans to be conducted over 2 weeks. At the end of each cycle, a new BURST cycle is assigned to each group. Students should receive 6 or more cycles in both the fall and winter semesters and small groups are reassigned after the second DIBELS testing. BURST is targeted toward struggling students (Tier 2) as a supplemental instructional program and students can exit treatment before receiving all 6 BURST cycles in a semester if they make sufficient progress.

This example below demonstrates the approaches to sampling just discussed, showing how electronic records generated routinely as part of BURST operations can be used at the design stage of an efficacy trial. The full set of electronic records contain DIBELS test scores for all students in all program schools, the grouping assignments of students, and records of assigned BURST cycles for students. However, because of the confidential nature of these data, as well as the proprietary nature of the program's operating algorithm for student assignment to BURST groups, the 2017 electronic records available to us from Amplify, Inc. were limited to include only (a) school average DIBELS test scores at the beginning, middle, and end of the year and (b) two school-level measures of implementation quality: the percentage of students assigned to any BURST cycles in a given semester and the percentage of students assigned to the recommended 6 (or more) BURST cycles. These confidentiality concerns are likely to be common and require researchers to build collaborative, trusting relationships with program developers, a point we return to in the discussion.

Simulation Study

As the theoretical discussion suggests, substantive or empirical considerations may lead to a number of adaptations to the basic strategy laid forth here. In order to help illuminate the implications of different options, we run a brief simulation study. The simulation samples from the sampling frame using the specified strategy under the assumption that 1% of schools will agree to participate (observed recruitment rates for the BURST study were just above 1%). One thousand independent samples were simulated. Criterion for the generalizability of the sample to specific key target populations were Tipton's (2014) B-Index and the average absolute standardized mean difference (SMD) between the sample and target population across covariates. Four target populations were identified: (1) all rural schools, (2) the BURST schools adopting BURST after 2015 (current user base; see below), (3) sampling frame schools with an IPS in the 90th percentile or above, and (4) current BURST users with an IPS in the 90th percentile or above. Specific sampling strategies contrasted are described in the text below. Note that we assume that nonparticipation is random because of the limited evidence to support any hypothesized model, although work that might lead to the creation of such a model is underway (see Tipton, Wang, Spybrook, & Fitzgerald, 2019). Simulations were also run where schools had a likelihood of joining the experiment determined from the recruitment process in the BURST experiment (rather than all schools having 1% chance of being recruited). Results were generally consistent, although all sampling approaches, unsurprisingly, showed strong ability to generalize to inference populations with high average propensity to join the experiment. These results are not favored due to concerns about the stability and accuracy of school-specific recruitment likelihoods.

BURST Case Study

In this section, we provide a demonstration of how to use electronic records to design an efficacy study sampling plan designed to generalize experimental results to both the PUB and schools likely to implement the program at high levels. The flow of this section mirrors that of Section 1 and follows the diagram in Figure 1.

Constructing a Population Frame

As Figure 1 shows, we combined three data sets to construct a population frame of schools that includes school-level information on demographic features, achievement data, district financial information, district testing data, and derived district and community characteristics. First, population-level school data were obtained from the CCD for 2009 through 2016, including data on student composition, school locale, school organization, and district finance data. Second, district test score data and community-level characteristics were obtained from the Stanford Data Education Archives (SEDA; Reardon et al., 2017), including community socioeconomic status (SES), segregation indexes, and district-level average growth in test scores shown by a cohort across grades. Third, the percentage of third graders proficient in math and English were obtained from SchoolDigger.com.[3] These percentages were standardized within state/year to form state/year rankings of schools. Third-grade scores were combined across math and English to form a school achievement index. Because BURST is an early reading program, the population frame was reduced to include only schools that served students in kindergarten, first grade, second grade, or third grade; were in one of the 50 states (or DC); were classified by the CCD as a “regular school” and not closed before 2016; and contained more than 0 students. This resulted in 54,683 schools in the population frame.

Sampling for the PUB

In this section, we provide an example of constructing an experimental sample to generalize to the PUB, including identifying the PUB, constructing a propensity score to estimate the likelihood of being in the PUB, stratifying the sampling frame, and sampling to recruit an experimental sample like the PUB. Electronic records are vital here for defining the PUB and testing for stability in the schools that are adopting BURST across time.

Defining the PUB

As discussed previously, the PUB should reflect the current BURST user base when the program is diffusing through homogeneous networks. The BURST electronic records

[3] SchoolDigger scrapes publicly available state data. We spot-checked test scores and found that CCD variables correlated with the CCD files above 0.995.

Table 2. Stratification of the sampling frame across different propensity score models.

