Method Note

American Journal of Evaluation
2019, Vol. 40(3) 414-430
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1098214018810519
journals.sagepub.com/home/aje
SAGE

Improved Generalizability Through Improved Recruitment: Lessons Learned From a Large-Scale Randomized Trial

Elizabeth Tipton[1] **and Bryan J. Matlen**[2]

Abstract

Randomized control trials (RCTs) have long been considered the "gold standard" for evaluating the impacts of interventions. However, in most education RCTs, the sample of schools included is recruited based on convenience, potentially compromising a study's ability to generalize to an intended population. An alternative approach is to recruit schools using a stratified recruitment method developed by Tipton. Until now, however, there has been limited information available about how to implement this approach in the field. In this article, we concretely illustrate each step of the stratified recruitment method in an evaluation of a college-level developmental algebra intervention. We reflect on the implementation of this process and conclude with five on-the-ground lessons regarding how to best implement this recruitment method in future studies.

Keywords

generalization, cluster-randomized trials, web-based tutors, developmental algebra, community colleges

Over the past 10 years, the Institute of Education Sciences has funded over 100 randomized control trials (RCTs) evaluating the efficacy or effectiveness of programs and curricula aimed at improving educational outcomes for children in pre-K, K–12, and postsecondary education. These RCTs typically include many schools, resulting in either cluster randomized designs (e.g., randomizing schools) or multisite designs (e.g., randomizing students within schools). When implemented well, the results of these RCTs are included in the What Works Clearinghouse (WWC), ideally providing policy makers with the information they need for making evidence-based decisions in practice.

[1] Teachers College, Columbia University, New York, NY, USA
[2] WestEd STEM Program, Redwood City, CA, USA

Corresponding Author:
Elizabeth Tipton, Teachers College, Columbia University, 2006 Sheridan Road, Evanston, IL 60201, USA.
Email: tipton@tc.columbia.edu

In the ideal, a policy maker—for example, a community college developmental education program director—could turn to the WWC to find evidence regarding developmental math curricula. However, while the WWC provides the results of RCTs, these studies may have taken place in contexts quite different from those that the program director is concerned with. This problem is one of *generalizability*, one facet of *external validity* (Shadish, Cook, & Campbell, 2002). Here, we use the term generalizability to refer to extrapolations across units (e.g., people and students) and sometimes settings (e.g., schools); in comparison, external validity also involves extrapolations across time, different versions of the intervention, and outcomes. Importantly, this generalizability problem is made even more difficult by the fact that the majority of RCTs conducted in education take place in convenience samples, with very little information on these samples provided in reports (Fellers, 2016; Olsen, Orr, Bell, & Stuart, 2013).

Now that RCTs are recognized as both possible and the ideal for making evaluations of educational interventions, questions of how to make generalizations from these findings to schools and students that are *not* in the studies have become increasingly important. This concern has led to a new wave of methodological developments focused on improving generalizations via improved site selection and recruitment (e.g., Tipton, 2014a, Tipton et al., 2014); methods for assessing the representativeness of a sample for different populations (e.g., Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2014b); and improved estimation of population average treatment effects (ATEs; e.g., Tipton, 2013; for an overview of these approaches, see Stuart, Ackerman, & Westreich, 2017; Tipton & Olsen, 2018). While this literature includes a variety of examples, its main focus has been on providing designs, estimators, and their properties. To date, there has been very little work in this area on the *implementation* of these methods in real studies (for exceptions, see Roschelle et al., 2014; Stuart & Rhodes, 2017; Tipton et al., 2016).

The goal of this article is to provide a case study of the use of methods for sample selection based on their implementation in a study advised and led by the authors of this article. This IES-funded RCT was designed to determine the effect of a web-based activity and testing system (WATS) in the population of community colleges in California. In order to address concerns with generalizability, the research team designed and implemented a stratified sample selection plan based on Tipton (2014a). This report begins by providing further context on this study, reviews the development of the sample selection plan, and then discusses how this plan impacted the resulting sample. We then provide a reflection on the implementation of this process, concluding with possible on-the-ground lessons learned, with an emphasis on the types of tools and communications necessary to ensure that the methods are implemented well. By providing details about the implementation experience, we hope to provide insights that may aid other researchers in implementing this method in similar types of evaluations.

The Intervention and Study

The case study we describe involved an evaluation of a widely used WATS for supporting community college developmental algebra learning. WATS are characterized by the delivery of adaptive problem sets that students complete in a self-paced way, individualized problem-solving feedback, and instructional media resources. To support teachers, WATS often provide feedback on student progress so that they can assist and monitor students' learning. WATS in many ways mimic aspects of human tutoring (e.g., provision of differentiated instruction, progress tracking, and motivational components). Human tutoring has been shown to be highly effective for improving student achievement, with effect sizes ranging from 0.4 to 2.0 (Bloom, 1984; Cohen, Kulik, & Kulik, 1982); thus, the use of WATS as a proxy to human tutoring may provide valuable scaffolding for community college students. However, WATS are more cost and labor effective than human tutoring, making them more practical to use at scale.

The particular WATS we investigated is freely available on any web-based browser. It has over 6 million users worldwide, making it one of the most popular WATS tools available. Despite its popularity, however, no large-scale study evaluating its efficacy had been conducted. This example study examined the use of this WATS in developmental algebra courses at the community college level. The focus was on this population because algebra is widely known to be a gateway course for later mathematics success. In addition, many college students exhibit marked difficulty in learning algebra (Grubb & Gabriner, 2013). As an illustration of this point, in California (where the study took place), 61% of first-year students enrolled in state universities had math skills below college level (Hindes, Hom, & Brookshaw, 2002) and as many as 87% of credits taken at some community colleges are comprised of basic math skills credits (California Community Colleges Chancellor's Office, 2012)—despite earning passing grades in their high school courses, students still seem to struggle to learn the material (Long, Iatarola, & Conger, 2009). Given the stark challenge facing preparedness for postsecondary mathematics, there is increasing interest in effective instructional methods to improve postsecondary mathematics outcomes.

The goal of the study was to estimate the average impact of this WATS platform in supporting community college students' Algebra I knowledge. In the study, both community colleges and instructors within these colleges were recruited, and then instructors in each school were randomly assigned to either use the WATS platform as a part of their instruction (treatment condition) or to teach using their normal practices (control condition). Those taking part participated for both semesters in the 2015–2016 school year, though efficacy was measured only in the spring semester (allowing the fall semester for "practice").

