RANDOM ASSIGNMENT WITHIN SCHOOLS: LESSONS LEARNED FROM THE TEACH FOR AMERICA EXPERIMENT

Abstract
Randomized trials are a common way to provide rigorous evidence on the impacts of education programs. This article discusses the trade-offs associated with study designs that involve random assignment of students within schools and describes the experience from one such study of Teach for America (TFA). The TFA experiment faced challenges with recruitment, randomization of students, and analysis. The solutions to those challenges may be instructive for experimenters who wish to study future interventions at the student or classroom level. The article concludes that within-school random assignment studies such as the TFA evaluation are challenging but, under the right conditions, are also feasible and potentially very rewarding in terms of generating useful evidence for policy.

Steven Glazerman
Mathematica Policy Research
Washington, DC 20024
sglazerman@mathematica-mpr.com
1. INTRODUCTION: EXPERIMENTATION IN EDUCATION

Education, like any other area of public policy, is characterized by big debates with strongly opposed advocates on either side. Does school choice improve school quality? Can merit pay improve teacher quality? Do alternatively certified or prepared teachers outperform traditionally prepared and certified teachers? In order for empirical evidence to influence these debates, it must be generated in a way that adheres as strictly as possible to scientific methods. The scientific method is based on establishing causal claims through experimentation, where the researcher holds constant a set of conditions while varying key parameters one at a time.

One of the most convincing ways to establish a causal claim is the randomized experiment, in which study subjects are assigned at random to different intervention groups and all the relevant outcomes for those groups are measured and compared after some reasonable interval. If the random assignment is done correctly, the only systematic difference between the groups should be their status. As the number of experimental units increases, the sources of nonsystematic variation—chance error—become trivial relative to the true causal difference that the experimenter seeks to identify. Such is the logic underlying many education experiments, where students, teachers, or schools are randomly assigned to implement an intervention or not.

What distinguishes education from many other areas of social experimentation is the fact that students learn in groups (classrooms), which are themselves nested in schools and school districts. Education interventions themselves vary considerably by whether they try to influence the educational process at the student, classroom, teacher, school, or even district level. Thus experiments need to be designed in such a way as to disrupt the normal education process as little as possible while exerting experimenter-led variation in the policy variable at the right level of analysis.

This article is about the lessons learned from conducting an experiment within schools at the classroom level. For this type of study, the strongest design was for the researcher to assign students or rosters of students to teachers using a randomization device. Because this approach assigns students by lottery to teachers, we refer to it as a within-school random assignment (WSRA) design. Another class of education experiment, between-school random assignment (BSRA) design, assigns each school to one treatment status or another. Figure 1 illustrates the difference between the two approaches.

WSRA experiments have a long tradition in experimental psychology. For example, the Journal of Experimental Child Psychology, which began publication in 1964, routinely published findings from randomized experiments conducted within schools. (See, e.g., Caron 1968; Whitehurst 1969; and Small and Lucas 1971.) Many of these studies tended to be small or focused...
on interventions that could be administered in small controlled doses to children, where the school setting itself was not a major variable in the experiment.

In the program evaluation literature, however, the proliferation of WSRA studies is a more recent phenomenon. One of the best-known within-school random assignment studies is the class size experiment in Tennessee known as Student/Teacher Achievement Ratio (STAR) (Finn and Achilles 1990). More recent examples include a study of National Board certification, a credential for accomplished veteran teachers, conducted in Los Angeles (Cantrell et al. 2008) and a national study of alternate routes to teaching (Constantine et al. 2009). Many more such studies are in the field today.