Propensity Score (PS) Stratum	Regularized BART Propensity Model *N* (Pct)	Non-Regularized BART Propensity Model *N* (Pct)	Linear Logistic Propensity Model *N* (Pct)
PS < min (PS) of Current User Base	7297 (13%)	16,343 (30%)	704 (1%)
Lowest Quintile	32,919 (60%)	30,304 (56%)	24,726 (45%)
Second Lowest Quintile	7832 (14%)	6764 (12%)	9843 (18%)
Middle Quintile	5683 (10%)	785 (1%)	10,887 (20%)
Second Highest Quintile	783 (1%)	305 (1%)	6129 (11%)
Highest Quintile	7 (0%)	20 (0%)	2205 (4%)
PS > max (PS) of Current User Base	0 (0%)	0 (0%)	27 (0%)
B-Index	0.74	0.39	0.92

Note. BART means the propensity score was estimated using Bayesian Additive Regression Trees. The B-Index comes from Tipton (2014) and measures overlap in the propensity score between the current user base and schools in the sampling frame. Rows labeled as quintiles are regions of common support in the propensity score while other rows are regions of no common support.

contain all schools using BURST as of 2017 along with the year in which adoption occurred. Thus, we can test whether the demographic characteristics of BURST adopters are changing across time. After consulting with the developers of BURST, we identified a marketing push that began in 2014, coinciding with a large group of new schools adopting BURST. Thus, we chose to focus on whether schools adopting BURST after 2014 (after the marketing push) were similar to those adopting BURST before 2014. If these two "adoption cohorts" are demographically similar, we have some confidence in specifying the PUB using the current BURST user base.

Table 1 shows characteristics of the sampling frame, BURST users, and the difference between earlier and later BURST adoption cohorts, which are extensive. While all BURST adopters are from larger-than-typical districts, schools adopting BURST in 2015 to 2017 are in much smaller districts than earlier adopters. Schools adopting BURST in 2015 to 2017 are also more likely to be charter schools, be in towns, have higher percentages of African American students and lower percentages of Hispanic students, and have slightly less money per pupil. While neither set of BURST adopter is similar to the full sampling frame, the schools that adopted BURST more recently are generally more similar to the sampling frame than earlier adopters.

These differences across adoption cohorts suggest that the kinds of schools adopting BURST have shifted over time, calling into question our ability to predict the PUB with the current user base. At the same time, schools that have adopted BURST are quite unlike the sampling frame as a whole. Thus, we want some way of targeting the sampling, but are not confident in predicting future adopters. Then, in this case, we might consider specifying a theoretical population (e.g., rural schools) and use the approach of Tipton (2013).[4]

For the sake of this demonstration, we choose to estimate a PUB. The shifts in marketing strategies were semipermanent and schools adopting BURST from 2015 to 2017 are stable, which we take as an indication that further exogenous impacts to the diffusion process are unlikely. Thus, we choose to use schools that adopted BURST between

[4] As nonparticipation is assumed random, the simulation finds that the Tipton (2013) approach and the random sampling approach are effectively equal so we show only random sampling.

2015 and 2017 as a synthetic data set to represent the PUB.[5] There is no correct way to specify the PUB, and researchers must make a decision based on their understanding of program diffusion (and in consultation with any marketing department). Due to the greater uncertainty in predicting the PUB caused by the shifting set of schools that adopt BURST, we hedge this decision to focus on later BURST adopters by regularizing the propensity score model. This regularization restricts the extent to which the propensity score estimates the PUB as unique from the sampling frame, in line with our belief that our ability to predict the PUB is weak. While this decreases our ability to generalize to schools similar to later BURST adopters (i.e., the predicted PUB), we are better able to generalize to the PUB if our expectations that the PUB looks like later BURST adopters is incorrect. The simulation shows results from both the regularized and nonregularized propensity score for comparison.

Estimating the Propensity to be in the PUB

At this point, we have defined the PUB in reference to schools adopting BURST after 2015. In this section, we describe estimating the propensity to be in the PUB. We chose to use Bayesian Additive Regression Tree (BART; Kapelner & Bleich, 2013) models to estimate this propensity score in order to capture complex relationships between observed variables and the likelihood of BURST adoption (i.e., nonlinear and containing interactions between covariates). BART models were run using the bartMachine package (Kapelner & Bleich, 2013) using tidyverse (Wickham, 2017) in R (R Core Team, 2018). As just discussed, we regularize (Hastie, Tibshirani, & Friedman, 2008, pp. 223–227) the BART model by setting alpha to 20% and beta to 6 (Chipman, George, & McCulloch, 2010).[6] This sets a prior that discourages complex relationships between predictors and the propensity to be in the PUB. We are not aware of specific methodological guidance on choice of regularization parameters that is applicable to this scenario. Instead, our simulation explores the impact of the regularization.

We further limited the prediction variables to the following: community SES, district size, SEDA poverty segregation index, SEDA average yearly cohort growth score, the school's magnet and charter status, the school's locale (i.e., urbanicity), the total students in a school, the proportion of African American students, the proportion of Hispanic students, a school achievement index (average of math and English percent proficient), a 1-year lagged school achievement index, and per-pupil district expenditures. Data were taken from the year immediately before adoption of BURST for schools using BURST and in 2017 for other schools. Standard BART missing data procedures were allowed to handle the small amounts of missing data (see Kapelner & Bleich, 2013).

[5] We could use a risk set matching approach to match each BURST school with schools in the population frame in their risk set within a specified caliper (see Rosenbaum, 2009). This would arguably create a more accurate propensity score for the PUB by incorporating the time-varying nature of BURST adoption into the analysis. Sampling could then follow the recommendations of Tipton (2013). This approach significantly complicates an approach already more complex than current practice without a clear payoff.

[6] This effectively sets the prior so that trees with depth 0 occur 80% of the time and trees with depth 1 about 20% of the time, limiting the growth of any tree. This restricts the complexity of prediction by limiting the extent to which interactions and nonlinear effects are modeled, unless the data strongly suggest these should occur (Chipman et al., 2010).

