Educational Evaluation and Policy Analysis
June 2016, Vol. 38, No. 2, pp. 318–335
DOI: 10.3102/0162373715617549
© 2016 AERA. http://eepa.aera.net

Estimates of External Validity Bias When Impact Evaluations Select Sites Nonrandomly

Stephen H. Bell
Robert B. Olsen
Abt Associates
Larry L. Orr
Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health

Evaluations of educational programs or interventions are typically conducted in nonrandomly selected samples of schools or districts. Recent research has shown that nonrandom site selection can yield biased impact estimates. To estimate the external validity bias from nonrandom site selection, we combine lists of school districts that were selected nonrandomly for 11 educational impact studies with population data on student outcomes from the Reading First program. Our analysis finds that on average, if an impact study of Reading First were conducted in the districts from these 11 studies, the impact estimate would be biased downward. In particular, it would be 0.10 standard deviations lower than the impact in the broader population from which the samples were selected, a substantial bias based on several benchmarks of comparison.

Keywords: *evaluation, impacts, external validity, generalizability, purposive sampling, convenience sampling*

Main Text

In recent years, a great deal of attention has been paid to the *internal* validity of impact evaluations of federal, state, and local programs and interventions: whether an evaluation yields an unbiased estimate of the impact of the program within the study sample. Much less attention has been paid to whether the evaluation yields an unbiased estimate of the impact of the program on the population of policy interest, a property known as external validity.

One potential threat to the external validity of impact evaluations comes from the choice of sites in which the evaluations are conducted. Impact evaluations of social programs are often carried out in multiple sites, such as school districts, housing authorities, local Temporary Assistance for Needy Families (TANF) offices, or One-Stop Career Centers. Ideally, each impact evaluation designed to inform policy decisions would include a representative sample of sites from the population of policy interest, selected at random to ensure statistical equivalence to the population except for chance differences. In practice, many impact evaluations select sites nonrandomly to meet certain criteria such as adequate capacity to support the evaluation or oversubscription to support random assignment–or to satisfy certain distributional requirements such as a mix of urban and rural sites. In addition, in most impact evaluations sites that have been selected can choose to opt out of the evaluation. All of these factors can result in an unrepresentative sample of sites participating in the evaluation.

General concerns about external validity have been expressed for many years. Shadish, Cook, and Campbell (2002) provide an overview of these threats. Concerns about external validity have also been raised in the area of clinical trials (e.g., Braslow et al., 2005; Humphreys, Weingardt, & Harris, 2007; Rothwell, 2005). While citing the advantage of experiments in producing impact estimates that are free from selection bias that can threaten internal validity, Orr (1999) notes that when conducting policy evaluations failing to obtain cooperation from sites can threaten the findings' external validity: "If the refusal rate is high among selected sites, bias can creep into the impact estimates via self-selection of sites" (p. 105). Heckman and Smith (1995) argue that the challenge of obtaining cooperation from sites in social experiments ". . . represents a systematic, institutional limitation on the use of experimental methods" (p. 104). At the same time, Imbens and Wooldridge (2009) note that many of the critiques of randomized experiments, including their frequent lack of external validity, apply to many studies without randomization.[1] Nonetheless, Banerjee and Duflo (2009) acknowledge that limited generalizability of experimental results from unrepresentative samples is a serious objection to reliance on studies of that design. This article provides one of the first analyses to address how serious a problem external validity bias might be in multisite evaluations of social programs.