2. THE TFA EXPERIMENT

Need for the Study
This article describes the lessons learned from one WSRA experiment in particular, an evaluation conducted in 2000–2004 of the impact of Teach for America (TFA) on student achievement and other outcomes (described in Decker, Mayer, and Glazerman 2004; also found in Glazerman, Mayer, and Decker 2006). TFA is a program that recruits prospective teachers primarily among top undergraduate institutions, provides training over the summer, and places them in under-resourced schools. It has been controversial because it allows recent college graduates without specific coursework or training in
pedagogy other than an intensive five-week summer institute that TFA provides to enter the teaching profession as lead classroom teachers. Critics have claimed that the program harms children by exposing them to underprepared teachers (Darling-Hammond 1994, 1996). Others have countered that traditional teacher preparation and certification routes may not necessarily increase quality and could keep talented individuals from entering the profession (Ballou and Podgursky 1998). By the time TFA had been operating for about ten years and had been serving fifteen regions (districts or clusters of districts) around the country, both sides of the debate had evidence to support their claims. Laczko-Kerr and Berliner (2002) analyzed data from five unnamed districts in Arizona and concluded that policies like TFA “appear harmful.” Using data from Houston, Raymond, Fletcher, and Luque (2001) found a mix of positive and statistically insignificant impacts, with lower variability in performance among TFA than other teachers. The authors concluded that TFA is a “viable and valuable source of teachers” (p. xiii). Darling-Hammond et al. (2005) reanalyzed the Houston data and found that teacher certification was the key factor that distinguished effective from ineffective teachers, with TFA teachers comparing more favorably to non-TFA teachers after they became certified. Each of those studies used nonexperimental methods. What happens when there are conflicting studies, high stakes, and no clear consensus on policy implications? This is precisely the situation in which a randomized experiment can be most helpful.

**Design of the Study**

In 2000, Mathematica Policy Research designed a study that would test the impact that TFA teachers had on their students, specifically the outcomes relative to what the impact would have been had the students been taught by the non-TFA teacher who would have been hired in the absence of the program. After conducting a one-year feasibility study, the research team mounted a pilot in 2001–2 in Baltimore city schools and then expanded the study nationwide in 2002–3.

The approach was simple: obtain a list of students entering each grade within each study school that had at least one TFA teacher and at least one non-TFA teacher and randomly assign the students to teachers. This will produce two parallel sets of classrooms—TFA and non-TFA—that are not systematically different in any way except for whether the teacher was a TFA teacher. Because we conducted random assignment within grades within schools, each combination of grade and school was considered a “block” in a block-randomized design (Cox 1958). A comparison of outcomes like test scores at the end of the year for both sets of classrooms within a randomization block should yield unbiased estimates of the impact of being
assigned to a TFA teacher. Aggregating across blocks provides an estimate of the overall impact. The Mathematica study administered new tests in all study grades in the fall and spring, using the fall test as a baseline. Teachers were also surveyed to gather background information and data on classroom management.

The TFA teachers in the evaluation were those who were already teaching in the study schools or expected to teach during the study year (2001–2 in the pilot district and 2002–3 in the other districts). Because TFA requires a two-year commitment, the study population was a mix of first- and second-year TFA teachers, with some who had more experience beyond their initial commitment. Forty-three percent of the final sample was made up of TFA teachers in their first year of teaching, with another 43 percent in their second year, 7 percent in their third year, and the remaining 7 percent with more experience (Decker, Mayer, and Glazerman 2004).

The non-TFA teacher in this design represents an approximation of the counterfactual state of the world, which cannot be observed. That is, for the students who are taught by TFA teachers, what would their experience have been if there had been no TFA program? An open question arose as to the most appropriate non-TFA teacher to approximate this counterfactual: the newly hired non-TFA teacher or the other teachers who were already teaching in the same schools. In both cases the chosen teacher might in fact be better than the marginal teacher who would have been hired but was not brought on, which means that the impact estimates from this study design may underestimate the true effect of TFA teachers.

Regardless, both types of non-TFA teachers were working with the uniquely disadvantaged population that TFA seeks to serve, and the characteristics and performance of these teachers had not been well documented to date. Schools recruited for the study typically had two or more pairs of mixed teaching teams (with both TFA and non-TFA). Because policy makers may be interested in both types of comparisons—TFA versus the new hire in the same school or TFA versus the mix of teachers who are already in the school—the study included schools that planned to have at least one grade level in which a TFA teacher was going to teach for the school year being studied, and there was at least one vacancy in the same grade. As noted below, the study focused on grades 1–5.

The Mathematica study focused on elementary schools in five urban regions (Baltimore, Chicago, New Orleans, Houston, and Compton) and one rural region (the Mississippi Delta). The final study sample included one hundred classrooms across eighteen schools. There were thirty-seven school-grade combinations (randomization blocks).
Study Findings

The study found that students of TFA teachers performed at least as well in reading and higher in math than students assigned to non-TFA teachers. The positive impact on math scores was equivalent to an effect size of 0.15, which translated to about 10 percent of a grade equivalent. The estimated impact on reading scores was very close to zero and was not statistically significant. The point estimate for reading scores corresponds to an effect size of 0.03. Both math and reading results were robust to a variety of specifications and sample definitions. Table 1 summarizes some of the main test score results and subgroup analyses.