In this article, we describe our experience using a stratified recruitment method for this WATS evaluation. Here, we focus on the recruitment of community colleges into the evaluation, a process that occurred prior to the first year of the study. In practice, recruitment involved both targeting community colleges and instructors within those colleges. In some cases, permission at a community college was garnered first and then recruitment of instructors took place. In other cases, potential instructors were targeted within community colleges, and once they agreed to take part, permission was sought from the college. While not the focus of this article, it is notable that the recruitment of instructors was not easy and required the inclusion of a second cohort (about 25% of the sample) who participated the following year in order to reach the required sample size. However, our recruitment efforts were initially successful in achieving a sample that was similar to the broader population in which we intended to generalize results. Thus, our hope is that by sharing our experience in implementing these methods in a single study, other researchers may be encouraged to implement these methods and can be more successful in doing so.

Designing a Recruitment Plan for Generalization

The recruitment plan used in this study involved a stratified sample selection plan. This plan followed strategies developed in Tipton (2014a) and included the following steps: (1) define an inference population, (2) select possible treatment effect moderator variables, (3) create and describe the strata, and (4) develop a within-stratum recruitment strategy. The first three steps were conducted solely among the methodologists on the team, whereas the fourth step involved collaborative efforts between the methodologists and recruiters involved in the project. In this section, we provide an overview of the development of the recruitment plan for the WATS evaluation. In later sections, we discuss lessons learned regarding the implementation of the within-stratum recruitment method and conclude with possible tips for how to improve the success of this method in future evaluations.

Step 1: Define an Inference Population

In an RCT, the goal is to estimate an ATE for an intervention. If the effect of a treatment or intervention varies across people or settings, then the treatment impact estimated in the RCT is specific to the sample. But the results of an RCT are typically used to guide practice in schools and for students *beyond* the sample. This process of generalizing requires researchers and policy makers to make connections between features of the sample and an inference population, a process that is difficult when delayed until after a study is completed.

For this reason, all statistical methods for improving generalizations begin by requiring researchers to define and enumerate an inference (aka target) population of their study (for a review, see Tipton & Olsen, 2018). This process requires planning by researchers *at the beginning of their study* to ask questions regarding the goals of the study, who will likely use the results, and what is feasible within the study (see Tipton & Peck, 2017). This process, while seeming straightforward, is typically more complex as it brings to light tensions between what is ideal (very broad generalizations) and what is practical (often much narrower generalizations).

In the WATS evaluation, there were a couple of possible inference populations and considerations. For all possibilities, the population was limited to community colleges offering semester-long developmental algebra courses in California. The state of California was selected in part because the state is large and diverse and in part because the study team sought to decrease variability that may result from differing high school mathematics standards and graduation requirements across multiple states.

Given available resources, the broadest population for this study would be one that included all community colleges, developmental algebra instructors, and developmental algebra students across the state of California. This population would be essential if the ATE estimated in the study was to be used to make policy decisions regarding all community colleges in California, as might occur if the state were to mandate that the curriculum studied were to be implemented in all developmental algebra community college courses. However, the program under study consisted of an online tool rather than a comprehensive curricular program that is uniquely developed for classroom implementation (e.g., such as Quantway; Sowers & Yamada, 2015); thus, it would be unlikely that a state, or even an entire community college, would mandate that the program be used in all classrooms.

The WATS being evaluated was freely available to instructors and students, thus making it was more likely that the outcomes of the study would be most useful to instructors most interested in using the program. This target population consisted of developmental math students and instructors *interested in using WATS* in community colleges in California. This meant that the ATE could be interpreted for making decisions by California community college instructors interested in the WATS. Importantly, this also meant that the evaluation would not be able to determine the effect of the program on instructors not interested in the program.

The latter population—which we ultimately chose for this study—brought with it some benefits and challenges. One benefit was that recruitment did not involve convincing community colleges to mandate that their instructors take part in the study. One challenge, however, was that it was not necessarily the case that the community college instructors interested in taking part in an *evaluation* of the WATS were the same as those interested in possibly *using* the WATS. This means that the ability to generalize to this population hinges on an assumption that these two groups are equivalent on average. Given the constraints of the study, however, this was determined to be the broadest possible inference population given the resources available.[1] To enumerate this population, the study team used the California Community College Chancellor's office Management Information Systems Data Mart (MIS Data Mart; http://datamart.cccco.edu/). MIS Data Mart is a publicly available database that contains information about all California community colleges. There are 118 California community colleges listed in MIS Data Mart, two of which are satellite campuses of larger

Table 1. Demographics on Population of 113 Community Colleges in California for the fall, 2014, Semester.

	Population ($N = 113$)	
College Characteristic	*M*	*SD*
Total student enrollment	20,168	13,149
Math basic skills student enrollment	1,018	769
Number of math basic skills sections offered	35.18	30.21
Total academic employment	518.59	304.14
Math basic skills full-time equivalent status	149.89	115.79
Proportion female	0.53	0.06
Proportion African American	0.07	0.08
Proportion Asian	0.13	0.10
Proportion Hispanic	0.42	0.17
Proportion White Non-Hispanic	0.30	0.16
Proportion U.S. citizen	0.86	0.09
Proportion aged 19 or less	0.26	0.06
Proportion aged 20–24	0.34	0.06
Proportion aged 25–39	0.26	0.05
Proportion aged 40 to 49	0.07	0.03
Proportion aged 50 above	0.08	0.05
Proportion first-time students	0.17	0.06
Proportion first-time transfer students	0.08	0.05
Proportion returning students	0.11	0.05
Proportion day time students	0.74	0.08
Proportion evening students	0.18	0.05
Proportion of students taking unit load from 0.1–9	0.47	0.10
Proportion of students taking unit load from 9–14.9	0.38	0.08
Proportion of students taking unit load of 15+	0.09	0.04
Proportion tenure track faculty	0.20	0.05
Proportion temporary faculty	0.50	0.08
Proportion math basic skills retention	0.83	0.06
Proportion math basic skills success (students with passing grades)	0.54	0.08
Median household income for county	61,102	14,291
Proportion of the county considered in poverty	0.17	0.05

campuses, two more are adult education centers, and one is an instructional video college. Because the adult education and instructional video colleges did not offer developmental algebra courses at the time of the study, they were excluded from participation. The two satellite campuses were considered a part of their parent campuses, resulting in a total population of 113 community college sites. Table 1 provides information on this population across a range of student and instructor characteristics (described in more detail, below).