Stratifying the Sampling Frame and Sampling

The next step is to stratify the sampling frame based on the propensity score. We do this by equally dividing the current user base into 5 strata. The cut-points between strata on the propensity score are used to stratify the sampling frame into the same 5 strata. Table 2 shows the resulting strata sizes that would result from both the regularized propensity score, the nonregularized BART propensity score, and a logistic propensity score that uses the same covariates. Table 2 divides the sampling frame into 5 strata where there is common support between the sampling frame and the current user base and 2 strata where common support is lacking. As the first row shows, the nonregularized propensity score estimates that 30% of the sampling frame has a propensity score lower than the lowest propensity score of the current user base, whereas only 13% fall in this stratum for the regularized propensity score (see also Figure 2 which also shows strata cut-points for the regularized propensity score). Importantly, schools in this

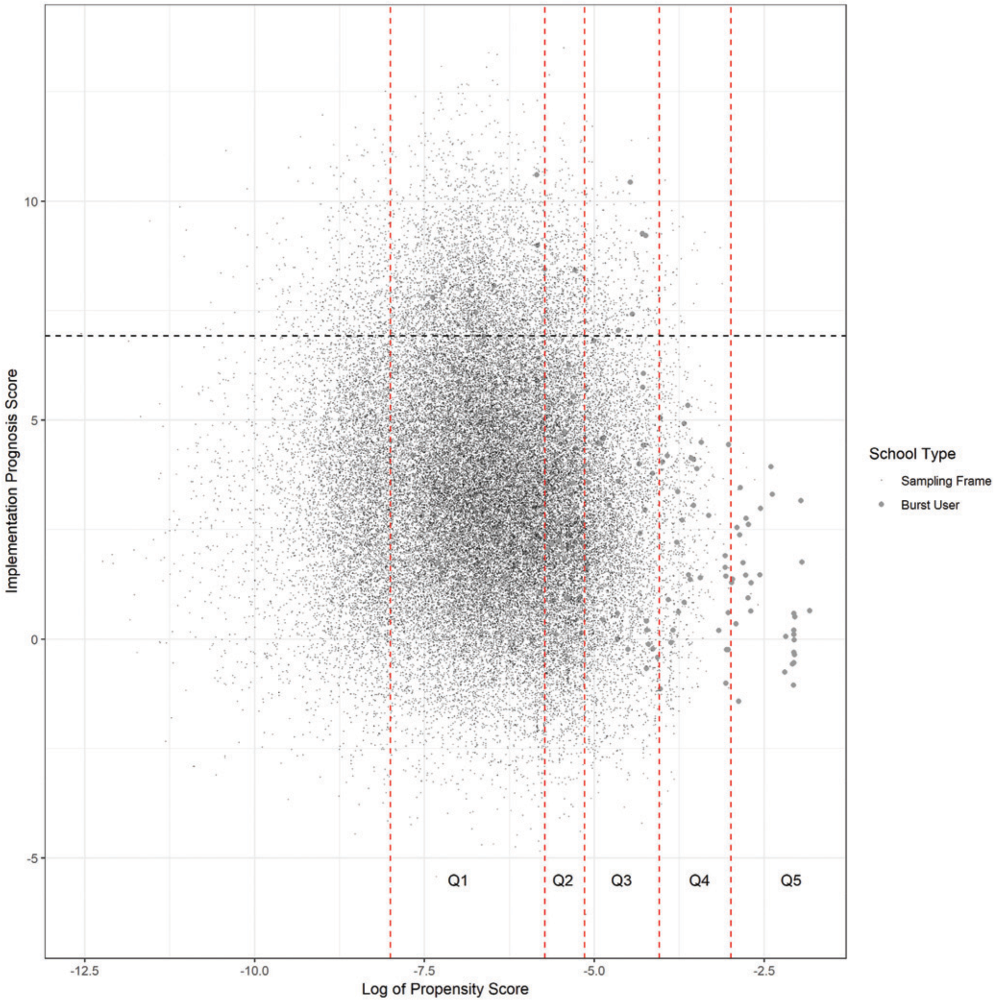

Figure 2. Implementation prognosis score by propensity to adopt BURST.

Table 3. Sampling frequencies across strata and sample.

Stratum	*N* Current BURST Users	Sample Needed	Sampling Frame			High-Prognosis Sampling Frame		
			N	To Recruit	Fraction	*N*	To Recruit	Fraction
0	0	0	7297	0	0%	1800	0	0%
1	33	12	32,919	12	0%	8960	12	0%
2	32	12	7832	12	0%	1625	12	1%
3	33	12	5683	12	0%	1139	12	1%
4	32	12	783	17	2%	122	24	20%
5	32	12	7	7	100%	0	0	–

stratum, conditional on the propensity score model, are not in the region of common support with the PUB (as represented by the current user base). When common support is lacking, the strongly ignorable sample selection assumption is violated and effects may not generalize to the PUB (Tipton, 2014).

Beyond the simulation results we discuss next, this provides a way of thinking about the need to regularize. The propensity score estimated without regularization suggests that about one-third of the sampling frame consists of schools that do not represent the PUB (and would lead to bias if included in the sample), whereas the regularized PUB reports that only 13% of the sampling frame does not represent the PUB. As discussed previously, researchers will have to use their judgement based on their confidence in predicting the PUB to make a decision about which model is most reasonable since the PUB is unknown (and unknowable). In this case, we are unsure enough about the PUB that excluding 30% of the sampling frame seems unwise, so we prefer using the regularized propensity score and sampling from regions of common support. The simulation shows the result of this decision compared to sampling equally across all 6 strata.