The pattern of impacts was similar for different subgroups of teachers. With the block randomized design it was possible to isolate the blocks where the counterfactual could be defined more specifically using novice controls, veteran controls, etc. For novice control teachers, for example, the impact on math was even larger (effect size = 0.26), and the impact on reading was still not significantly different from zero. For the comparisons involving a certified teacher only or an experienced teacher only, the estimated math impacts were still positive and statistically significant, and the reading impacts were still not significantly different from zero.

With the groups having been assigned at random, it was possible to measure other outcomes as well. The study measured grade retention, referral to summer school, absenteeism, and disciplinary incidents and found no significant differences between students of TFA versus non-TFA teachers. However, teacher self-reports (shown in table 2) revealed a significant difference in the percentage of teachers that believed physical conflicts among students were a problem (34 percent of TFA teachers versus 17 percent of control teachers) and the number of times teachers had to interrupt class to deal with student disruptions (24 percent of TFA teachers versus 14 percent of non-TFA teachers).

3. STUDY DESIGN CONSIDERATIONS

Within-School and Between-School Random Assignment

Conducting a randomized experiment to test the impact of a program like TFA posed some special challenges. The schools with which TFA was working were so unusual that it would have been difficult to generate a credible control group by finding separate schools that did not have a TFA teacher and using those schools as the point of comparison. Furthermore, it would not have been technically feasible to randomly assign teachers to schools because at the time there were more TFA candidates to place than there were schools able to accept them. If there were, say, one hundred schools with teaching vacancies
Table 1. Impacts of TFA Teachers on Student Test Scores

<table>
<thead>
<tr>
<th>Analysis Sample</th>
<th>Mathematics</th>
<th>Reading</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Impact Estimate</td>
<td>Standard Error</td>
<td>Blocks</td>
</tr>
<tr>
<td>Full Sample</td>
<td>2.43***</td>
<td>(0.73)</td>
<td>0.56</td>
</tr>
<tr>
<td>Teacher experience</td>
<td></td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>Novice TFAs versus novice controls</td>
<td>4.13***</td>
<td>(1.24)</td>
<td>1.06</td>
</tr>
<tr>
<td>All TFAs versus veteran controls</td>
<td>2.71***</td>
<td>(0.97)</td>
<td>0.45</td>
</tr>
<tr>
<td>Teacher certification</td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>All TFAs versus certified controls</td>
<td>1.92*</td>
<td>(0.94)</td>
<td>0.01</td>
</tr>
<tr>
<td>All TFAs versus uncertified controls</td>
<td>3.12**</td>
<td>(1.11)</td>
<td>1.01</td>
</tr>
<tr>
<td>Student gender</td>
<td></td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Girls</td>
<td>2.83***</td>
<td>(0.96)</td>
<td>0.14</td>
</tr>
<tr>
<td>Boys</td>
<td>1.95*</td>
<td>(1.02)</td>
<td>0.71</td>
</tr>
<tr>
<td>Student race/ethnicity</td>
<td></td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>African American</td>
<td>2.06</td>
<td>(1.81)</td>
<td>0.30</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1.89</td>
<td>(1.34)</td>
<td>2.10</td>
</tr>
<tr>
<td>Student initial achievement</td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Low</td>
<td>2.32**</td>
<td>(1.11)</td>
<td>0.51</td>
</tr>
<tr>
<td>Middle</td>
<td>2.08</td>
<td>(1.37)</td>
<td>-0.54</td>
</tr>
<tr>
<td>High</td>
<td>2.27*</td>
<td>(1.33)</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Note: Impact and standard error estimates are expressed in effect sizes. Impact estimates are regression-adjusted. The benchmark regression model controls for baseline test scores, gender, race/ethnicity, eligibility for free or reduced-price lunch, age (whether overage for grade), and percentage of students in the classroom who were not in the research sample.