Step 2: Select Possible Treatment Effect Moderators

Once a population is defined, the goal is to develop a strategy to recruit a sample that is compositionally similar to this population. Statistically, a simple random sampling approach would be ideal, since—assuming no nonresponse—the resulting sample would be similar, on average, to the population on *all* covariates. However, the dearth of RCTs using random sampling for selection speaks to some of the difficulties of this approach (see Olsen et al., 2013). In this particular study, there were several constraints making a random sampling plan difficult. First, conducting a random sample would have required within-school sampling frames that did not always exist, and the creation of these frames would have been costly. Second, it was nearly certain that randomly selected institutions or instructors might have been more difficult to recruit—both in terms of time and incentives—than instructors and institutions that recruiters were more familiar with.

While not as statistically ideal, the approach developed by Tipton (2014a) provided an alternative that balanced statistical concerns with bias and practical concerns regarding time and money. Using this approach, the study team was asked to define similarity between the sample and population with respect to covariates that would likely *moderate* the impact of the treatment. This required researchers to ask, why (and how) might the effect of this intervention differ across students and settings? If the team was able to enumerate *all* moderators, then the ATE estimated in the study would be an *unbiased* estimate of the population ATE (a "sampling ignorability" assumption; see Stuart et al., 2011; Tipton, 2013). In practice, the selection of moderators is limited by the data that are available and by the fact that it is impossible to know a priori how the impacts will vary before the study has commenced. The strategy suggested by Tipton (2014a) is aimed at *bias-robustness*—erring on the side of including covariates when their effect as a moderator is unknown.

In the WATS evaluation, we began by selecting characteristics that may be related to the study outcome (i.e., community college students' developmental algebra achievement). MIS DataMart provides a host of information on each California community college including student demographics, student enrollments in courses and programs, and information on community college faculty. We identified 28 variables we believed to be potentially related to the study outcome and that might moderate the treatment effect (see Table 1 for list of all variables examined, as well as means and standard deviations for each variable).

We created a database with information on these variables for each college,[2] using fall 2014 semester data (the most recently available semester at the time of the database's creation).[3] We also collected information on the median household income and the proportion of all ages in poverty in the county in which each community college resided, using 2013 U.S. census data. In total, the complete database included information related to 30 variables for the 113 community college sites.

Step 3: Divide the Population into Strata

At this stage, the goal is to develop a recruitment strategy that will result in a sample that is similar, on average, to the inference population on the set of moderators selected. In the WATS study, this required the sample to be like a miniature of the population of 113 community colleges in California (defined in Table 1). With 30 variables—many of which were continuous—creating strata directly based on the covariates would have resulted in far too many strata. Tipton (2014a) proposed that an alternative approach was to use k-means cluster analysis to divide the population into strata based on the covariates. Like stratification, the goal of k-means cluster analysis is to maximize the heterogeneity between clusters (here strata), thus minimizing the heterogeneity within clusters (strata). This method can be implemented in most statistical software; here we used the free program R (R Core Team, 2016).

While k-means cluster analysis can be used to generate any number of strata in an inference population, in practice, smaller values of k are more reasonable for recruitment. In the WATS evaluation, we attempted to choose the number of strata (k) so as to balance the goals of explaining the heterogeneity between strata, on the one hand, with the desire to choose a reasonable number of strata to guide recruitment efforts, on the other. Ultimately, we decided on a five-stratum solution, which captured 29% of the variation between strata. Although the amount of variance explained may be considered low by some standards, we chose this five-stratum solution in part because it represented a manageable number of strata for study recruiters and in part because increasing the number of strata beyond five did not yield large improvements in the amount of variance explained.

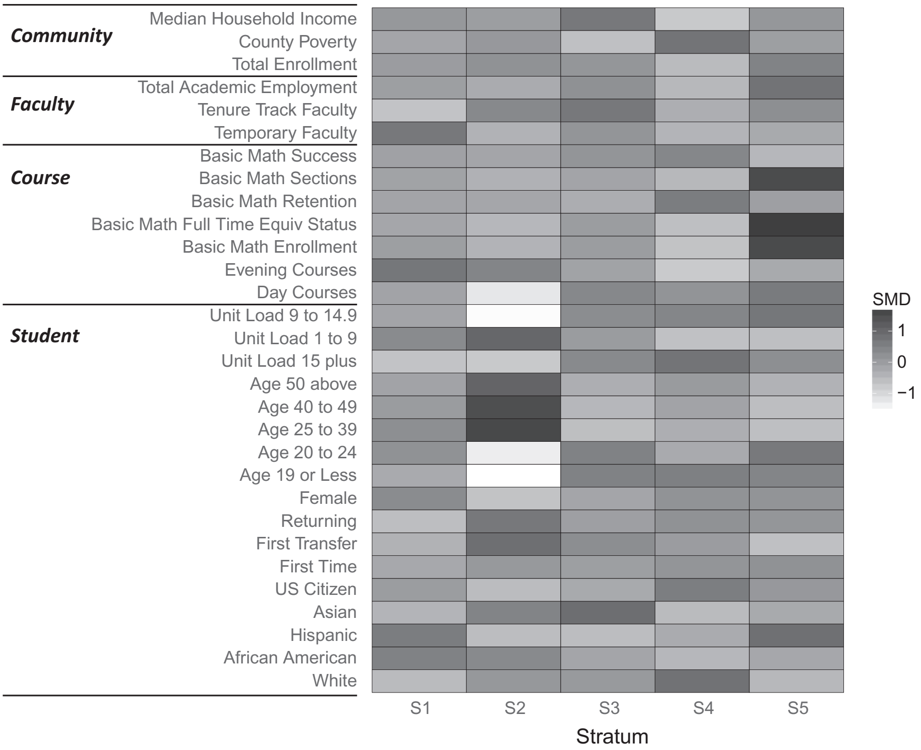


Figure 1. Heat map of the standardized mean differences (SMDs) on each college-level variable, for each stratum, grouped by community, course, faculty, and student characteristics. This representation was provided to recruiters, along with short descriptions that characterized each cluster, in order to support recruitment efforts.

