WORKING PAPER 21

MATHEMATICA
Policy Research

WorkingPAPER

BY BRIAN GILL, JOSHUA FURGESON, HANLEY S. CHIANG, BING-RU TEH, JOSHUA HAIMSON, NATALYA VERBITSKY-SAVITZ

Replicating Experimental Impact Estimates with Nonexperimental Methods in the Context of Control Crossover

October 2013

Abstract

A growing literature examines whether and in what context nonexperimental methods can successfully replicate the results of rigorous randomized experiments. Ideally, replications focus on the "intent-to-treat" (ITT) experimental impact estimate, the most causally rigorous measure. Some previous studies have been able to replicate an ITT experimental impact estimate by identifying a nonexperimental comparison group that closely resembles the experimental control group. But field experiments often experience control-group crossover, in which some control-group subjects ultimately receive treatment. If the intervention has an impact, control-group crossover will cause a nonexperimental impact estimate to differ from an experimental ITT estimate even if the nonexperimental method produces a valid estimate of the impact of participating in treatment. This paper develops a new replication approach that allows nonexperimental methods to be tested against rigorous experimental ITT impact estimates when the experiment includes substantial control crossover. We apply this new method of replicating ITT impact estimates in measuring the effects of charter schools on student achievement, testing three nonexperimental approaches that incorporate pre-treatment baseline measures of the outcome of interest. Using the new replication approach, all three methods produce results that are nearly identical to ITT experimental impact estimates.

Acknowledgements. The authors thank Philip Gleason for helpful advice and constructive feedback.

A growing literature examines whether and in what context nonexperimental methods can successfully replicate the results of randomized experiments (Fortson et al. 2012; Glazerman et al. 2003; Cook et al. 2008). A valid replication exercise requires that the experimental method and the nonexperimental method are measuring the same thing, i.e., that they do not differ in terms of the causal mechanism or the subject population (Cook et al. 2008). Ideally, replications focus on the intent-to-treat (ITT) experimental impact estimate, because it is the most causally rigorous measure. In some cases, replicating an ITT experimental impact estimate involves nothing more than replicating the experimental control group by identifying a nonexperimental comparison group that closely resembles it on observable characteristics.

In real-world field experiments, however, study subjects who are randomly assigned to the control group do not always comply with their assignments—some subjects assigned to the control condition ultimately receive treatment. Control-group crossover to treatment is common, for example, in studies that use school admissions lotteries to evaluate the impacts of oversubscribed schools: Some lottery losers subsequently find another way to be admitted to the school. Crossover (noncompliance with random assignment) by control-group members undermines the conventional approach to replication because a nonexperimental method necessarily excludes from its comparison group subjects who are receiving treatment. In other words, there is no way for the nonexperimental approach to identify a comparison group that fully resembles the experimental control group, because there is no way for the nonexperimental comparison group to include subjects receiving treatment.

If the intervention has a true impact (positive or negative), the mismatch between experimental control group and nonexperimental comparison group that results from control crossover will cause the experimental and nonexperimental impact estimates to differ even if the nonexperimental method is capable of producing a valid estimate of the impact of participating in treatment (i.e., the effect of treatment on the treated, or TOT). The conventional replication approach cannot succeed if there is substantial control crossover and a non-zero treatment effect.

When examining policy impacts, replication exercises are particularly important, because many policy-relevant questions cannot readily be addressed with randomized experiments. Nonexperimental methods remain essential tools in policy research and evaluation—and they need to be tested against the most rigorous, gold standard, randomized experimental methods to provide confidence in their validity and rigor. Given the importance of validating nonexperimental methods, there is a clear need for a replication approach that can be used in field experiments that involve control crossover. This paper develops a new approach that allows nonexperimental methods to be tested against rigorous experimental ITT impact estimates even when the experiment includes substantial control crossover.

We apply this new method to examine whether three nonexperimental approaches—all incorporating pre-treatment baseline measures of the outcome of interest—can replicate ITT impact estimates of charter schools on student achievement when there are high rates of control crossover. Using the new replication approach, all three methods produce results that are nearly identical to ITT experimental impact estimates.

The next section explains why the ITT impact estimate from a randomized experiment (rather than a TOT impact estimate) is the appropriate standard to be used in a replication exercise, particularly when control crossover exists. In the subsequent section, we derive the nonexperimental equivalent of an experimental ITT estimate when some control students attend charter schools. The following sections describe the data and the estimation methods and present results. Finally, we discuss the implications of the findings, both for future replication exercises and for the use of nonexperimental methods when experimental data are unavailable.

Why Replicate ITT?

For purposes of causal inference, randomized experimental designs are clearly superior to alternative designs. Properly designed and implemented experiments—often difficult and expensive, and thus uncommon—produce impact estimates that support stronger causal conclusions, by creating treatment and control groups that are equivalent in expectation on observed and unobserved characteristics prior to receiving an intervention (Shadish et al. 2002; Murnane and Willet 2011). Any statistically significant difference between group outcomes can be attributed to the impact of the intervention.

In field experiments, noncompliance with random assignment is common. Some subjects randomly placed in the treatment group may decline treatment, while other subjects randomly placed in the control group may find alternate ways to receive treatment. Noncompliance routinely occurs, for example, in studies that rely on randomized admissions lotteries of oversubscribed schools to implement an experimental research design in measuring the impacts of the schools. Charter schools, in particular, often select students by lottery if they have more applicants than available spaces. Various studies have taken advantage of data from admissions lotteries to estimate school impacts (e.g., Hoxby et al. 2009; Gleason et al. 2010; Abdulkadiroglu et al. 2011; Angrist et al. 2011; Dobbie and Fryer 2011a; Dobbie and Fryer 2011b; Bifulco 2012; Tuttle et al. 2013).

When crossover occurs, rigorous ITT impact estimates will understate the intervention's TOT impact, which is often of greater interest to policymakers and other stakeholders. To estimate the TOT impact of receiving treatment, researchers typically use the random assignment as an instrument in a two-stage, least-squares analysis (Angrist et al. 1996). This provides an estimate of the TOT effect of program participation on students who comply with their random assignment.[1]

In the presence of crossover, researchers seeking to validate nonexperimental methods sometimes attempt to replicate the TOT impact estimate based on the two-stage instrumental variable (IV) approach (e.g., Abdulkadiroglu et al. 2011; McKenzie et al. 2007, as cited in Cook et al. 2008). This approach eliminates the control crossover problem, because the crossover subjects are not included in the IV-based TOT impact estimate.