Table 3 shows the sampling plan when using the regularized BART propensity score and planning for a sample of 60 schools. The high-prognosis sampling frame shown will be discussed below. While the goal is to sample 12 schools from each of the 5 strata in the region of common support, there are only 7 schools in the highest stratum. We make up for the lack of schools available in this stratum by sampling extra schools from the fourth stratum (see “To recruit” column).[7] While other approaches exist (e.g., combining or reducing stratum), we prefer this simple approach because it does not involve reassigning sampling frame schools to stratum on the fly when problems arise during sample recruitment.

Results From the Simulation Study

Table 4 shows the results of the simulation study. Here, we focus on the first four rows, which show the result for random sampling from the whole sampling frame, PS Stratified–Regularized BART (the recommended approach just discussed), and three deviations from the approach we took. The first deviation is including schools in the region of no common support (stratum 0 in Table 3; row PS Stratified–Include no

[7] It is common when oversampling from a stratum to down-weight estimates to ensure that a stratum is not overrepresented to the sample average. We do not recommend that here as the strata are intended to ensure that the PUB is broadly sampled rather than serving to provide a precise estimate of the PUB. Rather, we recommend using post hoc adjustments (Stuart et al., 2015).

Table 4. Results of the simulation study.

	Tipton's (2014) B-Index				Absolute Standardized Mean Differences			
Sampling Approach	Rural Schools	Current User Base (Predicted PUB)	High IPS	Current User Base and High IPS	Rural Schools	Current User Base (Predicted PUB)	High IPS	Current User Base and High IPS
Random	0.82 (0.75, 0.87)	0.76 (0.70, 0.81)	0.37 (0.28, 0.45)	0.60 (0.52, 0.67)	0.31 (0.25, 0.39)	0.23 (0.17, 0.31)	0.49 (0.41, 0.57)	0.47 (0.40, 0.55)
PS Stratified–Include no common support	0.77 (0.70, 0.83)	0.89 (0.87, 0.91)	0.33 (0.23, 0.41)	0.64 (0.56, 0.70)	0.40 (0.33, 0.49)	0.15 (0.10, 0.22)	0.52 (0.45, 0.60)	0.53 (0.45, 0.61)
PS Stratified–Regularized BART	0.77 (0.71, 0.83)	0.88 (0.84, 0.92)	0.33 (0.23, 0.41)	0.65 (0.58, 0.72)	0.47 (0.39, 0.56)	0.11 (0.08, 0.17)	0.54 (0.47, 0.63)	0.56 (0.48, 0.64)
PS Stratified–Nonregularized BART	0.76 (0.70, 0.82)	0.89 (0.85, 0.92)	0.33 (0.24, 0.41)	0.65 (0.57, 0.72)	0.49 (0.40, 0.58)	0.12 (0.08, 0.17)	0.56 (0.48, 0.65)	0.57 (0.49, 0.66)
PS Stratified–Logit	0.77 (0.70, 0.82)	0.82 (0.78, 0.86)	0.34 (0.25, 0.42)	0.61 (0.52, 0.67)	0.41 (0.35, 0.50)	0.13 (0.09, 0.18)	0.52 (0.45, 0.60)	0.54 (0.47, 0.61)
High Implementation Prognosis	0.98 (0.96, 0.99)	0.72 (0.65, 0.77)	0.98 (0.96, 0.99)	0.78 (0.72, 0.83)	0.34 (0.29, 0.39)	0.59 (0.54, 0.64)	0.11 (0.08, 0.16)	0.30 (0.25, 0.37)
PS Stratified–Include only high IPS	0.95 (0.92, 0.97)	0.79 (0.73, 0.84)	0.70 (0.63, 0.76)	0.75 (0.68, 0.80)	0.31 (0.25, 0.38)	0.45 (0.40, 0.51)	0.21 (0.16, 0.27)	0.32 (0.28, 0.38)
PS Stratified–Sample by high IPS	0.90 (0.86, 0.93)	0.90 (0.87, 0.93)	0.63 (0.58, 0.69)	0.76 (0.70, 0.82)	0.39 (0.33, 0.46)	0.31 (0.25, 0.37)	0.32 (0.27, 0.38)	0.37 (0.31, 0.44)

Note. Cells show either the B-Index or standardized mean difference between the simulated samples and the indicated populations along with the 90% confidence intervals across 1,000 simulations. All simulations assumed 1% recruitment rate with recruitment independent of all school characteristics. IPS is the implementation prognosis score. High IPS schools are in the 90th percentile or higher in the sampling frame. PS Stratified–Include no common support indicates stratifying used the regularized BART propensity score and sampling 10 schools from each of the 6 strata in Table 3. PS Stratified–Regularized BART is the recommended approach and is the sampling strategy shown in the sampling frame columns of Table 3. PS Stratified–Nonregularized BART indicates sampling after stratifying using the nonregularized BART propensity score (see Table 2) and sampling 12 schools from the 5 strata of common support. PS Stratified–Logit indicates sampling after stratifying using the logistic propensity score (see Table 2) and sampling 12 schools from the 5 strata of common support. High IPS indicates sampling schools with the highest implementation prognosis score first. PS Stratified–Include only high IPS indicates sampling using PS Stratified–Regularized BART, but restricting sampling to schools in the 75th percentile or higher of the implementation prognosis school (see high prognosis sampling frame in Table 3). PS Stratified–Sample by high IPS indicates using the PS Stratified strata and recruiting schools with the highest implementation prognosis scores within each stratum.