*Significantly different from zero at the .10 level, two-tailed test; **Significantly different from zero at the .05 level, two-tailed test; ***Significantly different from zero at the .01 level, two-tailed test.
Table 2. Impacts on Teacher Reports of Classroom Problems

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Control Mean&lt;sup&gt;a&lt;/sup&gt;</th>
<th>TFA Mean</th>
<th>Impact</th>
<th>p-value</th>
<th>Number of Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher reports a serious problem with attendance/tardiness (percentage)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student tardiness</td>
<td>12.9</td>
<td>16.2</td>
<td>3.3</td>
<td>0.669</td>
<td>96</td>
</tr>
<tr>
<td>Student absenteeism/class cutting</td>
<td>8.6</td>
<td>17.1</td>
<td>8.6</td>
<td>0.237</td>
<td>96</td>
</tr>
<tr>
<td>Teacher reports a serious problem with behavior (percentage)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical conflicts among students</td>
<td>17.1</td>
<td>34.3</td>
<td>17.1&lt;sup&gt;*&lt;/sup&gt;</td>
<td>0.073</td>
<td>96</td>
</tr>
<tr>
<td>Verbal abuse of teachers</td>
<td>4.3</td>
<td>14.3</td>
<td>10.0</td>
<td>0.107</td>
<td>96</td>
</tr>
<tr>
<td>General misbehavior (for example, students talking in class, refusal to follow classroom rules)</td>
<td>22.9</td>
<td>30.0</td>
<td>7.1</td>
<td>0.460</td>
<td>96</td>
</tr>
<tr>
<td>Problems in the most recent week (average number)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Students tardy or absent without excuse</td>
<td>4.5</td>
<td>6.6</td>
<td>2.1</td>
<td>0.108</td>
<td>94</td>
</tr>
<tr>
<td>Teacher interrupted class to deal with student disruptions</td>
<td>13.7</td>
<td>24.0</td>
<td>10.2&lt;sup&gt;*&lt;/sup&gt;</td>
<td>0.061</td>
<td>94</td>
</tr>
<tr>
<td>Teacher sent child out of the room</td>
<td>2.4</td>
<td>2.6</td>
<td>0.3</td>
<td>0.795</td>
<td>95</td>
</tr>
</tbody>
</table>

Source: Glazerman, Mayer, and Decker 2006.

<sup>a</sup>Control group means and impacts are regression adjusted. The regression model controls for baseline test scores, gender, race/ethnicity, eligibility for free or reduced price lunch, and age (whether overage for grade), as well as percentage of students in the classroom who were not in the research sample.

<sup>*</sup>Statistically significant at the 0.10 level, two-sided test.
and fifty TFA teachers to place, such a design would be attractive. Without the possibility of assigning TFA teachers (treatment) to schools, it became natural to consider assigning students to TFA teachers within schools.

Within-school study designs are not always necessary or desirable. On the checklist of concerns for any randomized experiment is the possibility of contamination of the control group. If teachers work closely together, share students, or the students share what they pick up from their classroom with students from other classrooms, the effects of the intervention under study would spill over into the control classrooms, and the differences between the two groups would be biased toward zero.

**Conditions Favoring Within-School Randomization**
There are several factors that make a within-school design attractive.

**Internal Validity Is More Important Than External Validity**
There is a trade-off involved with selecting a study sample in that a small homogeneous sample will have fewer confounding influences to worry about, yet it may be less likely to represent some broader population of interest to policy makers. In the case of TFA, most of the concern about prior attempts to estimate TFA impacts, which were drawn from single districts or states, centered on the internal validity and the comparability of the comparison group.

With any WSRA design, there will be a challenge recruiting schools in which random assignment is feasible, given the organization of the school in terms of how students are normally assigned to teachers. For example, schools with a single classroom per grade had to be excluded, as did schools where students moved around for reading groups and whose teachers team–taught or looped from one grade to the next, staying with their same group of students every year (unless the study was restricted to the entry grade of the loop). Once these restrictions are imposed, the researcher loses the ability to generalize to all possible school settings if it is suspected that any of these school organization features are related to the effectiveness of the intervention (TFA teachers).