[1] In charter-school studies, researchers sometimes can minimize crossover from assignment to control by carefully monitoring admissions offers from the waitlist, so that any student offered admission prior to the start of school—including students who were not offered admission at the time the lottery was conducted—is defined as a treatment student (Gleason et al. 2010; Abdulkadiroglu et al. 2011). Even when this is done, control crossover still occurs. In Gleason et al. (2010), about 6 percent of control students attended the study charter schools during the first follow-up year (see also Abdulkadiroglu et al. 2011; Dobbie and Fryer 2011b). These students might have been admitted through an administrative error, favoritism, or an admission offer after the beginning of the school year. This definitional minimization of control crossover often cannot be used for two reasons. First, schools with oversubscribed lotteries often ultimately offer admission to all students not admitted at the time of the lottery, meaning there are no randomized students who never received an offer (that is, no control students). Second, some schools do not have good records on whether the randomization order was followed when admitting students from the waitlist, after the lottery. Consequently, some studies (Abdulkadiroglu et al. 2011; Furgeson et al. 2012) define treatment as receiving an offer at the time of the lottery; students admitted off the waitlist to replace those admitted students who do not enroll become control crossover. These studies have high rates of control crossover, approaching 50 percent (Furgeson et al. 2012).

Replicating IV-based TOT impact estimates is an imperfect solution to the crossover replication challenge. The two-stage, IV-based analysis is less causally rigorous than the experimental ITT analysis. In particular, the validity of the IV-based TOT impact estimate depends on the "exclusion restriction." When control crossover exists, the exclusion restriction requires that control-group subjects who cross over to treatment must have the same average outcomes as treatment-group subjects who would have received treatment regardless of whether they had been assigned to treatment. In other words, *always-takers* should have the same average outcomes regardless of treatment assignment, because treatment assignment, independent of enrollment, should not affect outcomes.

The charter-school context provides illustrations of why the exclusion restriction might fail, invalidating IV-based TOT impact estimates. One threat to the exclusion restriction in the charter school context is that treatment always-takers might receive a different treatment than control always-takers, leading to different average outcomes. Because crossover students often receive admission offers later than treatment students, it is possible that their enrollment in the charter school begins after the start of the school year and that they experience a different treatment and consequently different outcomes than lottery winners. Data from one study (Gleason et al. 2010) indicates that some control-group crossover students received their admission offers after the school year started, sometimes as late as February.[2] Over the course of multiple years, fractional treatment becomes a more prominent issue that affects the treatment group as well as control crossovers; IV-based TOT analyses (e.g., Abdulkadiroglu et al. 2011) deal with this by measuring the treatment effect per year of treatment, even though treatment students are more likely to experience treatment early in the study period while control crossovers are more likely to experience treatment later in the study period (as a result of being admitted in a later grade). Charter schools may have different effects in early grades than later grades, but the IV-based TOT approach assumes equivalence. Finally, many studies now have multiple treatment charter schools with participants in several lotteries, and in these studies, treatment always-takers might attend different schools than control always-takers.

There is a second reason that an IV-based TOT estimate is not ideal as a standard to replicate, even if the estimate is unbiased—the impact estimates produced by IV-based TOT and nonexperimental approaches involve different groups of treatment subjects. The two-stage IV approach estimates impacts only for a specific subset of subjects, known as *compliers*: those who receive treatment only if they are randomly assigned to treatment. The approach estimates complier average causal effects (CACE). Not all subjects in treatment are compliers, since some would have enrolled in treatment even if they were not assigned to treatment (always-takers). Standard nonexperimental approaches cannot distinguish between compliers in treatment and always-takers in treatment; they estimate the impacts on all subjects who are receiving treatment. In short, when crossover occurs, an attempt to replicate IV-based TOT impacts violates one of the basic principles of replication—that the two methods should be measuring the same thing for the same subjects.

Replicating Experimental ITT Using NXP

This section describes how we use nonexperimental methods to estimate an impact that should be comparable to the gold-standard experimental ITT impact for the same subjects, as illustrated using charter schools and their admissions lotteries.

[2] Unfortunately, we cannot determine how often this occurred in our data.

In basic form, an experimental ITT estimate is simply the average outcome for treatment subjects minus the average outcome for control subjects. In the absence of crossover, replicating experimental impact estimates is therefore equivalent to identifying a group of subjects that proxies for the experimental control group (given that the treatment group is defined to be the same in experimental and nonexperimental analyses). When some control-group subjects cross over to enroll in treatment, however, replication is more complicated. It is impossible for a nonexperimental approach to replicate the part of the experimental control group that crosses over, because the nonexperimental approach finds comparison subjects only among the population that did *not* enroll in treatment schools. Nonetheless, a nonexperimental approach can attempt to replicate the experimental ITT estimate by splitting the estimate into components corresponding to the different groups of subjects who actually receive treatment. (Note that we can identify which subjects are always-takers in the control group—they are the ones in treatment—but not in the treatment group, because participating subjects are a mix of always-takers and compliers who cannot be distinguished.) Below we derive the nonexperimental equation to be used to match experimental ITT impact estimates. We use the same notation from Equation 1, plus an additional term: p_k^E= proportion of experimental group k that received treatment, where $k \in \{T, C\}$. Note that $p_T^E = p^A + p^{CP}$and $p_C^E = p^A$.

Not all of the μ parameters can be estimated. In fact, only μ_T^N and μ_C^A are observable. However, we can always estimate the *p* parameters because the percentage of always-takers, compliers, and never-takers in each treatment group is the same due to randomization, and p^A and p^N are observed (p^A is the proportion of control cross-overs, and p^N is the proportion of treatment subjects who do not enroll), enabling the calculation of p^C.