PUB: predicted user base; IPS: implementation prognosis score; PS: propensity score; BART: Bayesian additive regression.

common support), the second is using the nonregularized propensity score (PS Stratified–Nonregularized BART), and the third is using a logistic propensity score (PS Stratified–Logit), which we view as a more intensive form of regularizing. Cells in Table 4 shows median and 90% confidence intervals for the B-Index (Tipton, 2014) and average absolute SMDs across covariates.

As Table 4 shows, all four stratification approaches lead to higher B-Index values and lower SMD than purely random sampling when the inference population is the predicted PUB, but show lower ability to generalize to both rural schools and high IPS schools. Here, we intentionally use rural schools because being a rural school was relatively independent of being in the PUB, highlighting the trade-off from sampling to generalize solely to the PUB.

Table 4 also highlights some modest differences among the stratification approaches. The logit model and sampling outside the area of common support lead to samples more similar to rural schools and high IPS schools according to the SMD metric, but samples less like the PUB, as we predicted in discussion above. These minor differences are not present in the B-Index estimates, although the logit model does seem to produce samples with lower B-Indexes relative to the current user base than other stratification approaches. Regularizing the BART model leads to very modestly lower SMD with rural schools and high IPS schools as compared to the nonregularized model. That said, all differences between stratified models are likely small enough to be discounted. Future work must explore whether this is a unique feature of this case study. Given these simulation results, we might prefer the logit propensity score given that this propensity score more evenly divided schools across strata (and hence makes meeting sampling quotas easier).

Sampling Schools Likely to Implement the Program With Fidelity

In this section, we demonstrate the steps needed to sample schools likely to implement the program at high levels so as to test the program's theory of action. We start by deciding on a measure of implementation, then show how to estimate an IPS. We then demonstrate approaches to sampling using the IPS, discussing sampling concerns and simulation results.

Identify a Measure of Implementation

The first step is identifying a measure of program implementation using the electronic records. While this may be highly complicated for electronic records that contain large amounts of information, the BURST electronic records contained only two possible implementation measures: the percentage of students receiving any BURST cycles and the percentage receiving the recommended 6 or more cycles per semester. According to the BURST theory of action, students should receive 6 cycles in order to get sufficient support to master their knowledge deficits. Thus, we take the percentage of students receiving the recommended 6 cycles as the implementation metric.

Create a Model Predicting Implementation in Current User Base

After selecting the implementation metric, we create a model to predict this outcome using observed variables. The goal here is to obtain the best predictor so we again use

Table 5. Difference between high and lower implementation prognosis schools.

	Implementation Prognosis Score (IPS)		
	Low	High	Standardized Difference
County socioeconomic status	−0.06	−0.02	0.032
N students in district	21,454	4753	−0.421
Segregation across FRL status	0.09	0.03	−0.707
District yearly cohort test score growth (in grade units)	0.96	0.98	0.196
Total district expenditures	$12,902	$12,118	−0.144
Magnet school	4%	1%	−0.135
Charter school	9%	5%	−0.154
School in city	37%	10%	−0.602
School in rural	20%	48%	0.647
School in suburb	34%	23%	−0.248
School in town	9%	20%	0.323
N students	502	382	−0.42
Percent free-lunch students	55%	60%	0.175
Percent Hispanic	25%	19%	−0.236
Percent African American	18%	9%	−0.382
Achievement index (3rd grade)	−0.05	0.16	0.247
Lagged achievement index (3rd grade)	−0.02	0.09	0.132

Note. Achievement index is the average of the percentage of students proficient in math in 3rd grade and the percentage of students proficient in English in 3rd grade, after standardizing these variables within state and year. The cutpoint between low and high IPSs is set at the 75th percentile of scores. Column for the low and high prognosis scores show population means.
FRL: free/reduced price lunch; IPS: implementation prognosis score.

BART models to predict the percentage of students receiving the recommended BURST cycles, selecting model parameters with cross-validation to minimize out of sample prediction error. The BART model has an out-of-sample pseudo R-squared of 0.39, suggesting that observable features of schools have some, but not an overly strong, association with BURST usage.

Predict IPS in All Schools

The last step to create the IPS is to estimate the score in the sampling frame. The model just discussed, which was built from the electronic records (i.e., using information on the current user base), can be used to predict the implementation level that schools will enact. This assumes that the relationship between implementation and observed variables will be the same in the experiment as the current user base. As discussed above, we call this predicted score the IPS.

We should attend to the schools dropped when removing the low IPS schools from the sampling frame. If these schools form a unique subpopulation, we will be unable to generalize to this subpopulation (without assuming that treatment effect heterogeneity is independent of IPS). This is the consequence of focusing on high IPS schools. Table 5 contrasts schools with IPSs in the 75th or higher percentile and those with lower IPSs. High-IPS schools are smaller, are from smaller districts, have low levels of segregation across poverty levels, are more likely to be in rural areas or towns, have fewer minority students, and are higher-achieving. These high-IPS schools are also quite different from the average current BURST user (see Tables 1 and 5). Sampling only high-IPS schools, then, will mean sampling schools that are not like the PUB.