**The Object of Study Is a Classroom-Level Phenomenon**
By focusing on elementary schools with self-contained classrooms, the research team believes the impact of the TFA teacher will be confined to the immediate classroom. This may not be true of all studies of teacher interventions. For example, many teacher professional development models explicitly call for teacher teams to work in concert or at least to participate in interactive
workshops together. Interventions that target individual students in ways that are not expected to affect other students can also be studied using WSRA, but such phenomena are rare. Examples might include school-based health interventions or the use of learning aides meant to be used in the home, such as calculators or videos.

**There Is a Large Variance Component Attributable to the School Level**

Empirical evidence from other studies as well as the TFA data itself suggest that many factors that determine student test scores vary considerably between schools. For example, residential segregation and neighborhood effects, unique school-specific factors such as the quality of the facility, or the principal’s leadership could exert a common effect on test scores. One measure of this between-school influence is the intra-class correlation, which is the percentage of total variance in test scores that is explained by the variation between schools. Using national survey data, Hedges and Hedberg (2007) have shown that intra-class correlations for student test scores in low socioeconomic status (SES) schools range from 0.06 to 0.15 for reading and 0.11 to 0.17 for math in grades similar to the TFA study (1–5), even when controlling for student demographic characteristics and prior test scores. In other words, conducting WSRA removes as much as 18 percent of the variation in test scores from the treatment contrast. More recently, Schochet (2008) provided an explicit framework for quantifying the trade-off in statistical precision for within- versus between-school experimental designs using variance parameters derived from the TFA experiment as well as others. Xu and Nichols (2010) provide additional empirical estimates of such variance components using state administrative data in North Carolina and Florida.

**There Is a High Fixed Cost per School**

The TFA study conducted in-school testing twice during the year. This would have been very costly if there had been many schools in the study. It was more efficient to limit the number of testing sites. The researchers could have invited all students in the study to an off-site testing location, as was done for students who left their original schools, but the participation rates with such a testing strategy are typically much lower than in-school testing and create a new transportation and time burden for test takers and their families. Other school fixed costs include providing student records and negotiating with the school leadership for access to the school and its students to conduct special testing and surveys of the teachers. For studies that use district-administered tests these considerations might not apply, unless the researchers were to add classroom observations to the data collection plan.
Conditions Favoring Between-School Randomization Design
At the same time, there are several conditions under which a WSRA design would be less appropriate than a BSRA design.

Collaboration within Schools Is Important or There Is Concern about Spillover between Classrooms
Any study of teacher effectiveness should consider the degree to which teachers collaborate in order to determine if the impacts can be isolated to specific classrooms. If teachers work together, move students back and forth between classrooms, or co-teach to an unusual degree, it would be problematic to conduct random assignment within grades as we did for the TFA study. Our focus on the elementary level, exclusion of districts that used certain school reform models that required frequent regrouping of students throughout the day or week, and interviews with district and school staff suggested that such collaborating within schools was not a concern for the TFA study. For a study of the Talent Transfer Initiative, being conducted by Mathematica in ten school districts, the goal was to understand whether assignment to treatment had any influence on collaboration, so that is an example of where it was necessary to use a between-school design.

The Intervention Is Diffuse
For the TFA study, the intervention was restricted to particular teachers. It was the placement of a teacher from a specific pathway into the classroom. Other interventions might be more diffuse, such as a curriculum change or accountability framework. In such instances the intervention is something that may be difficult or impractical to restrict to one set of classrooms within a school and exclude from others.

Allocation of Resources across Classrooms within the Same School Is Part of the Phenomenon Being Studied
In some cases the researcher might wish to learn whether there is a behavioral response to the intervention that involves the strategic allocation of resources between classrooms. For example, the Talent Transfer Initiative was meant to encourage veteran teachers who had demonstrated effectiveness in the classroom to transfer to low-achieving schools. One of the mechanisms by which such a program could have an impact is that the veteran teacher who fills a given vacancy may by design require less support than the person who might otherwise have filled that position—someone new to the teaching profession, for example—and therefore the mentoring support that the new hire would have needed can be reallocated to other teachers in the grade or
the school. Similarly, a principal might choose to assign more challenging students to the proven-effective veteran transferring in and leave the students with fewer learning and behavioral challenges to the other teachers. In order to study such a phenomenon, the classroom being directly affected as well as those that are indirectly affected all become part of the treated group, and a control group must be formed from somewhere else, notably a paired control school.