We define external validity bias from nonrandom site selection (referred to as simply external validity bias from this point forward) as the difference between the average impact in the population and the average impact in a nonrandom sample of sites on which study findings are based. In prior work on external validity bias, Olsen, Orr, Bell, and Stuart (2013) have shown formally what many researchers understand intuitively—that evaluations can yield biased impact estimates for the broader population of interest if (a) impacts vary across sites in the population and (b) sites are selected or included in the study nonrandomly. Their exploration of the issue provides evidence that the vast majority of social experiments are conducted in nonrandomly selected sites and derive a formula for the external validity bias that can result from such a mechanism of site inclusion in an evaluation. However, the formula itself does not yield any insight on how large the bias might be in practice. Many papers have estimated the internal validity bias from nonexperimental comparison groups (e.g., Dehejia & Wahba, 1999; Fraker & Maynard, 1987; Heckman, Ichimura, & Todd, 1997; LaLonde, 1986; Shadish, Clark, & Steiner, 2008; Smith & Todd, 2005), in what has become known as "design replication" research. However, we are not aware of any studies that have estimated the external validity bias from conducting an impact evaluation in an unrepresentative sample of sites.

Contribution and Empirical Strategy

In this article, we provide what we believe are the first empirical estimates of the magnitude of the external validity bias that can result from nonrandom site selection for rigorous impact evaluations. To obtain such estimates, we use data that allow us to calculate both the impact of an intervention in a nonrandom subset of sites typical of impact evaluations and the impact of the same intervention in the larger population from which those sites are drawn. No single study can directly provide these data without additional assumptions or information. Studies that select sites nonrandomly can provide a measure of the impact for the included sites and can be used to explore treatment effect heterogeneity but cannot be expected to provide a measure of impact for the larger population free from external validity bias.[2] Conversely, studies that select sites randomly can provide a measure of the impact for the full population that is free from external validity bias but cannot provide a measure of the impact that would have resulted if the study had selected sites purposely, at least not without strong assumptions regarding which sites would have participated in an evaluation that used a purposive sample.

To estimate external validity bias from selective samples of sites, our empirical approach involves (a) calculating a program's impact on a defined population of interest for which data are available in all sites, (b) within that population measuring program impacts for nonrandom samples of sites that have been used in actual evaluations of educational interventions, and (c) estimating external validity bias as the difference between the measured impact in the population and the measured impacts in the nonrandom samples.

Inputs to this analysis come from two sources:

1. A study of Reading First conducted by Abt Associates (Gamse et al., 2011), which collected data on all public school districts and Title I schools in a defined population (15 states in the United States)[3] and
2. Lists of school districts that participated in 11 rigorous impact studies that selected sites nonrandomly and assessed the impacts of 11 distinct educational interventions (not including Reading First).

The data from the Reading First evaluation included *all* public school districts that were eligible for the Reading First program in the studied states. The data include both Reading First schools and non–Reading First schools, and they cover multiple years, including the year in which Reading First was first implemented. These data allow us to estimate the impact of the program in a well-defined population of states using a comparative interrupted time series (CITS) model. We use a CITS model to estimate impacts because, in our judgment, it is the strongest analytic approach we can take toward estimating the program's effect with the data available. While we would have preferred to use data from a randomized experiment, we were unable to identify any experiments that were suitable for the empirical exercise conducted for this article.[4]

The 11 rigorous impact evaluations that selected sites nonrandomly were carried out by a number of different organizations and range in size from a few public school districts to several dozen. These 11 evaluations allow us to identify the school districts that were actually included in real-world impact evaluations when sites were selected nonrandomly and school district officials could decide whether to participate. Using the same CITS model as used in the full population, we estimate the effect of Reading First for each of the 11 subsets of districts that participated in these 11 rigorous impact studies. We take the difference between the effect in the full population of school districts and the estimated effect in each of the 11 nonrandom samples of districts to produce 11 estimates of external validity bias from nonrandom site selection. Finally, to obtain an overall test of the existence of external validity bias across the 11 samples, we use meta-analytic methods.

In the remainder of the article, we provide a more complete description of both the data and methods used in the analysis. In addition, we report and interpret our estimates of external validity bias. Finally, we consider opportunities for future research.

Data

As noted earlier, our empirical strategy for estimating external validity bias requires data on a population of interest—data that can be used to measure the impact of some program both in the full population and in specific sites within that population—as well as data that identify the samples of sites used by studies with nonrandom site selection from within this population.