Sampling for the High–Implementation Prognosis Schools

The next step is sorting schools in decreasing order based on their IPSs. Schools are then recruited starting at the top of the list and moving down. This sampling approach maximizes the likelihood of recruiting schools that we think should implement the program with fidelity. However, it is likely to recruit a narrower range of schools into the experiment than sampling for the PUB because no stratification occurs. Therefore, we would generally recommend combining this approach with a stratification approach, as we discuss in the next section.

Simulation Results

Table 4 shows the simulation results from the sampling plan just discussed in the high-implementation prognosis row. The approach just discussed demonstrates substantially better ability to generalize to the inference population formed by schools in the 90th percentile of the IPS compared to other methods. However, this comes at the cost of substantially reducing the ability to generalize to the PUB. Interestingly, this approach just discussed also has the strongest ability to generalize to rural schools, a result driven by the high IPSs in rural schools, which is likely unique to this case study.

Sampling for Both the PUB and Likely High Implementers

In this section, we begin to jointly consider the two goals of efficacy studies identified in Section 1. This involves planning an experiment to generalize to the portion of the potential user base likely to use the program at high levels. We start by exploring the relationship between the propensity score and the IPS. When these scores are relatively independent, it is possible to use both scores to design sampling with relatively little loss in generalizability to either population of interest. However, when there is a strong relationship between the two, sampling for IPSs will invariably mean sampling only a portion of the full range of the propensity score.

Figure 2 displays the relationship between the IPS and the propensity score (created using ggplot2; [Wickham, 2016]). The vertical dashed lines divide the graph into the 5 quintiles used to stratify the sampling frame in Section 3.2. The horizontal dashed line marks the 75th percentile for the IPS. Sampling frame schools are represented by small dots and current BURST users by large dots. Note that by sampling only above the dashed line (i.e., high-IPS schools), we are sampling from a region of the graph apart from the majority of the current BURST users, especially for the top 2 quintiles. This finding suggests a trade-off between the goal of sampling for the PUB and sampling for high-implementation schools in the BURST efficacy study, which the simulation shows as sampling for the PUB did not lead to high ability to generalize to high-IPS schools and vice versa.

The slight negative relationship between the prognosis score and the propensity score displayed in Figure 2 is surprising. The schools most likely to adopt BURST tend to use it at relatively low levels. This could be because BURST is a supplemental program and so most likely to be adopted by schools with a relatively small percentage of students in need of the program while schools with more intensive needs (and that would

Table 6. Differences between user base in the top quintile and lower quintiles.

	Low Strata	Highest Stratum	Standardized Difference
County socioeconomic status	−0.18	−0.88	−0.703
N students in district	26,141	101,860	1.24
Segregation across FRL status	0.10	0.21	1.304
District yearly cohort test score growth (in grade units)	0.98	1.06	0.854
Total district expenditures	$12,031	$12,595	0.123
Magnet school	3%	0%	−0.198
Charter school	15%	3%	−0.351
School in city	34%	75%	0.879
School in rural	26%	19%	−0.171
School in suburb	25%	0%	−0.647
School in town	15%	6%	−0.249
N students	479	435	−0.182
Percent free-lunch students	57%	80%	0.767
Percent Hispanic	27%	16%	−0.4
Percent African American	21%	60%	1.312
Achievement index (3rd grade)	0.03	−0.70	−0.742
Lagged achievement index (3rd grade)	0.01	−0.61	−0.711

Note. Achievement index is the average of the percentage of students proficient in math in 3rd grade and the percentage of students proficient in English in 3rd grade, after standardizing these variables within state and year. The low strata column shows mean scores for the four strata with the lowest propensity scores. The highest stratum column shows mean scores for the stratum with the highest propensity score.

treat more students) are likely to adopt a nonsupplemental instructional program. This insight, which is clear from the electronic records (and which we note in hindsight), would suggest that instead of a school-randomized trial examining the impact of a school's adoption of BURST, it might be more prudent to randomize at a lower level to more directly test the impact of BURST on students. After all, the electronic records suggest that few schools are using BURST with more than a small percentage of students, making detecting school-level effects difficult.

Table 3 shows the result of restricting the sampling frame to the 75th percentile or higher on the IPS in the high-prognosis sampling frame columns. There are no schools that can be sampled from the top quintile and likely not enough to meet the sampling quota from the fourth quintile. This raises the question of what schools in the top quintile look like. Are we willing to avoid sampling for this portion of the PUB in order to sample high-IPS schools? Table 6 contrasts the current user base schools in the top quintile with those from lower quintiles. Table 6 shows that the schools in the top stratum are in larger districts, are poorer, are more segregated communities, are in cities, have larger percentages of African American students, and have lower achievement levels but higher yearly district score growth rates.

This highlights the trade-offs in decisions on the sampling approach. Maintaining the focus on schools with IPSs in the top 25% will lead to schools not representative of the PUB being sampled in the top two strata. Maintaining the focus on the PUB will lead to schools with relatively low IPSs being included in the experimental sample. As discussed previously, we have two ways of proceeding depending on which goal we wish to emphasize. To emphasize selecting schools with high IPSs, we first restrict the sampling frame to schools with IPSs in the top 25% and then use the stratification approaches discussed in Section 3.2. This results in the sampling plan in the high-prognosis sampling frame of Table 3. To emphasize selecting for the PUB, we can use the same stratification approach discussed in Section 3.2 and sample within strata based on schools with the highest IPSs (which leads to the sampling plan in columns under sampling frame in Table 3). Both of these approaches are meant to generalize the experiment to schools in the PUB that are likely to use the program with high levels of fidelity, although they place slightly different emphases in defining this population.