The experimental ITT estimate can be expressed as follows.

$$(2)\ ITT^{EXP} = (p^A\mu_T^A + p^{CP}\mu_T^{CP} + p^N\mu_T^N) - (p^A\mu_C^A + p^{CP}\mu_C^{CP} + p^N\mu_C^N)$$
$$= p^A(\mu_T^A - \mu_C^A) + p^{CP}(\mu_T^{CP} - \mu_C^{CP}) + p^N(\mu_T^N - \mu_C^N)$$

The next step in the derivation requires us to assume that the exclusion restriction—that is, assignment affects outcomes only through attendance—holds for the subset of the analysis group who are never-takers (i.e., who would not participate in treatment regardless of whether they are randomly assigned to treatment or control). We earlier expressed doubts about the validity of the exclusion restriction when used for purposes of converting an ITT impact estimate to a TOT estimate in the context of crossover, but the reasons for doubting the exclusion restriction in that context are related to the group of always-takers, not the never-takers. Even if the exclusion restriction fails with respect to always-takers (because among the always-takers, lottery winners may experience different treatment than lottery losers), it is reasonable to assume that the restriction holds with respect to never-takers: Never-takers do not experience treatment of any kind, so the assumption that they are unaffected regardless of treatment assignment is plausible.

If never-takers are unaffected by treatment assignment, then $\mu_T^N - \mu_C^N = 0$, so

$$(2')\ ITT^{EXP} = p^A(\mu_T^A - \mu_C^A) + p^{CP}(\mu_T^{CP} - \mu_C^{CP}).$$

In the nonexperimental approach, every treatment participant from the experimental treatment group is matched with an observationally similar subject who is not receiving treatment. Treated subjects from the experimental treatment group have the following mean outcome, which we denote by μ_T^E:

(3) $\mu_T^E = \frac{p^A}{p^A+p^{CP}}\mu_T^A + \frac{p^{CP}}{p^A+p^{CP}}\mu_T^{CP}$,

which is a weighted average of the always-taker mean and the complier mean in the experimental treatment group, with weights equal to their proportional representation among treatment enrollees in the experimental treatment group. μ_T^E is observed, but μ_T^A and μ_T^{CP} are not observed because always-takers and compliers cannot be distinguished.

Consider the matched comparison sample in the nonexperimental analysis. Let μ_D denote the mean outcome in the matched comparison sample, with the subscript "D" connoting that the students in this sample come from traditional district schools and thus are not receiving treatment.

The nonexperimental TOT impact, in terms of the above notation, is:

(4) $TOT^{NXP} = \mu_T^E - \mu_D = \left(\frac{p^A}{p^A+p^{CP}}\mu_T^A + \frac{p^{CP}}{p^A+p^{CP}}\mu_T^{CP}\right) - \mu_D.$

The matched comparison sample consists of: (1) untreated subjects (in the current study, these are students enrolled in local district schools) who are matched to treatment group always-takers, and (2) untreated subjects who are matched to treatment group compliers. Because we cannot distinguish always-takers and compliers in the treatment group, we also cannot distinguish the two subgroups of the matched comparison sample. Nevertheless, μ_D can be expressed as a weighted average of the unobservable mean outcomes of the two aforementioned subgroups:

(5) $\mu_D = \left(\frac{p^A}{p^A+p^{CP}}\mu_D^{MA} + \frac{p^{CP}}{p^A+p^{CP}}\mu_D^{MCP}\right)$,

where μ_D^{MA} is the mean outcome for untreated subjects (district students) matched to treatment group always-takers, and μ_D^{MCP} is the mean outcome for untreated subjects matched to treatment group compliers. Substituting (5) into (4) and rearranging terms gives:

(6) $TOT^{NXP} = \frac{p^A}{p^A+p^{CP}}(\mu_T^A - \mu_D^{MA}) + \frac{p^{CP}}{p^A+p^{CP}}(\mu_T^{CP} - \mu_D^{MCP}).$

The first bracketed expression in equation (6) for the nonexperimental estimate is, in general, not the same as the first bracketed expression in equation (2') for the experimental ITT estimate. If treatment has an impact, then $\mu_D^{MA} \neq \mu_C^A$ because the outcome of the treated subjects (here, charter enrollees) in the experimental control group will reflect that impact, but the matched comparison subjects' outcome will not.

However, if the untreated subjects who are matched to treatment group compliers are a perfect proxy for the experimental control group compliers (which is plausible because neither group receives treatment), then $\mu_D^{MCP} = \mu_C^{CP}$, which means:

(6') $TOT^{NXP} = \frac{p^A}{p^A+p^{CP}}(\mu_T^A - \mu_D^{MA}) + \frac{p^{CP}}{p^A+p^{CP}}(\mu_T^{CP} - \mu_C^{CP})$.

Because our objective is to transform the nonexperimental estimate to be as analogous as possible to the experimental ITT in equation (2'), we multiply (4') by (p^A+p^{CP}) to eliminate the denominators on the right-hand side of equation (6'). Moreover, as noted, $p_T^E = p^A + p^{CP}$. Therefore,

(7) $p_T^E \times TOT^{NXP} = p^A(\mu_T^A - \mu_D^{MA}) + p^{CP}(\mu_T^{CP} - \mu_C^{CP})$.

Equation (7) is very close to the experimental ITT expression in equation (2'). The only difference is that equation (7) has $(\mu_T^A - \mu_D^{MA})$ instead of $(\mu_T^A - \mu_C^A)$. As noted, this is a substantial difference because the mean outcome of always-takers in the experimental control group, μ_C^A, reflects the contribution of treatment, while the mean outcome of untreated subjects (enrolled in district schools) who are matched to treatment group always-takers, μ_D^{MA}, does not reflect any contribution of treatment since these students did not receive treatment (did not enroll in charter schools).

To resolve this discrepancy, we conduct another matching exercise in which treated subjects in the experimental control group (that is, the control group always-takers) are matched with untreated subjects who are similar on observed characteristics. Let TOT^{EC} denote the difference in outcomes between treated always-takers in the experimental control group and their matched counterparts who are not in treatment. We assume that the comparison subjects who are matched to treated subjects in the experimental control group have the *same mean outcome* as comparison subjects who are matched to always-takers in the experimental treatment group. This assumption is reasonable because both groups of untreated comparison subjects are being matched to always-takers from the lottery, and randomization means the always-takers should be equivalent at baseline. Under this assumption:

(8) $TOT^{EC} = \mu_C^A - \mu_D^{MA}$ leads to:

(9) $p_C^E \times TOT^{EC} = p_C^E(\mu_C^A - \mu_D^{MA}) = p^A(\mu_C^A - \mu_D^{MA})$.