One of the benefits of this cluster-analysis approach for stratification is that the strata can be used to help better understand and describe the population of the evaluation. In the WATS evaluation, to help in describing these strata, we created a heat map (Figure 1) of the relative density of characteristics of the clusters for each college characteristic. From this heat map, we generated short descriptions of each stratum, which we outline below:

- **Stratum 1.** Represented 25% of colleges. These were colleges with a total student enrollment near the average (across all community colleges in the state) whose students tended to take more credits in the *evening* relative to colleges in other clusters. Stratum 1 colleges had more Temporary Faculty, and more Hispanic students and African American students.
- **Stratum 2.** Represented 15% of colleges. These colleges served primarily students aged 25 and above who *took fewer credits* and more commonly were evening students.
- **Stratum 3.** Represented 22% of colleges. These were colleges with a total student enrollment near the state average where students were more commonly *Asian, younger, and enrolled full time during the day.*
- **Stratum 4.** Represented 23% of colleges. Stratum 4 represented *smaller colleges* that had a higher proportion of *white students* that tended to be younger, mostly full time, and took fewer evening courses.

- **Stratum 5.** Represented 15% of colleges. These were *larger colleges* that had more Hispanic and younger students. Students tended to take *more daytime courses*, with more fulltime loads and many remedial mathematics courses and high remedial math enrollment.

Step 4: Develop a Recruitment Plan

After dividing the population into five strata, Tipton (2014a) and Tipton et al. (2014) provide two options for recruitment. One option—best implemented when refusal bias is low—is to randomly select sites within each stratum for recruitment (see Tipton & Olsen, 2018, for further discussion). Another strategy is to rank sites within each stratum from most to least "typical" (closest to the stratum average) and to prioritize more "typical" sites for recruitment. In the WATS evaluation, we followed the latter strategy. Recruitment target enrollments for each stratum were set (proportional to that in the population of schools; see Tipton, 2014a, as well as Tipton & Peck, 2017, for a discussion of other possible approaches). This resulted in the following goals for recruitment of colleges: eight to nine in Stratum 1; five to six in Stratum 2; seven to eight in Stratum 3; eight to nine in Stratum 4; and five to six in Stratum 5.

Up until this point, the development of the recruitment strategy involved only the methodologists within the team (i.e., the authors of this article). The next step involved communicating the stratified recruitment method to recruiters, who were not a part of its initial conceptualization, but who nevertheless would be integral to its success. This process began by conveying the basic rationale underlying the strategy to recruiters who—at the time of the method's introduction—had already begun recruiting colleges into the study. The recruitment method was framed as a way to prioritize recruitment efforts (e.g., to recruit colleges within a certain stratum that presently had low representation in the sample).

As mentioned above, target numbers of colleges were identified for each stratum and were provided to the recruitment team. However, rather than providing these recruitment targets to recruiters alone, we recognized a need to update our recruiting targets dynamically (as each college was recruited), so that recruited proportions of colleges in each stratum did not stray drastically from the target proportions. To support this process, we created a visual representation that showed the proportions of recruited colleges in each stratum relative to their targets—the proportions updated each time a new college was recruited. In this way, we ensured recruitment goals were always aligned with recruiting colleges that would help achieve the desired target proportions.

To further support recruitment efforts, we provided the team with the stratum descriptions (described previously) and an overview of the goals for recruitment and for the study. Recruiters were told that they could use these short descriptions as a part of their recruiting narrative to prospective colleges. For example, we suggested that recruiters could say statements like,

> Your college serves a student population with a high proportion of Hispanic students and with high remedial math enrollment. We would really like a college *like yours* to participate, so that results can inform how students in these demographic categories learn best.

Or,

> We're trying to get an accurate picture of how this works in a wide range of different settings and *your school represents* colleges that have more remedial mathematics enrollment, and we want to make sure our study captures that experience.

Following recommendations for best recruitment practices (e.g., Roschelle et al., 2014), recruiters attempted to obtain buy-in from all community college stakeholders, including at both the administrative and instructor levels. While administrators were not the direct participants of the

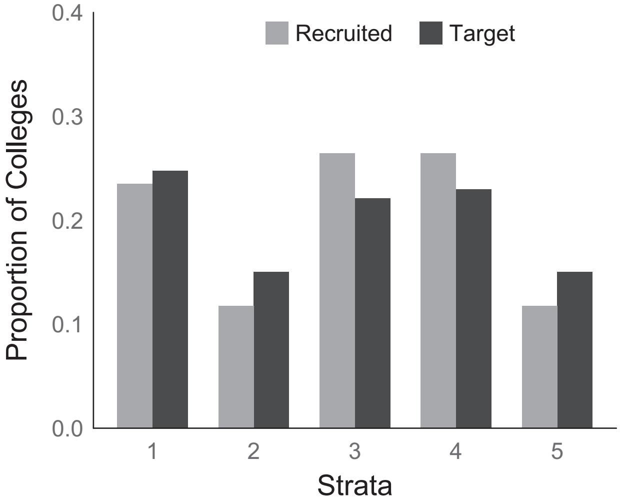


Figure 2. Proportion of recruited colleges relative to the target proportion at the beginning of the study.

intervention, they held the authority to make decisions that could ultimately impact participation (Terrell & Bugler, 2018; e.g., assigning instructors to teach developmental math in back-to-back semesters); thus, obtaining administrative buy-in was deemed a priority. Toward these ends, recruiters sent e-mails to past contacts as well as "cold" calling and e-mailing instructors whose contact information was scraped from college websites. At the same time, recruiters made phone calls, sent e-mails to college administration staff, and sent physical recruitment flyers to the colleges. Thus, sites and instructors were recruited simultaneously, and sometimes occurring in a "bottom-up" fashion, after instructors expressed interest in participation, and sometimes occurring in a more "top-down" fashion, after administrative staff provided approval for the study.

Incentives also were used to encourage instructor participation. These incentives included an honorarium of up to US$1000 for participating in the study, and a US$25 Visa gift card incentive for each referral of a colleague at their college who consented and was qualified for participation. Finally, recruiters were encouraged to keep track of the schools that they contacted and reasons that were given for agreeing or refusing to participate in the study. As colleges agreed to join the study, this information was tracked by the study team (see Lessons Learned section for more information).