Simulation Results

Table 4 again shows the results of these two options. The row PS Stratified–Include only high IPS shows the results from sampling only the top 25% of high-IPS schools while the row PS Stratified–Sample by high IPS shows the results of focusing on generalizing to the PUB. As Table 4 shows, the two approaches provide strong ability to generalize to the current user base with high IPS while retaining the ability to generalize to the current user base. Sampling by high IPS within strata leads to higher SMDs and about equal B-Indexes when generalizing to the current user base compared to the pure stratification approaches while improving the ability to generalize to the high IPS and current user base and high-IPS populations. Stratifying after keeping only the top 25% of schools on IPS leads to a sample less generalizable to the current user base, but more generalizable to the high IPS and current user base and high-IPS populations. Thus, both approaches can be seen as a compromise between targeting the two individual populations of interest as expected.

Discussion

This article demonstrates an approach for designing efficacy studies that addresses two important goals. First is the goal of estimating a treatment effect that generalizes to the population of schools most likely to use the experimental results. We argue that this is the set of schools considering adopting the program being studied, the PUB. Second, the goal of testing the program's theory of action. This requires the program to be implemented to high levels so that it has an opportunity to work. One pathway to supporting this goal is recruiting schools likely to implement the program at high levels. In showing how to design an experiment addressing these goals, we discussed the decisions necessary to manage trade-offs demonstrating the trade-offs with a basic simulation study. There are other important considerations, however. Efficacy studies can be done under ideal conditions (Flay, 1986), which can include additional support and training for schools, another dimension along which trade-offs are required. Providing extra supports to schools may improve the test of the program's theory of action but may preclude generalizing experimental results to schools buying the program on the open market since most schools do not receive this extra support.

Defining the PUB

The first challenge was defining the set of schools likely to adopt BURST in the future, the PUB. Because practitioners highly privilege evidence generated in settings like their own (Nelson et al., 2009), ensuring that the experiment includes schools in the PUB is vital if the results are to be used. However, identifying the PUB is nontrivial as innovations diffuse through markets in complex ways. We rely on the fact that innovations tend to diffuse through homogeneous networks (Rogers, 2003) and verify this assumption using electronic records. This assumption has limitations, which our case study highlighted as the diffusion of BURST was affected by an intentional marketing push. This creates uncertainty in the PUB, which can be incorporated in propensity score models by regularizing model fit or sampling in the region of the propensity score with no common support, although the simulation suggested quite modest effects of these adjustments. At a certain point, uncertainty in the PUB should lead researchers to specify an inference population. The simulation shown here would allow researchers to empirically explore the likely impact of different choices.

Post Hoc Generalization

While this article focused on the design of experiments to address both goals of efficacy trials, there is a growing literature on the post hoc generalization of experimental effects (e.g., Stuart et al., 2015). This literature combines nicely with the approaches discussed here and should be adopted in any attempt to generalize to an inference population. Further, both the propensity score used to stratify the sampling frame and the IPS will be associated with the likelihood of recruitment into the experiment while the IPS was designed to be theoretically predictive of treatment effect heterogeneity. Balancing these scores between the experimental sample and inference population using post hoc generalization approaches, then, should improve generalization.

The design decisions discussed here maximize overlap between the experimental sample and inference populations, reducing the potential bias and variance inflation introduced by post hoc generalizations (Tipton, 2014), while the post hoc generalization smooths out differences caused by recruitment challenges and allows for targeting generalization to populations not part of the planned recruitment process. The simulation suggested a trade-off in how well different designs might generalize to different target populations and the post hoc approaches can alleviate this trade-off supporting generalizing to multiple target populations, although when there is less overlap in the experimental sample and the nontargeted inference population, generalizing may lead to more bias and variance inflation than when there is more overlap.

Empty Sampling Frame Strata

A second challenge that arose in this article, which we expect to be a general challenge, is a lack of schools in the top stratum of the sampling frame. This is a model-dependent challenge and occurs when the propensity score models suggests there is a subset of schools currently using the program for which no similar schools in the sampling frame exist. This could occur when the program has penetrated widely into specific niche markets, which may not be rare due to homogeneous diffusion processes, or if the sampling frame is reduced due to other experimental considerations (e.g., based on school size to ensure high power). The researcher should examine the nature of current program users who fall into this empty stratum and decide whether this implication of the propensity model is reasonable. If it is not reasonable, then some regularization of the propensity model might be appropriate. If it is reasonable, then the researcher should characterize this population for which no schools in the PUB appear to exist.

Role of Funders in Supporting Better Experimental Design

The challenges discussed in this article highlight the complexity of planning for an experiment. Current practice is unlikely to shift and adopt this complexity without support and pressure, which can come from funders. After all, in current practice, which ignores the complexities associated with intentional sampling, the recruitment process still often proceeds well into the intended start of the study, costing more than expected (Spybrook, Puente, & Lininger, 2013; Tipton et al., 2016). Further, current Institute of Education Sciences requirements to identify a sample before applying for a grant may hinder this careful sampling process by disincentivizing the extra work to craft a careful sample, especially when experience often finds schools becoming disinterested between initial recruitment before the submission of grants and the start of an experiment.