Finally, subtracting equation (7) from equation (5) gives:

(10) $p_T^E \times TOT^{NXP} - p_C^E \times TOT^{EC} = p^A(\mu_T^A - \mu_D^{MA}) + p^{CP}(\mu_T^{CP} - \mu_C^{CP}) - p^A(\mu_C^A - \mu_D^{MA})$
$= p^A(\mu_T^A - \mu_C^A) + p^{CP}(\mu_T^{CP} - \mu_C^{CP})$
$=ITT^{EXP}$.

Therefore, $p_T^E \times TOT^{NXP} - p_C^E \times TOT^{EC}$should be similar to the experimental ITT impacts. Note that the better the matches created by the nonexperimental matching process, the more likely the equality in equation 10.

This equation has the virtue of allowing the effects of treatment on the control crossover subjects to differ from the effects on subjects randomized to treatment (in this study, lottery winners who attend charter schools). To replicate experimental ITT results nonexperimentally, we subtract the estimated nonexperimental effects on the crossover subjects (scaled by the proportion of all control subjects who are crossovers) from the effects on the treatment subjects who receive treatment (scaled by the proportion of all of the randomized treatment subjects who receive treatment).

Data

Description

We use student-level administrative data provided by state departments of education, school districts, and charter-school management organizations (CMOs). The data were collected for a study of charter schools operated by charter management organizations (CMOs) (Furgeson et al. 2012). CMOs aim to improve charter performance by leveraging well-regarded charter school models, and create and operate multiple charter schools under a common structure and philosophy. The sample includes data from four jurisdictions. The treatment schools were 12 oversubscribed middle and high schools across seven CMOs in the 2006–2007, 2007–2008, or 2009–2010 academic years. Three of the schools had oversubscribed lotteries in multiple years.

We focus on two outcomes: reading (labeled English language arts in some states) and math scores on state achievement tests one year after students enrolled in school following participation in a lottery (year 1). Where statewide statistics are available, we standardize test scores using state-level means and standard deviations for each grade and cohort. Otherwise, we use district-level means and standard deviations for test score standardization.

Student characteristics available for the experimental and nonexperimental analyses are: baseline reading and math test scores (including missing test score indicators), sex, race/ethnicity (African American, Hispanic, white/other), baseline free- or reduced-price lunch (FRPL) eligibility status, English language learner (ELL) status, special education status (IEP), and an indicator of whether a student attended a charter school in the baseline year.[3]

Diversity of Sample

The six CMOs included in the sample are quite diverse. The sample CMOs have schools in three of the four U.S. Census regions: Northeast, South, and West. Table 1 provides baseline (pre-enrollment) student characteristics for the CMOs. Prior to enrolling in the sample CMO middle schools, standardized student test scores were as low as –0.08 and –0.11 in reading and math, respectively, and as high as .63 and .53 (where zero is the average test score in the locality), respectively. Special education rates for the CMOs (measured in the year prior to enrollment) range from 5 percent to 14 percent of their enrolled students. The percentage of students who were English language learners varies from 4 percent to 33 percent. Finally, the CMOs are diverse in terms of race/ethnicity: the percentage of African American students ranges from 11 percent to 81 percent, and the percentage of Hispanic students ranges from 17 percent to 75 percent.

[3] One district did not have reliable information on students' free- or reduced-price lunch status. One district did not include information on students' ELL status.

Table 1. Baseline Statistics for Treatment Charter Schools

CMO	Reading Score	Math Score	Percentage African American	Percentage Latino
1	-0.08	-0.11	Medium	Medium
2	-0.07	0.04	Low	High
3	-0.07	0.16	Low	High
4	0.02	-0.04	Medium	Low
5	0.05	-0.03	High	Low
6	0.63	0.53	Low	Low

Notes: CMOs are ordered first by reading score (lowest to highest) and then math score. If the percentage of African American or Latino students is between 0 and 33 percent, the CMO is labeled low. CMOs with percentages between 34 and 67 percent are labeled medium, and CMOs with percentages greater than 67 percent are labeled high. Exact percentages of African American and Latino students are not reported to avoid identifying CMOs.

Experimental Method

Sample and Baseline Equivalence

The experimental sample frame consists of students who applied to an oversubscribed charter (CMO) school that used a random lottery to admit students. The treatment group is composed of applicants offered admission to a participating CMO school *at the time of the lottery*.[4] Applicants not offered admission at the time of the lottery form the control group. All students who provided consent (obtained prior to the lottery), were in the correct application grade at the time of the lottery, were randomized in the lottery, and had baseline test scores are included in the analysis. A student who applied to more than one of the sample schools could receive an offer to one of the schools even if the student was among the lottery losers at the other school(s). In these cases, we treat all schools sharing applicants as a single site.

To ensure the validity and power of the impact estimates, included sites meet each of the following criteria:

1. The overall and differential attrition rates are lower than the maximum thresholds defined by the U.S. Department of Education's What Works Clearinghouse (liberal attrition standard, Handbook version 2.1);
2. If we did not observe the lottery and consequently were unsure of the randomization validity, any difference between treatment and control average baseline test scores is less than 0.25 effect size, and demographic differences are less than 25 percentage points;[5]

[4] The enrollment rates of students admitted at the time of the lottery are substantially higher than those of students rejected at the time of the lottery, meaning that this measure provides enough random assignment so that we might plausibly expect to observe an impact. Angrist et al. (2013) used a similar approach. An approach in which assignment is based on whether students were ever admitted to a CMO school was not possible for two reasons. First, many schools with oversubscribed lotteries ultimately admitted all students who were not admitted at the time of the lottery, meaning there were no randomized students who never received an offer (that is, no control students). Second, many schools did not follow the randomization order when admitting students after the lottery.

[5] The effect size measure was Hedge's g. Relatively large baseline differences were allowed because some of the sites were small and could have moderate baseline differences even if the randomization was valid.

3. The difference between treatment and control groups in enrollment rates in the treatment schools is at least 20 percentage points (ensuring that the lottery assignment predicts treatment well enough that it would be plausible to observe an impact).