Evaluating the Success of This Plan

Recruitment staff contacted both administrative staff and instructors at community colleges through e-mails and phone calls. Community colleges were considered in the study when at least one eligible instructor from the college provided consent to participate in the study, and when their college approved of the participation. In total, the recruitment process for this project took a total of 9 months, with the majority of the community colleges signing on between 6 and 7 months into the recruitment, which was approximately 2 months before the study began. Whereas the number of colleges recruited was close to our original goal, the target number of instructors within those colleges was not reached in the first year of recruiting, motivating the study team to recruit a second cohort of participants (largely within the already recruited colleges) the following year. However, we focus on the first cohort of the study recruitment in the present article.

Table 2. Means, Standard Deviations, and Absolute Standardized Mean Difference Between the Inference Population and the Study Sample, for Each Variable in Each Stratum.

	Sample		*ASMD*					
College-Level Variables	*M*	*SD*	Overall	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Total enrollment	21,146	13,527	0.07	0	0.17	0.11	0.53	0.48
Math basic enrollment	974	703	0.06	0.05	0.41	0.03	0.66	1.53
Math basic sections count	38	33	0.10	0.12	0.37	0.15	0.48	1.52
Total AcadEmp	568	322	0.16	0.05	0.27	0.21	0.49	0.8
Math basic FTES	143	92	0.06	0.18	0.48	0.01	0.61	1.74
Female	0.54	0.04	0.01	0.29	0.67	0.16	0.17	0.15
African American	0.05	0.04	0.24	0.5	0.34	0.16	0.47	0.21
Asian	0.14	0.13	0.11	0.42	0.45	0.89	0.54	0.25
Hispanic	0.4	0.17	0.09	0.6	0.58	0.57	0.27	0.84
White Non-Hispanic	0.32	0.16	0.12	0.53	0.07	0.04	0.81	0.49
U.S. citizen	0.86	0.10	0.02	0.02	0.56	0.24	0.57	0.07
Aged 19 or less	0.26	0.06	0.07	0.26	1.57	0.50	0.53	0.45
Aged 20–24	0.34	0.05	0.03	0.19	1.33	0.49	0.25	0.67
Aged 25–39	0.25	0.04	0.08	0.24	1.57	0.6	0.31	0.62
Aged 40–49	0.07	0.02	0.02	0.01	1.48	0.48	0.13	0.59
Aged 50 above	0.08	0.05	0.01	0.14	1.07	0.33	0.02	0.39
First time	0.16	0.05	0.17	0.22	0.05	0.04	0.12	0.2
First transfer	0.08	0.05	0.01	0.38	0.87	0.27	0.01	0.62
Returning	0.10	0.05	0.30	0.59	0.68	0.06	0.17	0.11
Day	0.75	0.07	0.12	0.13	1.24	0.38	0.16	0.64
Evening	0.18	0.05	0.07	0.71	0.44	0.11	0.79	0.24
Unit 1–9	0.47	0.08	0.06	0.33	1.02	0.01	0.64	0.61
Unit 9–14.9	0.38	0.06	0.01	0.15	1.54	0.28	0.43	0.71
Unit_15plus	0.09	0.04	0.11	0.68	0.77	0.33	0.76	0.24
Tenure TrackFac	0.21	0.05	0.10	0.68	0.35	0.69	0.31	0.24
Temporary Fac	0.50	0.08	0.07	0.68	0.4	0.12	0.43	0.25
Math basic retention	0.84	0.05	0.13	0.14	0.19	0.29	0.61	0.07
Math basic success	0.56	0.08	0.25	0.09	0.16	0.12	0.4	0.48
MedHouse income	60,029	14,457	0.08	0.08	0.03	0.68	0.78	0.09
County pov	0.17	0.05	0.04	0.16	0.07	0.62	0.75	0.05

Note. ASMD = absolute standardized mean difference; FTES = full-time equivalent student; *M* = mean; *SD* = standard deviation.

Overall, recruitment for Cohort 1 of participants yielded a study sample of community colleges similar to the population in terms of stratum allocation, as seen in Figure 2. As the figure indicates, recruitment was easier in Strata 3 and 4 than in the other strata, with the most difficulties in recruitment occurring in Strata 2 and 5. More importantly, the average values for the covariates in the total sample were quite similar to those in the population, as shown in Table 2.

Taking a step back, the purpose of the stratified recruitment plan was to garner a final sample of community colleges that was similar, on average, to a population of community colleges in California on a set of potential moderators (i.e., those in Figure 1). As Tipton (2014a) notes, when the sample is similar to the population on these moderators, the ATE estimated can be generalized to the population. Therefore, it is important to determine whether the sample is in fact sufficiently similar to the population to warrant these generalizations. Tipton (2014b) proposed a summary statistic—the generalizability index—that can be used for this purpose. In this study, the generalizability index was calculated to be .93—indicating that the sample is 93% similar to the population on these moderators. This value is considered "very high"; in fact, values this high indicate that the sample is as similar to the population on this set of 30 variables as would be expected in a random sample of the same size. Another metric, proposed by Stuart et al. (2011), focuses on the average absolute standardized mean difference (ASMD). In this study, the ASMD across these covariates was 0.09, which is small, indicating a high degree of similarity between the resulting sample and population.

Finally, we note that this ASMD is likely a conservative estimate, as many variables consisted of proportions between 0 and 1, and small standard deviations in these variables resulted in large ASMD estimates. Taking the absolute mean difference in these proportion variables (without standardizing) yielded an average difference of 0.01 (e.g., the sample proportion was 0.09, and the population proportion was 0.10). The average ASMD for all other continuous covariates (excluding proportions) was 0.09. Altogether, these assessment statistics indicate that the resulting sample of community colleges could be interpreted as broadly representative of the population of community colleges in California that serve students in developmental algebra courses (on the moderators indicated).

Lessons Learned

In the previous section, we skipped directly between the plan for recruitment and the (largely positive) results of these efforts. In between these two steps, however, there was a lot of on-the-ground learning and experimentation regarding the best strategies for recruitment. Our internal process involved frequent communication between the recruiters and the methodologists, and our processes evolved over time—adapting based on feedback regarding what was and was not working. Many of these lessons were learned through trial and error and included problems not addressed in the statistical literature on sample selection or recruitment. The goal of this article, therefore, is to share our experiences, thereby improving the likelihood that others can implement approaches aimed at improving the generalizability of the results of their experiments.