4. IMPLEMENTATION OF THE EXPERIMENT

There were several steps involved in implementing the TFA study design. Most critical were the recruitment of sample members, the randomization itself, and the monitoring of compliance with random assignment. As with any study of this type, data collection and analysis were important and contain lessons for future study as well.

Recruitment

A major challenge for the research team associated with the experimental study of TFA was the recruitment of study subjects. First we needed to identify districts and schools where TFA teachers could be found alongside non-TFA teachers who were responsible for the same types of students learning the same material. This narrowed the prospective study subjects to a set number of districts—TFA was operating in fifteen regions at the time—and within those districts to only the elementary schools with self-contained (homogeneous) classrooms (i.e., classrooms with no ability grouping or special status, such as bilingual instruction). To limit the prospects even further, we aimed to identify at least some of the schools for the study with one or more TFA teachers or vacancies that would be filled by a TFA teacher and at least one additional vacancy or first-year non-TFA teacher in the same grade, to allow for a comparison between TFA and novice non-TFA teachers.

Once we identified these prospective study sample members, we needed districts’ and schools’ cooperation in order to conduct the random assignment and their permission to conduct additional testing of their students, but school staff had little incentive to provide such cooperation. Participation in the study would mean that an outside researcher would have to be involved in the assignment of students to teachers, and time would have to be devoted to extra achievement tests. There was no added experimental intervention or funds that the school would receive, since it already had agreements in place to be given the chance to hire TFA teachers. This is a challenging recruitment exercise because school staff often pride themselves on their judgment at assigning
students to teachers, even if they claim that such assignment is random or done to achieve balance.

To secure participation we relied on monetary incentives (initially $1,000 per school in the pilot, increased to $2,000 per study grade per school) and a persuasive message about the benefits of fair allocation of students to classrooms and of the special testing for diagnostic purposes (since we tested students in the fall as well as the spring). We were able to tell principals who claimed to have an assignment process that was fair to all teachers that we could guarantee such fairness by conducting the assignment ourselves. We also reassured school principals and teachers that the student random assignment process would accommodate any special requirements or constraints that were necessary for the welfare of the students and the smooth operation of the school. For example, if there were a student whose disability required him or her to be in the classroom nearest the restroom or assigned to a teacher with special training, such requests could be accommodated on a case-by-case basis. In other words, the possibility of making exceptions to the protocol, if they were made by the research team, was worth the cost of having to exempt what turned out to be a very small number of students from the study.

Regular communication with principals was critical to ensure compliance with the study protocol and to establish trust and rapport. This communication is best conducted with at least one in-person meeting and regular follow-up with the same point of contact; otherwise the research team may get unfortunate surprises. For example, it was not uncommon for principals to forget about the study or for a principal to leave and be replaced by someone unfamiliar with the study or the commitments that his or her predecessor had made on behalf of the school. In more than one case the school went forward with roster assignments without the study team’s knowledge, and for one school there was no way to change it since the teachers and parents had all been notified. To recover from that particular situation we obtained an agreement to “just make a few changes” to the purposively assigned rosters (which the school claimed were random). We took the intact student rosters, randomly assigned them as we would have, and told the school which teachers to change. This gave the school the feeling that they had control over many of the assignments while ensuring that the final allocation was indeed random.

Two additional factors contributed to the success of the recruitment effort: the experience of having done a feasibility study and then a pilot study, and the assistance of the TFA organization. Throughout the feasibility and pilot phases of the study we learned about stakeholder concerns, refined the messaging, and built in design modifications to match the study protocol to conditions in the schools. Once the full-scale study was under way, we were able to use TFA staff and their relationships with schools and districts to gain introductions,
smooth communication, and earn trust—all critical in recruitment and the ongoing implementation of the study.

It is worth noting that negotiating permission to conduct random assignment and ongoing monitoring of the experiment’s integrity are nontrivial challenges associated with an experimental study, and they are uniquely challenging in a WSRA study. What made these steps especially challenging for the TFA study was the fact that participation in the study did not come with extra funding to implement a new intervention and that participating schools had to allow their students to be administered a special standardized test in addition to those already being administered by the district. These factors may not be at issue for future studies where program funding can be used as an inducement or where the study would rely on district-administered tests. However, the effort associated with monitoring compliance will always be required of a field trial and will always require some kind of roster validation if the assignment of students to teachers is under experimental control.