Application of these criteria left 579 treatment and 809 control students with baseline data who were eligible for the reading analysis, and 331 treatment and 574 control students with baseline data who were eligible for the math analysis. In the reading impact analysis, we excluded 52 treatment and 74 control students because we were unable to obtain outcome data or they were in the wrong grade in the outcome year,[6] leaving a final analysis sample size of 527 treatment and 735 control students. In the math analysis, we excluded 13 treatment and 26 control students for the same reasons, leaving a final analysis sample size of 318 treatment and 548 control students. Overall attrition in the reading sample was 9 percent, with no differential attrition between the treatment and control conditions. Overall attrition in the math sample was 4 percent, with a 1 percentage point difference between attrition in the treatment and control conditions. The low attrition levels in this study are unlikely to significantly bias impact estimates, according to the What Works Clearinghouse attrition standards (What Works Clearinghouse 2011).

Consistent with minimal bias, baseline statistics of observable characteristics indicate that the final treatment and control groups are very similar for both reading and math impact analyses, with no statistically significant differences (all p-values>.10). Table 2 presents baseline statistics for students included in the math and reading analysis samples.

[6] These students without outcome test scores most likely attended a private school or an independent charter school that did not provide data to their district. The students in the wrong grade in the outcome year either repeated or skipped a grade in the outcome year.

Table 2. Baseline Statistics for Treatment and Control Groups

Characteristic	T Mean	C Mean	Diff (T-C)	p-value
Reading Analysis (n= 527 T and 735 C)				
Baseline Reading Score	0.11	0.07	0.04	0.497
Baseline Math Score	0.08	0.02	0.06	0.328
Prebaseline Reading	0.16	0.08	0.08	0.183
Prebaseline Math	0.13	0.12	0.01	0.949
% Male	0.51	0.50	0.01	0.653
% Black	0.43	0.43	0.00	0.926
% Hispanic	0.49	0.51	-0.02	0.648
% White/Other	0.07	0.06	0.01	0.425
Baseline % FRPL	0.76	0.75	0.01	0.850
Baseline % LEP	0.09	0.10	-0.01	0.859
Baseline % IEP	0.06	0.08	-0.02	0.313
Baseline % Charter Attendance	0.09	0.11	-0.02	0.465
Math Analysis (n= 318 T and 548 C)				
Baseline Reading Score	0.12	0.09	0.03	0.696
Baseline Math Score	0.08	0.02	0.06	0.313
Prebaseline Reading	0.14	0.09	0.05	0.454
Prebaseline Math	0.10	0.08	0.02	0.817
% Male	0.53	0.47	0.06	0.128
% Black	0.61	0.62	-0.01	0.857
% Hispanic	0.37	0.36	0.01	0.823
% White/Other	0.02	0.02	0.00	0.883
Baseline % FRPL	0.83	0.82	0.01	0.907
Baseline % LEP	0.06	0.05	0.01	0.542
Baseline % IEP	0.06	0.08	-0.02	0.437
Baseline % Charter Attendance	0.04	0.02	0.02	0.208

Notes: These tables include only students who were included in the reading and math analyses. Students with missing reading outcome data are excluded from the reading analysis sample; likewise, students with missing math outcome data are excluded from the math analysis sample. Some students have missing baseline and pre-baseline test scores; these scores are imputed in the analysis, but imputed values are not included in this table. Reading and math test scores are standardized using the state district mean and standard deviation. All statistics are weighted to account for admission probabilities.

Experimental ITT Estimation and Weights

To estimate an experimental ITT impact, we compare outcomes of applicants offered admission at the time of the lottery to those of applicants rejected at the time of the lottery, controlling for students' previous test scores and demographic characteristics.[7] The impact estimation model is:

$$(11) \qquad y_{ij} = \propto + X_i\beta + \delta T_i + S_j\theta + (T_i \times S_j)\varphi + \epsilon_{ij},$$

where y_{ij} is the reading or math test score outcome for student i in site j; α is the intercept; X_i is a vector of student achievement and demographic characteristics (see Table 2); T_i is a binary variable for treatment status, indicating whether student i was admitted at the admission lottery; S_j is a vector of indicators identifying the site j that the student applied to (i.e., site—all schools sharing applicants—fixed effects); ε is a random error term that reflects the influence of unobserved factors on the outcome; and β, δ, θ, and φ are vectors of parameters or parameters to be estimated. Each site is defined by a common city, lottery year, and grade level, and thus the fixed effects control for these factors, and also allow student achievement scores within site to be correlated. The estimated coefficient on treatment status, δ, and the interactions between site and treatment status, φ, represents the impact of admission to a CMO school at the time of the lottery.[8] Students are weighted to account for admission probabilities (see Furgeson et al. 2012 for details).

When students are missing baseline or pre-baseline test scores, we include a missing data indicator in the model and set each missing test score to the state or district-level mean, which is zero by design. For students missing demographic variables (race/ethnicity, gender, FRPL, LEP, IEP, baseline charter status), we recode the missing values for these covariates to the mode across all students in the sample (not an English language learner, no IEP, not attending a charter school at baseline, receiving free/reduced-price lunch, female, and Hispanic). In some cases, missing data indicators could not be included because they were perfectly collinear. We do not impute outcome test scores, and students who are missing either a math or reading test score in the follow-up year are excluded from the analysis when that test score is the outcome variable.

Nonexperimental Methods

Following the derivation discussed earlier, in order to replicate the experimental ITT impact estimate, we require two sets of nonexperimental impact estimates: the impact of charter school enrollment on treatment group enrollees and the impact of charter school enrollment on control group enrollees (crossovers). The treatment group enrollees are lottery winners who attended an experimental charter school. The control-group enrollees are lottery losers who attended an experimental charter school. For both groups of enrollees, comparison groups are identified among students who did not attend an experimental charter school.

[7] As student admission to CMO schools was randomly determined, we could simply compare the mean outcomes of the treatment and control groups. However, to obtain more precise impact estimates, we adjust for baseline student characteristics in a regression model.