To distill these experiences into lessons, the authors jointly reflected on the development of the sample selection strategy previously described including what went well, what didn't, and adjustments made. In order to verify the authenticity of our experiences and obtain additional perspectives, we conducted informal interviews with three of the recruiting staff involved in the project. These interviews were approximately 30 min each and included both a set of questions and informal discussion. These were then distilled into lessons learned—five in total—with an eye toward supporting researchers in conducting future studies. Importantly, this process was not intended to serve as a formal qualitative study. Future work might intentionally study this process through the use of an external evaluator. Ideally, such insights will better prepare researchers for challenges in implementing the stratified recruiting method and provide strategies for improving communication within research teams during recruitment phases.

Lesson 1: Learning About the Population can Take Time

Building the population frame in this study required adequate planning and effort. Even though the population had been decided upon conceptually (California community colleges) well before the start of recruiting, there was substantial work involved in reconciling this ideal population with how data on the population were represented in the data source (MIS Data Mart). For instance, some research was needed to identify colleges that did not offer developmental algebra during the study semester, and these colleges were excluded from the population frame. In addition, data pulls were not straightforward, as there were often missing data for some colleges, requiring data to be pulled from previous semesters and compiled for those colleges. To ensure accuracy, a statistical script was generated to reproduce the database from code.

Once a complete population frame was created, the methodologists conducted cluster analyses on the population frame, decided on the optimal stratum size, and created visual representations (e.g., Figure 2) and verbal descriptions of each stratum to present to recruiters. This process involved multiple meetings and discussions within the methodological team. In total, the entire process—starting from the building the database to the point where strata were ready to present to recruiters—took place over the course of approximately 2 months. Future studies and evaluations should build in adequate time to complete this process, ideally before recruiting begins (although in the present study, this process was completed in parallel with the start of recruitment).

It is worth noting that for studies conducted in the K–12 population, researchers can substantially reduce this time by using a free web-based tool called The Generalizer (www.thegeneralizer.org). The Generalizer provides a step-by-step walk through of the process of building the population frame and identifying relevant strata from which to recruit (Tipton & Miller, 2015). The website ultimately provides researchers with a list of all the schools that could be recruited into the study, organized by their stratum, and ranked based on their stratum similarity (using public information from the Common Core of Data and from the American Community Survey). Using The Generalizer, the entire process of deciding on an inference population, conducting cluster analyses, and identifying strata can be completed within a few minutes. The Generalizer can also be used post-study to determine the degree to which the final sample is similar to different inference populations. However, even if a study involves a K-12 population and can utilize The Generalizer for developing a recruitment plan, we recommend building in ample time to discuss and decide upon study inclusion criteria, covariates, and the optimal number of strata within the research team.

Lesson 2: Need for Buy-In From Recruiters

In the present study, the recruitment team underwent a reorganization in personnel right before the time that the recruitment method was ready to be communicated (after the research team had identified strata, etc.). This reorganization in staffing provided an opportunity to modify the recruitment processes to incorporate the stratified recruitment method. Overall, this team consisted of seven recruiters, all of whom had previously recruited participants as a part of other educational research studies.

Previous to this stratified sample recruitment design, recruiters were typically given a target sample size (e.g., 40 schools), with the only priority being achieving the total sample within the allotted time period. In theory, this new plan added additional constraints—requiring specific numbers of schools within strata—thus possibly making the recruitment job more difficult than before. For this reason, the study team determined that communicating the generalizability goals effectively to the recruiters would be central to motivating them to implement the stratified recruitment methods well.

Simply providing the recruiters with targets for different strata was not, on its own, effective—in fact, the recruiters initially pushed back, arguing that these additional goals were simply not possible. In reaction to this, a meeting was held with the recruiters and the methodologists, with the goal of making clear the goals of generalizability. Recruiters asked many questions about the value of the recruitment method, as well as potential roadblocks and strategies to incentivize participation. This dialogue helped to create shared understanding of the importance of the recruitment method in attaining the research goals. This meeting turned out to be central to this work, since based upon it, the recruiters indicated that they felt both supported and motivated to produce a study with meaningful results.

Our informal interviews with the recruitment team suggested that it was generally not challenging to understand the rationale behind the need for stratum information in recruiting. Recruiters indicated that it was "easy to understand the point," and why the targets were different for different strata.

While the rationale was not difficult, one recruiter suggested that dictating the recruitment method could result in recruiters feeling micromanaged or that their expertise is not valued. In order to avoid these potential issues, the recruiter suggested framing the method as a tool for prioritizing recruitment efforts, as it was in the present study, rather than a restriction imposed on how they do their job. Such attention to the way in which the method is communicated and the goals framed may support its easier adoption.

Lesson 3: Strategizing is a Dynamic Process

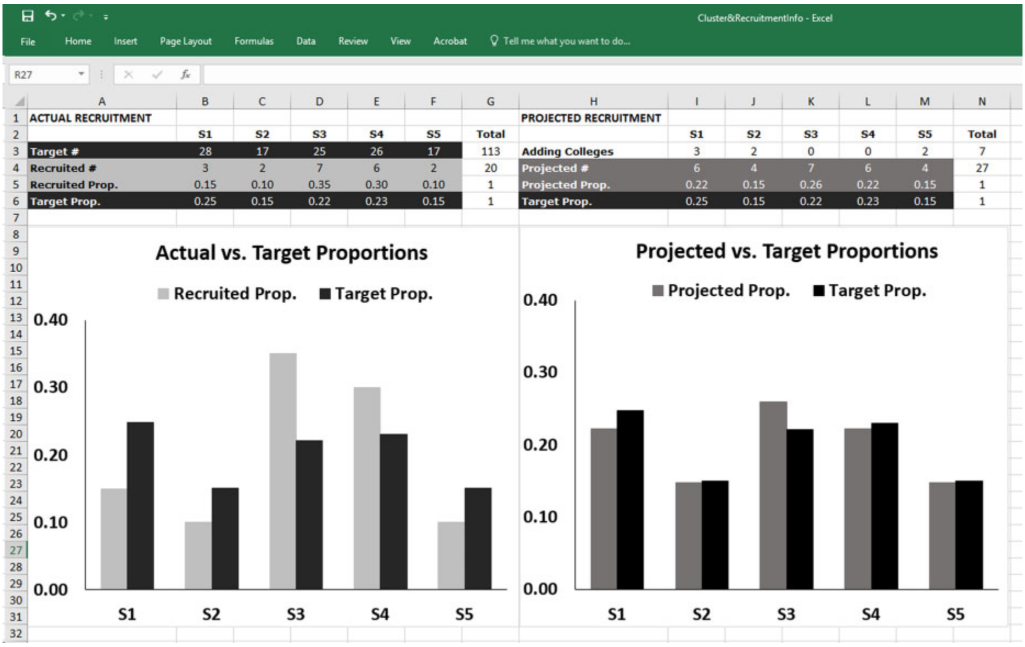
After the meeting with recruiters, the team was initially provided with stratum recruitment goals and lists of schools and told to begin recruiting. The recruitment team then began recruiting in Stratum 1 and Stratum 2, aiming to meet those goals before moving to Strata 3–5. As this recruitment progressed, however, it became clear that this strategy would not be effective in terms of overall similarity between the sample and population.