Especially for evaluations of the types of programs likely to have electronic records, such as computer-based programs, funders could create the expectation that these records will be used to design an experiment. Funders might even consider small initial grants with the goal of studying electronic records and planning more detailed experiments. These studies could inform other aspects of experimental design. For example, the electronic records from BURST indicate that most schools have only a small proportion of students enrolled in the program. This is more consistent with a few teachers adopting and using the program rather than full school adoption, which, if true, would raise the possibility of research designs randomizing at the teacher level or at least highlight the importance of tracking individual teachers' usage of BURST, rather than school-level usage. This sort of insight is often not available to researchers during the initial design phases of an experiment, but would become clear after careful analyses of electronic records, analyses that require time.

Obtaining Buy-in for the Sharing of Electronic Records

Not all providers will be willing to share electronic records with researchers. After all, these records might have confidential information, such as student data, or contain proprietary data. It is important, then, for researchers to both build trust with providers and craft clear data use agreements that ensure the concerns of providers are addressed, such as not revealing the number of schools using a program (e.g., Tipton et al., 2014). It is also important to highlight the key benefits that sharing this information will give to providers. For example, the use of the IPS should enable the targeting of schools likely to have high-quality implementation for efficacy studies. This, in turn, makes it more likely that the efficacy study will lead to positive results (assuming the theory of action is valid), a clear benefit to the provider who is selling and promoting their program. Targeting the sample to schools deciding whether to adopt the program may, similarly, make the experiment more valid for this set of schools, increasing uptake of the program (assuming positive experimental results). Funding agencies can again step in here to incentivize the sharing of electronic records, either by making it a condition of receiving grants or explicitly favoring applications where this sharing occurs.

Conclusion

Overall, this article shows how electronic records created in the course of the daily operation of programs can be used to plan more useful efficacy studies, improving both the generalization of treatment effects to relevant populations and testing the program's theory of action. The planning process, however, is not straightforward and requires managing multiple trade-offs regarding which experimental goals to emphasize and specifying clearly a target population. The approaches discussed in this article provide a framework for managing these trade-offs. This should help researchers to make more intentional decisions to design more useful efficacy experiments.

Funding

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305A120811 to the University of Michigan. The opinions expressed are those of the authors and do not represent the views of the institute or the U.S. Department of Education.

ARTICLE HISTORY

Received 8 August 2018
Revised 12 June 2019
Accepted 17 June 2019

References

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites non-randomly. *Educational Evaluation and Policy Analysis*, *38*(2), 318–335. doi:10.3102/0162373715617549

Bowers, J., Fredrickson, M., & Hansen, B. B. (2010). RItools: Randomization inference tools (Version 0.1–11) [R package].

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. doi:10.1214/09-AOAS285

CLEVER.com. (2018, May). *Edtech usage myths vs. fact*. Retrieved from https://www.edsurge.com/news/2019-02-05-myth-vs-fact-how-much-do-you-know-about-edtech-usage-in-schools-infographic

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: National Implementation Research Network.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, *15*(5), 451–474. doi:10.1016/0091-7435(86)90024-1

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Retrieved from http://dibels.uoregon.edu/

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481–488. doi:10.1093/biomet/asn004

Institute of Education Sciences. (2016). *Building evidence: What comes after an efficacy study?* (Technical Working Group Meeting Summary) (p. 17). Washington, DC.

Kapelner, A., & Bleich, J. (2013). bartMachine: Machine Learning with Bayesian Additive Regression Trees. ArXiv:1312.2171 [Cs, Stat].

Nelson, S. R., Leffler, J. C., & Hansen, B. A. (2009). *Toward a research agenda for understanding and improving the use of research evidence* (p. 80). Portland, OR: Northwest Regional Educational Laboratory.

R. Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & DiSalvo, R. (2017). *Stanford Education Data Archive (Version 2.0)*. Retrieved from http://purl.stanford.edu/db586ns4974

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Mumbai, India: Free Press.

Rosenbaum, P. R. (2009). *Design of observational studies*. Berlin, Germany: Springer.

Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students Placed at Risk (JESPAR)*, *22*(3), 178–184. doi:10.1080/10824669.2017.1334560

Spybrook, J., Puente, A. C., & Lininger, M. (2013). From planning to implementation: An examination of changes in the research design, sample size, and precision of group randomized trials launched by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, *6*(4), 396–420. doi:10.1080/19345747.2013.801544

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. doi:10.1080/19345747.2016.1205160

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*(3), 475–485. doi:10.1007/s11121-014-0513-z

Tipton, E. (2013). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, *37*(2), 109–139. doi:10.1177/0193841X13516324

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. doi:10.3102/1076998614558486

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Castilla, V. R. d. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, *9*(sup1), 209–228. doi:10.1080/19345747.2015.1105895

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114–135. doi:10.1080/19345747.2013.831154

Tipton, E., Wang, Q., Spybrook, J., & Fitzgerald, K. (2019). Assessing the relevance of IES funded goal 3 and 4 studies to important policy populations. In *The past, present, and future of recruitment and generalization in education.* Presented at the SREE 2019, Washington, DC.

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

Wickham, H. (2017). *tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2. 1.* Vienna, Austria: R Core Team.