[8] The overall impact is estimated by $\sum_1^j weight_j \times \delta + \varphi S_j$, where *weight* indicates the weight for site j based on the number of students in the site and their admission probabilities (see Furgeson et al. 2012 for details). It's possible to estimate an overall impact without the interaction terms (that is, $y_{ij} = \propto + X_i\beta + \delta T_i + S_j\theta + \epsilon_{ij}$). We prefer to use the model with the interaction terms to estimate overall impacts to be consistent with the NXP approach that must use site-specific impact estimates. (Because the NXP estimates subtract the control enrollee impacts from the treatment enrollee impacts, impacts must be estimated individually for each site and then aggregated.) Although the no-interactions model has identical point estimates to the ones we estimate using Equation 10, it has more precision. As a sensitivity check, we estimated overall impacts using both models; the conclusions from hypothesis tests were unchanged.

There is reason for optimism about the validity of nonexperimental methods in studies of educational interventions for which test scores are the outcomes of interest. Such studies can often utilize pre-intervention test measures and include comparison students from the same community (Cook et al. 2008). Because pre-treatment test scores are highly correlated with post-treatment test scores, including these measures as covariates and/or matching variables might enable nonexperimental approaches to sufficiently account for selection into charter schools. Moreover, selecting comparison students from the same community makes the groups more similar on unobserved characteristics associated with geography. Indeed, three recent studies (Bifulco 2012; Fortson et al. 2012; Tuttle et al. 2013) that compared experimental ITT estimates to nonexperimental estimates using pre-intervention measures found results that suggest cautious optimism about the performance of nonexperimental approaches—but none of those studies included substantial control crossover.

We estimate impacts using an OLS regression model; covariates are included to improve statistical precision and to control for any remaining differences in baseline characteristics. (For the two matching approaches described below, this follows the creation of matched samples.) The regression model is identical to the model used in ITT experimental analysis (Equation 11). For the nonexperimental approaches, however, the treatment indicator, *T*, corresponds to each of the two enrollee groups. In estimating impacts, enrolled students are weighted to account for the probability of winning a lottery admission offer (replicating the experimental impact estimation). The matched comparison students are assigned the analysis weight for the enrolled students to whom they are matched. The experimental admission probability weights are rescaled so that a given site has the same weight in both the experimental and the nonexperimental approaches. This weighting ensures that any potential differences between experimental and nonexperimental estimated impacts can be attributed to the approaches themselves rather than differences in weights. To calculate an overall impact estimate, the site-specific estimates were aggregated using Equation 12.

$$(12)\ \sum_{1}^{\mathrm{j}} \mathrm{weight}_{\mathrm{j}} \left[\left(p_j^T \times TOT_j^T\right) - \left(p_j^C \times TOT_j^C\right)\right],$$

where weight indicates the weight for site j based on the number of students in the site and their admission probabilities (same for both experimental and nonexperimental), p indicates the percentage of treatment or control enrollment at site j, and TOT indicates the estimated impact for treatment or control enrollees at site j.

Propensity Score Matching

The first step in the propensity-score matching (PSM) approach is to estimate a propensity score for each student in the sample. To determine the appropriate propensity score model for each of the two enrollee groups, we use a forward model selection procedure for the logistic regression. Because baseline math and reading test scores are some of the strongest predictors of later outcomes, we specify that the model-building procedure begins with the model containing the two baseline test scores and corresponding missing test score indicators. At each subsequent step, the forward procedure adds a term from a specified set of potential covariates to optimize model fit to the data. The procedure can select from a list of 52 potential covariates: the 11 observed baseline covariates, 39 two-way interactions of these covariates, and 2 interactions of test scores with themselves (i.e., quadratic terms). These models fit the data well, as indicated by the Hosmer and Lemeshow Goodness-of-Fit test *p*-values (0.45 for treatment enrollees and 0.78 for control enrollees).

After estimating the propensity scores, we identify comparison students whose estimated propensity scores are similar to those of each treatment student (that is, comparison students who had similar probabilities of enrolling in CMO schools). The selection uses caliper matching, whereby a given treatment student is matched to all comparison students with estimated propensity scores within a specified range (or caliper), rather than merely selecting a specified number of nearest neighbors. The sampling occurs with replacement. The matching procedure is implemented separately for each jurisdiction. To improve statistical precision, we select multiple comparison students for each treatment student.

For math outcome samples, the matched comparison students on average have similar baseline (pre-treatment) math and reading test scores as the treatment students (Table 3). They also have similar distributions on all demographic covariates, with the exceptions of race/ethnicity and baseline charter school attendance.[9] The results were similar for reading outcome samples (not shown).

Table 3. Baseline Statistics for Treatment-Group Enrollees and Propensity-Score Matched Comparison Group (Math Outcome)

Prior Student Achievement or Student Characteristic	Treatment (n= 200) Mean/Percentage	Comparison (n= 5,905) Mean/Percentage	Difference
Baseline Math Score	0.10	0.07	0.02
Baseline Reading Score	0.10	0.06	0.04
Race/Ethnicity			
African American	0.58	0.30	0.28**
Hispanic	0.40	0.48	-0.08**
White/Other	0.02	0.22	-0.20**
Male	0.60	0.59	0.01
Free/Reduced-Price Lunch	0.84	0.87	-0.03
Special Education	0.05	0.05	0.00
English Language Learner	0.06	0.11	-0.05*
Attended Charter School at Baseline	0.05	0.02	0.03**

*Significantly different from zero at the .10 level, two-tailed test
**Significantly different from zero at the .05 level, two-tailed test

[9] There are only a few students who were not African American or Hispanic in the treatment group. While the race/ethnicity variable was selected by the model selection procedure, the associated coefficients had large standard errors. As a result, race/ethnicity were excluded from the propensity-score matching model, resulting in an imbalance between treatment and matched comparison students. However, race/ethnicity is a covariate in the impact estimation model. In addition, the PSM estimates are almost identical to those from the exact matching approach, which includes race/ethnicity as a matching characteristic. This suggests that our approach is robust to the exclusion of this variable from the propensity model. Similar problems occurred with baseline charter school attendance.