Randomization, Batch, and Rolling
Another challenge was how to conduct random assignment in this particular setting. Schools vary considerably in the timing of their assignment of students to classrooms, and even if one could identify a date for each school, the assignment process is rarely completed in one motion. Student lists change. Students who are registered for the following year may fail to show up. As a general principle, it is good to conduct random assignment as late as possible, ideally in the instant just before implementation of the program—in this case, the beginning of the school year—to minimize the chances that sample members will re-sort themselves and undermine the experimenter’s control over the assignment process. However, stakeholders often want to know their assignments as early as possible so they can prepare.

To meet these challenges, we conducted random assignment as early as the school staff needed, using a batch assignment process, and then followed up with a rolling random assignment of any students who matriculated after that point, through the end of the first week of school. Most of the new enrollees and no-shows are identified during the first two days of school. We assumed that no-shows were not influenced by their treatment status, because they were not typically informed before the start of the school year about the teacher to whom they were assigned or that teacher’s pathway into the profession.

Batch assignment was straightforward. We received a list of students with identifying information, a list of teachers, and a set of constraints, if any. The constraints are the requirements listed by the school principal or other stakeholder if such requirements were part of the negotiation over recruitment
into the study. We then sorted the students by random number and assigned them by random number to each one of the available teachers until all teachers had an equal number of students. If the constraints specified two students who were to be assigned to the same classroom (a rare request), they received the same number. If the constraint specified that students should be separated (a common request with twins), such a request was easily accommodated. In the case of twins, such assignments have the beneficial effect of further reducing the between-class variance because the same family (and possibly set of genetic traits) is then represented in both groups. In rare cases we had to accommodate exceptions that required the pairing of a particular student with a particular teacher. Such students were exempted from the study and assumed not to affect the teacher’s overall performance with the other students.

For any students enrolling after the initial batch assignment, rolling random assignment was done as follows. We gave each school a toll-free telephone number that was staffed by a member of the research team who had a randomized list of assignments to teachers for each school-grade team. School staff wanted to maintain a numerical balance so class sizes would be even, but we endeavored to remove the predictability from the assignment process. We therefore preferred to assign groups of two or three students called in at once but would allow classes to go out of balance by up to two students at any given time, with the understanding that arrivals after the cutoff date (one week after the start of classes) would always fill the smallest classroom if there were any differences.

Given possible temptations by school staff to reassign students or bypass the random assignment process, we conducted student roster validations throughout the year. If any students were found to be in a different class than their experimentally assigned one, we followed up with the principal to understand why and determine whether it was possible to have the student sent back to the correct classroom.

### Analysis and Realized Power

One of the advantages of a randomized experiment is the transparency of the analysis. Instead of relying on obscure statistical conditions or econometric results, an experimenter can compare the mean outcomes for the treatment and control groups to estimate the program impact. In reality, analysis and interpretation of experimental data are often more complicated than this simple mean difference. Covariates are often required to reduce the error variance and increase the precision of the treatment effect estimates. Study subjects may fail to receive the treatment to which they were assigned and instead be exposed to a different treatment condition. Study subjects may attrit from the study at
different rates depending on the study arm. When this happens for reasons that are specific to the study itself, there could be a pattern of missing data that biases the impact estimates. Furthermore, the study design might call for clustering of units, stratified sampling, variation in the treatment assignment probabilities, or some other constraint that requires a corresponding analytic adjustment to ensure unbiased estimation of the original parameter of interest.

The TFA experiment required estimation of fixed block effects. We randomly assigned students or groups of students within grade-school combinations, or randomization blocks. It was therefore necessary to estimate the block effects to remove their influence from the estimate of the program impact. This model uses a degree of freedom for each block but results in a study that is very efficient because the within-block variation is considerably less than the unrestricted variation in outcomes (test scores), even conditioning on covariates (prior test scores, student race/ethnicity, free or reduced price lunch eligibility, special education status, limited English proficiency status, and over age for grade). The intra-class correlations were 0.17 and 0.11 for reading and math, respectively. The resulting precision made it possible to detect impacts of at least 0.06 and 0.08 standard deviations in reading and math, respectively, using a two-sided test and a 10 percent significance level. This achieved precision was even better than the study design had anticipated. The target minimum detectable effect developed during the design phase was 0.11 standard deviations.