To see why, consider a hypothetical study with three strata (A, B, and C), and recruitment goals of 10, 10, and 10, respectively, with an overall goal of recruiting 30 sites. Now imagine that recruitment begins in Stratum A and is very effective—in fact, the team is able to recruit 12 schools (instead of 10). They then move to Stratum B and again are overly successful, recruiting 12 schools (instead of 10). Here, the problem arises: When the recruitment team gets to the last stratum (C), instead of aiming to recruit 10 sites they will now only need to recruit $10 - 2 - 2 = 6$ sites in order to achieve a final sample of 30 sites (needed for adequate power). Despite good efforts, the proportion of the sample in the three strata now differ markedly from those in the population—instead of allocations of 33/33/33 as in the population, the sample is now allocated 40/40/20, with the last stratum greatly underrepresented in the experiment.

In anticipation of this problem, we learned that the recruitment process is more effective when it is communicated as a *dynamic process*, involving frequent checking of the distribution of recruited sites across the strata against the targets. In order to facilitate this process, the research team created a shared document that the recruiters updated each time a new site was recruited. This document was created using commonly used software and formulas (i.e., Microsoft Excel) and was set up to visually display the stratum proportions in the population (the recruitment targets) and their up-to-the-minute success in matching these targets. The document was also set up to allow recruiters to project the stratum proportions if new colleges joined the study from different strata. Figure 3 provides an example screenshot from this document (and we include a template of this document in the Supplemental Material, available in the online version of this article). Importantly, recruiters were instructed to check the visual representations of the current stratum proportions frequently, and recruiting priorities were adjusted according to the most up-to-date stratum proportions. This strategy helped to ensure that the sample proportions did not end up wildly inconsistent with the target proportions.

Lesson 4: Stratum Info Can Help Guide Recruitment

Tipton (2014a) suggests that an additional benefit of the stratification approach is that it can be used to *describe* the population as well, potentially helping with recruitment efforts. In the study, we followed this approach, providing descriptions of the strata and discussing with the recruitment team how these could be leveraged during recruitment (e.g., "we are really hoping that schools like yours are represented in our study so that we can appropriately generalize our findings to schools that offer lots of remedial courses").



ACTUAL RECRUITMENT						
	S1	S2	S3	S4	S5	Total
Target #	28	17	25	26	17	113
Recruited #	3	2	7	6	2	20
Recruited Prop.	0.15	0.10	0.35	0.30	0.10	1
Target Prop.	0.25	0.15	0.22	0.23	0.15	1

PROJECTED RECRUITMENT						
	S1	S2	S3	S4	S5	Total
Adding Colleges	3	2	0	0	2	7
Projected #	6	4	7	6	4	27
Projected Prop.	0.22	0.15	0.26	0.22	0.15	1
Target Prop.	0.25	0.15	0.22	0.23	0.15	1

Figure 3. Example of an excel spreadsheet that allowed recruiters to (a) view the currently recruited proportions of strata relative to their target proportions and (b) projected strata proportions if new colleges were added to the study.

In this study, it is unclear how/whether these narratives were effective. On the one hand, the method provided talking points for recruiters that helped to convey that a good deal of consideration had gone into why the college's participation would be beneficial with regard to generalizing the results. On the other hand, the recruitment staff believed that other practical variables were more important to colleges, such as whether the college had the infrastructure to support computer use needed for the WATS, or whether they were tied to using a particular curriculum that was not perceived as being aligned with the study tool. If anything, the recruitment staff believed the stratum descriptions and goals for generalization may have been most effective when talking to administrative audiences.

While the effectiveness of the descriptions was unclear, the recruiters all indicated that the stratum descriptions were helpful *to them* in "getting a picture" of the types of colleges they would be talking to. The team used these descriptions and the lists of schools to strategize and learn about the types of schools—and to think through in advance different contexts, motivations, and how the program might be framed as an opportunity to meet the needs of the college. As a whole, they felt that this process helped them to better target their message in recruitment and to anticipate problems and concerns.

Lesson 5: Stratification didn't Make Recruitment Harder

Nearly, all RCTs are conducted in convenience samples (Olsen et al., 2013). These samples are typically recruited based upon previous contacts, proximity, and the availability of large sample sizes; they are rarely recruited strategically or with awareness to an inference population (or lists of all possible sites in that population). As methodologists have begun encouraging those planning studies to think about issues of generalizability during the planning and recruitment stage, an important concern has been that doing so will not be possible given the resource constraints in practice. In this study, in fact, we began to use this approach tenuously, with concern, too, that the recruiters would find implementation of this approach too difficult.

Surprisingly, the recruiters interviewed were unanimous that the method did not create additional work or make recruiting harder (compared to their previous experiences recruiting without strata). In fact, recruiters indicated that the method was helpful as it provided clear priorities on where to invest their recruitment efforts. In practice, recruiters combined their typical strategy—of leveraging relationships and local knowledge—with the strata, for example, beginning recruitment efforts not always with the most "typical" college within each stratum,[4] with those where they had contacts. In other words, the recruitment method did not change recruitment techniques, per se—recruiters utilized the skills that optimize the chances of success, including leveraging prior contacts, exercising strong social-relational skills, and so on—rather, the method provided guidance on which sites to pursue next. Where the strata were most useful, then, was recruiting colleges after recruiters had worked through their list of usual colleges. In typical recruitment efforts, this part of recruitment would have been very unstructured—and having the targets, descriptions, and goals afforded the recruiters more guidance during this process.