Table 4. Baseline Statistics for Control-Group Enrollees and Propensity-Score Matched Comparison Group (Math Outcome)

Prior Student Achievement or Student Characteristic	Treatment (n= 124) Mean/Percentage	Comparison (n=3,695) Mean/Percentage	Difference
Baseline Math Score	-0.01	0.03	-0.03
Baseline Reading Score	0.05	0.08	-0.03
Race/Ethnicity			
African American	0.56	0.32	0.25**
Hispanic	0.42	0.46	-0.04**
White/Other	0.02	0.22	-0.20**
Male	0.40	0.40	0.00
Free/Reduced-Price Lunch	0.89	0.84	0.05
Special Education	0.08	0.10	-0.02
English Language Learner	0.02	0.02	0.00
Attended Charter School at Baseline	0.01	0.02	-0.01

*Significantly different from zero at the .10 level
**Significantly different from zero at the .05 level

Exact Matching

Exact matching (EM) uses comparison group students who exactly match treatment students on a set of demographic characteristics and have very similar baseline test scores (e.g., see Woodworth and Raymond 2009). To be selected, the comparison students must exactly match the treatment students on the following categorical characteristics: baseline charter school attendance, sex, race/ethnicity, FRPL eligibility status, LEP status, IEP status, grade in outcome year, cohort, and jurisdiction. Exact matching on continuous characteristics—such as baseline math and reading test scores—would rarely identify matches, so we define a comparison student to be an exact match if his or her test score falls within 0.10 standard deviation of the treatment student's baseline test score in the same subject. We managed to find matches for 95 and 97 percent of the treatment students for the math and reading analysis samples, respectively. Following the creation of the matched comparison group, impacts are estimated using the same regression model used in the experimental and PSM analyses.

OLS Regression with Baseline Achievement Covariates

The OLS-only approach does not attempt to create a matched comparison group of students. Instead, the approach uses the entire population of non-CMO students in the local jurisdiction as comparisons, relying entirely on covariates to adjust for baseline differences between treatment students and other students. We use the same OLS regression model used to estimate impacts in all of the other approaches.

Results

The nonexperimental approaches successfully replicate average experimental ITT impact estimates for the 12 charter schools in the replication sample (Table 5). Propensity-score matching produces ITT impact estimates within 0.01 of the experimental ITT estimate in math and 0.03 of the experimental ITT estimate in reading. Neither of the differences is statistically significant, and they differ in opposite directions (i.e., they do not consistently under- or over-estimate impacts).[10] The last two rows of Table 5 show that exact matching and OLS, like propensity-score matching, produce impact estimates that are very close to experimental impact estimates.

Table 5. Experimental and Nonexperimental Impact Estimates in 12 CMO Schools

	Reading		Math	
	Impact	SE	Impact	SE
Experimental ITT	0.00	0.04	0.09**	0.04
Propensity Score ITT Based	0.03	0.06	0.08	0.07
Exact Matching ITT Based	0.02	0.06	0.08	0.07
OLS ITT Based	0.02	0.06	0.07	0.07

**Indicates statistically distinguishable from zero with 95 percent confidence.

Moreover, nonexperimental impact estimates at the site level are very similar to experimental site estimates, as the high correlations in Table 6 indicate. At the site level, PSM ITT impact estimates correlate with experimental ITT impact estimates at 0.97 in math and 0.90 in reading. Site-level results are also similar to experimental impacts for exact matching (0.96 for math and 0.90 for reading) and OLS (0.99 for math and 0.88 for reading). For each nonexperimental method, the impact estimates do not consistently under- or overestimate experimental ITT impacts (i.e., there is no evidence that they are positively or negatively biased).

Table 6. Experimental and Nonexperimental Impact Estimates at the Site Level

	Reading				Math			
Lottery Site	EXP	PSM	EM	OLS	EXP	PSM	EM	OLS
A	0.05	0.12	0.08	0.10	0.31	0.36	0.38	0.34
B	0.14	0.17	0.16	0.13	n/a	n/a	n/a	n/a
C	-0.15	-0.21	-0.25	-0.18	-0.11	-0.08	-0.11	-0.13
D	-0.10	-0.04	-0.01	-0.03	0.05	0.03	0.06	0.03
E	0.11	0.11	0.11	0.08	0.19	0.14	0.11	0.14
F	-0.06	-0.11	-0.08	-0.09	-0.03	-0.08	-0.07	-0.06
G	-0.10	-0.01	-0.04	0.00	n/a	n/a	n/a	n/a
Correlation with EXP	n/a	0.90	0.90	0.88	n/a	0.97	0.96	0.99

n/a means not applicable

[10] When calculating the NXP standard errors, we make two assumptions, each of which has an opposite effect on the size of the standard errors. First, we assume zero covariance between the sampling errors of the treatment and control enrollees' impact estimates. Because the comparison students who are being matched to the two groups overlap—especially in the OLS analysis that incorporates all students in the same grade in the district—the covariance is actually positive. This results in the over-estimation of the SEs and p-values. Second, we assume no covariance between the sampling errors of site-specific impact estimates. Several sites occur within the same district, and thus comparison students overlap between sites, especially in the OLS analysis, creating a positive covariance between the sampling errors of the site impact estimates. This results in the under-estimation of the SEs and p-values.

Conclusion

In this paper we develop a new approach to examine whether nonexperimental methods can replicate the most causally valid ITT experimental estimates when there is substantial control crossover. Our findings suggest that nonexperimental panel approaches (propensity-score matching, exact matching, and OLS regression) that follow subjects over time and incorporate pre-treatment measures of the outcome can produce impact estimates that replicate ITT experimental estimates with a high degree of accuracy. Moreover, the nonexperimental estimates are neither higher nor lower than the experimental estimates, implying no systematic bias. Experimental ITT impacts remain the gold standard due to their transparent, minimal assumptions, but the repeated validations of nonexperimental methods indicate that well-conducted nonexperimental studies with baseline measures of the outcome of interest can often achieve sufficient internal validity.

The finding that nonexperimental methods using pre-treatment measures of the outcome can successfully replicate experimental results is not novel. But this is the first study to demonstrate that replication of rigorous ITT experimental impact estimates is possible using nonexperimental approaches *even in the context of substantial control-group crossover*, a common feature of many field experiments.