Another advantage to the block-randomized design is the ability to estimate impacts for treatment-defined subgroups. An example of a treatment-defined subgroup could be a treated unit where implementation was especially strong or weak. In the case of TFA, a treatment-defined subgroup might be the set of TFA teachers in their first year as opposed to their second year or being “alumni” (teaching beyond their two-year commitment). Another would be TFA teachers who had their regular teacher certification as opposed to TFA teachers who were teaching with emergency or provisional credentials. Impacts for such subgroups are usually difficult to estimate because for any given treatment unit that implemented the program in a certain way it is often impossible to tell which control unit would have implemented the intervention in that way. For the block-randomized design, we can use the a priori block definitions so that every treatment unit has its own predefined control unit, and the pair or group can be included or dropped from the sample collectively. For example, we defined blocks in the TFA experiment as either containing a first-year TFA teacher or not and estimated separate impacts for these two groups. Results of these analyses are included in table 1.

Using the same logic we can define blocks based on the characteristics of the control unit to estimate impacts for a particular counterfactual. The most
notable example in the TFA study was estimating impacts for novice control teachers. We estimated the impact within the subset of blocks in the study where the control teachers included a novice.

5. LESSONS LEARNED
One of the lessons learned from the TFA experiment is that within-school experiments that involve the random assignment of students or student rosters to teachers are challenging, but they can be both feasible and useful. Under the right conditions, such a study design can be a powerful tool in understanding the causal relationships between important policy interventions and policy outcomes. With just eighteen schools carefully selected from a national sample of districts, it was possible to achieve a level of statistical precision using within-school randomization to detect impacts as small as 0.06 standard deviations, which amounts to just two or three percentile points, depending on where the students start in the test score distribution.

This article highlights some of those conditions that make such a study design effective, including the differences between schools being relatively large and there being high fixed costs per school. It also highlights some of the challenges and potential solutions. One of the challenges includes school recruiting, for which using monetary incentives, a flexible study protocol, and regular communication with school leaders can be effective. Another challenge is conducting random assignment, which is possible given a careful combination of batch and rolling random assignment. The rolling random assignment requires hands-on involvement at the beginning of the school year. Regular periodic roster checks are useful for avoiding crossover between the treatment and control groups, or at least documenting the timing and reason for such crossover. The WSRA design would be less preferred in cases where the intervention is diffuse, not confined to a classroom, or prone to spillover between classrooms or grades within a school. In such cases BSRA may be a necessary alternative.

While it can add costs to a study’s design to conduct random assignment within schools (with its added components of recruiting complexity, communication during the randomization process, and roster validation), these steps improve the research in ways that would strengthen any study. The most costly aspects of conducting any study like this would not necessarily be a function of the study design (random assignment versus some other kind of comparison group or pre-post design) but of the data being collected. Administering student tests in schools and in off-site locations for makeup tests and school leavers, as was done in the Mathematica study, requires considerable effort, including the consent gathering and on-site proctoring as well as extra
costs associated with off-site testing of mobile students (renting space and compensating families for transportation and child care during the testing). Even studies that do not administer new tests may require other expensive data collection, such as surveys, with labor-intensive telephone follow-up or direct observation of classrooms by trained observers. In the end, the effort required to conduct a rigorous within-school experiment in education can be worth it if the design is well tailored to the phenomenon being studied. For its modest size, the Mathematica TFA experiment has had a strong influence in policy debates, shifting much of the discussion from the question of whether TFA harms students to new questions, including whether TFA teachers’ effectiveness in their early years is enough to offset their higher exit rates (Kane, Rockoff, and Staiger 2008), whether their effectiveness extends to high school (Xu, Hannaway, and Taylor 2011), and how TFA can be complemented by other strategies to attract and retain high-quality teachers to hard-to-staff schools (Levin and Quinn 2003).

I would like to thank Paul Decker, who led the original effort to design and implement the evaluation of Teach for America, and Mary Grider, both of whom provided useful comments on the article, along with Tom Downes and other participants of the Association for Education Finance and Policy annual conference.

REFERENCES