That said, while stratification didn't make recruitment itself harder, it also didn't make recruitment easier. The study still encountered difficulties achieving adequate sample size, and ended up recruiting two cohorts as a result (a process that is common in randomized trials). In addition, recruitment in some strata was more difficult than in others. For example, recruiters more readily recruited colleges in Strata 3 and 4 than in Strata 2 and 5.

An additional benefit, however, of the recruitment approach is that with information on recruitment tracked, the team can conduct later analyses to better understand the relationship between recruitment, attrition, and the strata. This information may be useful in answering question such as, is attrition higher in strata that are harder to recruit? And, although in the present study very few colleges were not contacted or outright refused participation, future studies could conduct analyses to determine the extent to which sites that agree to be in the study differ from those that were recruited (e.g., Tipton et al., 2016).

Conclusion

Research design in applied sciences involves a balancing act between adhering to methodological ideals and dealing with the practical realities involved in carrying out the research. In general, there are many resources that provide guidance on how to design, analyze, and report education studies (e.g., What Works Clearinghouse, 2014). However, there are fewer resources that provide guidance on how to *implement* these methods. The goal of the present work was to provide a case study for how a stratified recruitment method had been implemented in a previous education RCT, and thus to extrapolate lessons for others conducting RCTs. We have concretely illustrated each step of this approach, making clear the challenges that arose in the course of its implementation. We have particularly emphasized how communication of the approach evolved within the complex ecosystem of research teams, which are often made up of members with varied backgrounds and expertise. Our lessons learned exemplify how clear communication between teams is a driver of how successfully the method is actually enacted.

Overall, the method was successful and resulted in a sample of community colleges that represented well the population of community colleges in California on over 30 characteristics. Given the known difficulties in recruitment in RCTs, in the beginning, this goal of representing a population seemed an ambitious and possibly unattainable aim. However, despite a steep learning curve, over the course of the study, we were pleased to find the goal was achievable. By providing some practical guidance and lessons learned on how to implement the stratified recruitment method, we hope to encourage others conducting RCTs to include concerns with generalization in their study designs. Doing so requires a shift in organizational practice and, as we have shown in this article, ongoing feedback and a focus on communication within and across project teams are essential. Furthermore, this approach requires strategizing in terms of recruitment practices, in essence turning a process that was previously ad hoc into one that is increasingly systematic. We hope this article will encourage others as well to share their on the ground experiences with recruitment and generalization.

Authors' Note

Elizabeth Tipton is now at Northwestern University. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education.

Acknowledgments

We thank Shandy Hauk for helpful comments on an earlier version of this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140340 to WestEd.

ORCID iD

Elizabeth Tipton https://orcid.org/0000-0001-5608-1282

Supplemental Material

Supplemental Material is available in the online version of this article at http://journals.sagepub.com/home/aje.

Notes

1. We thank a reviewer for noting that these trade-offs occur not just in RCTs seeking to generalize but also in the definition of target populations in sample surveys as well. For example, see Weitzman, Guttmacher, Weinberg, & Kapadia, 2003.
2. Data for satellite sites and their parent campuses were averaged for each variable.
3. When information on variables were missing in the fall 2014 data, spring 2014 data were used.
4. As indicated by ranking of college within strata based on similarity to the stratum average, which was provided to recruiters by the methodologists.

References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*, 4–16.

California Community Colleges Chancellor's Office. (2012). College level report. Table B8–credit/noncredit math basic skills FTES by age categories in 2010–11. Retrieved from http://extranet.cccco.edu/Portals/1/TRIS/Research/Accountability/Basic%20Skills/2012/Table%20B8_FTES_Math.pdf

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*, 237–248.

Fellers, L. (2016). Does IES funded research represent US schools well? An evaluations of issues of generalizability in grant funded research between 2005–2014. Doctoral dissertation, New York, NY: Columbia University.

Grubb, W. N., & Gabriner, R. (2013). *Basic skills education in community colleges: Inside and outside of classrooms*. New York, NY: Routledge.

Hindes, V. A., Hom, K., & Brookshaw, K. (2002). Making WAVES. *Proceedings of the Annual International Conference on the First Year Experience*.

Long, M. C., Iatarola, P., & Conger, D. (2009). Explaining gaps in readiness for college-level math: The role of high school courses. *American Education Finance Association*, *4*, 1–33.

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.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roschelle, J., Feng, M., Gallagher, H., 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.

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

Sowers, N., & Yamada, H. (2015). Pathways impact report. Retrieved September 1, 2015, from https://www.carnegiefoundation.org/wp-content/uploads/2015/01/pathways_impact_report2015.pdf

Stuart, E. A., Ackerman, B., & Westreich, D. (2017). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*.

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*, 369–386.

Stuart, E. A., & Rhodes, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation Review*, *41*, 357–388.

Terrell, J. H., & Bugler, D. (2018). If you build it, will they come? Lessons learned in recruiting students for randomized controlled trials in postsecondary settings. *Educational Research and Evaluation*, *23*, 272–289.

Tipton, E. (2013). Improving generalizations from experiments using propensity score 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*, 478–501.

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

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Ruiz de Castilla, V. (2016). Site selection in experiments: A follow-up evaluation of site recruitment in two scale-up studies. *Journal of Research on Educational Effectiveness*, *9*, 209–228.

Tipton, E., Hedges, L. V., Vaden-Kiernan, M., Borman, G. D., 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*, 114–135.

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

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*, 516–524.

Tipton, E., & Peck, L. R. (2017). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review*, *41*, 326–356.

Weitzman, B. C., Guttmacher, S., Weinberg, S., & Kapadia, F. (2003). Low response rate schools in surveys of adolescent risk taking behaviours: Possible biases, possible solutions. *Journal of Epidemiology & Community Health*, *57*, 63–67.

What Works Clearinghouse. (2014). *Procedures and standards handbook (Version 3.0). National Center for Education Statistics*. Washington, DC: Institute of Education Sciences.