The extent to which nonexperimental studies might produce unbiased estimates of impacts on outcomes for which baseline measures are unavailable remains an open question (e.g., Glazerman et al. 2003). Additional research on this issue is merited, because some outcomes of interest are not measured repeatedly and therefore cannot be included as baseline control variables. In the context of schooling, these include attainment outcomes such as high-school graduation and enrollment in college. Whether baseline test scores can adequately control for selection related to student attainment is as yet unknown—a question to be addressed in future studies that have admissions lottery data and a long post-lottery time series.

With the growth of detailed administrative data on program participants, validation of nonexperimental methods will continue to be important for the development of knowledge of the impacts in the field of complex policies and interventions. In education and many other policy arenas, randomized experiments are unlikely to be able to address many critical policy questions. In some instances, exclusive reliance on randomized experimental results could lead to mistaken conclusions. For example, charter schools that are sufficiently oversubscribed to use admissions lotteries could be an unusually effective group of charter schools (as suggested in Abdulkadiroglu et al. 2013). Small-scale experiments may fail to capture systemic effects that become evident only at scale, as, for example, when the state of California used the results of a class-size reduction experiment to motivate a statewide class-size reduction policy, failing to anticipate unintended teacher labor-supply effects that occurred only as a result of large-scale implementation (see Stecher and Bohrnstedt 2001). This paper should help future studies to validate nonexperimental methods that will be important in policy research, in various contexts.

Bibliography

Abdulkadiroğlu, A., J. D. Angrist, S. M. Dynarski, T. J. Kane, and P. A. Pathak. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *The Quarterly Journal of Economics,* vol. 126, no. 2, 2011, pp. 699–748.

Angrist, J. D., S. A. Cohodes, S. M. Dynarski, P. A. Parak, and C. D. Walters. "Charter Schools and the Road to College Readiness: The Effects on College Preparation, Attendance, and Choice." Boston: The Boston Foundation and the NewSchools Venture Fund, 2013.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association,* vol. 91, no. 434, 1996, pp. 444–455.

Angrist, J. D., P. A. Pathak, and C. D. Walters. "Explaining Charter School Effectiveness." NBER Working Paper 17332, 2011.

Bifulco, R. "Can Nonexperimental Estimates Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison." *Journal of Policy Analysis and Management,* vol. 31, no. 3, summer 2012, pp. 729–751.

Cook, T. D., W. R. Shadish, and V. C. Wong. "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Public Analysis and Management,* vol. 27, no. 4, 2008, pp. 724–750.

Dobbie, W., and R. G. Fryer. "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics,* vol. 3, no. 3, 2011a, pp. 158–187.

Dobbie, W., and R. G. Fryer. "Getting Beneath the Veil of Effective Schools: Evidence from New York City." NBER Working Paper 17632, 2011b.

Fortson, K., P. Gleason, E. Kopa, and N. Verbitsky-Savitz. "Horseshoes, Hand Grenades, and Treatment Effects? Reassessing Bias in Nonexperimental Estimators." Oakland, CA: Mathematica Policy Research, March 2013.

Fortson, K., N. Verbitzy-Savitz, E. Kopa, and P. Gleason. "Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates (NCEE Technical Methods Report 2012-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, April 2012.

Furgeson, J., B. Gill, J. Haimson, A. Killewald, M. McCullough, I. Nichols-Barrer, B. Teh, N. Verbitsky-Savitz, M. Bowen, A. Demeritt, P. Hill, and R. Lake. "Charter-School Management Organizations: Diverse Strategies and Diverse Student Impacts." Mathematica Policy Research, January 2012.

Gill, B., P. M. Timpane, K. E. Ross, D. J. Brewer, and K. Booker. *Rhetoric Versus Reality: What We Know and What We Need to Know about Vouchers and Charter Schools*. Santa Monica, CA: RAND, 2007.

Glazerman, S., D. M. Levy, and D. Myers. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *The ANNALS of the American Academy of Political and Social Science,* vol. 589, no. 63, 2003, pp. 63–93.

Gleason, P., M. Clark, C. C. Tuttle, C. C., and E. Dwoyer. "The Evaluation of Charter School Impacts: Final Report" (NCEE 2010-4029). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2010.

Hoxby, C. M., and S. Murarka. "Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement" (No. w14852). NBER Working Paper 14852, 2009.

Hoxby, C. M., S. Murarka, and J. Kang. "How New York City's Charter Schools Affect Achivement, August 2009 Report." Second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project, September 2009.

Lalonde, R. "Evaluating the Econometric Evaluations of Training with Experimental Data." *American Economic Review,* vol. 76, 1986, pp. 604–620.

McKenzie, D., J. Gibson, and S. Stillman. "How Important Is Selection? Experimental Versus Nonexperimental Measures of Income Gains from Migration." Washington, DC: World Bank., 2007.

Murnane, R. J., and J. B. Willet. *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* New York: Oxford University Press, 2010.

Shadish, W. R., T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin, 2002.

Smith, J., and P. E. Todd. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics,* vol. 125, no. 1, 2005, pp. 305–353.

Stecher, B., and G. Bohrnstedt. "Class-Size Reduction in California: A Story of Hope, Promise, and Unintended Consequences." *Phi Delta Kappan,* vol. 82, May 2001, pp. 670–74.

Tuttle, C. C., P. Gleason, and M. Clark. "Using Lotteries to Evaluate Schools of Choice: Evidence from a National Study of Charter Schools." *Economics of Education Review,* vol. 31, no. 2, 2012, pp. 237–253.

Tuttle, C. C., B. Teh, I. Nichols-Barrer, B. P. Gill, and P. Gleason. "Student Characteristics and Achievement in 22 KIPP Middle Schools." Washington, DC: Mathematica Policy Research, 2010.What Works Clearinghouse. *Procedures and Standards Handbook version 2.1.* U.S. Department of Education's Institute of Education Sciences, 2011.

Woodworth, J. L., and M. E. Raymond. *Charter School Growth and Replication.* Stanford, CA: CREDO, 2013.

About the Series

Policymakers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers policymakers and researchers access to our most current work.

For more information about this paper, contact Brian Gill, senior fellow, at bgill@mathematica-mpr.com.

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

PRINCETON, NJ - ANN ARBOR, MI - CAMBRIDGE, MA - CHICAGO, IL - OAKLAND, CA - WASHINGTON, DC

MATHEMATICA
Policy Research