Data on a Population

To estimate impacts of an educational program, this article uses data that were collected for Abt Associates' most recent study of the Reading First Program, a federal program to improve reading instruction. Reading First was authorized by the No Child Left Behind Act and ended in 2008. Annual appropriations for the program were more than US$1 billion per year between 2004 and 2007. The Reading First program awarded grants to states, and states were instructed to give funding priority to Title I schools with the lowest reading proficiency and highest poverty rates.[5] States distributed funds to nearly 6,000 Reading First schools in more than 1,800 school districts. Schools that received Reading First funds were expected to use evidence-based reading curricula, offer professional development to teachers to help them implement evidence-based instruction in reading, and screen and monitor students to identify early reading difficulties (for more details, see Gamse et al., 2011).

The Reading First study obtained data from 15 states: Arizona, Arkansas, Colorado, Delaware, Florida, Georgia, Hawaii, Minnesota, Mississippi, North Carolina, South Carolina, South Dakota, Texas, Utah, and Wisconsin. Each of these states first implemented Reading First in either 2003–2004 or 2004–2005. These states were selected because they collected individual-student-level achievement and attendance data over a continuous time span that began before Reading First was implemented and continued through the 2007–2008 academic year and maintained these data in a way that enabled linkage of records for a given student over time.

Importantly for our research, the Reading First study did *not* select subsets of sites nonrandomly. Because it relied on extant data and required no active cooperation from participating schools and districts, data were collected for *all* Title I schools in the participating states. In those states, the study collected data on all students enrolled in Grades K–6 in Title I schools. This included schools that actually received Reading First funding, as well as other Title I schools that did not receive Reading First funding. Student scores on state reading assessments were collected for students in Grades 3 through 6 (and earlier grades in some states).

For the present analysis, we limit attention to a subset of the data by

- **Focusing on reading achievement in Grade 3 as the outcome of interest.** The Reading First program served students in kindergarten through Grade 3. We use scores on the state test of Grade 3 reading achievement as our outcome measure (dependent variable) because these scores should reflect the impact of the program after students had received a "full dose" of the intervention, but before any decay occurred that could reduce the effect of the program.
- **Restricting the sample to states with data for which we can measure the impact of Reading First in the population and for one or more samples from studies with nonrandom site selection.** For five states, the Reading First study did not collect data for years before Reading First was implemented, which is required for estimating the effects of the program using the CITS model described in the next section. For one additional state, we can measure the impact of Reading First in the population, but we cannot estimate its impact in any of the 11 study-specific samples (see the next section): None of those school districts included in the sample contain any Reading First schools. This leaves nine states with data that can be used in our analysis: Arizona, Colorado, Delaware, Florida, Georgia, Hawaii, North Carolina, South Carolina, and Texas.[6]
- **Removing from the sample those students who changed schools between kindergarten and third grade.** This step seeks to remove students who were enrolled in Reading First in some years but not others so that we can measure the effect of receiving a "full dose" of Reading First. Unfortunately, our data do not allow us to distinguish these students from other students who moved between schools—that is, from students who moved between two Reading First schools or from students who moved between two non–Reading First schools. Therefore, we exclude all students who changed schools between kindergarten and third grade from the analysis file; 22% of students were excluded on this basis.[7]

Table 1 provides sample sizes for the final data set used to estimate the impacts of Reading First, both overall and for each of the nine included states. Table 2 compares the nine included states with the nation as a whole (all 50 states and DC).

From these data, we construct a longitudinal school-level file for the analysis. This file includes school-level measures of the dependent variable used in the analysis—average Grade 3 test scores on the state reading test—and each of the independent variables used in the analysis (e.g., the percentage of students who are African American—see the full list below). As different states used different test instruments, students' Grade 3 test scores were first standardized by subtracting the state mean and dividing by the standard deviation of all Grade 3 student reading scores in the same state and year. Each of the school-level measures in the longitudinal file is the simple average of the student-level measures for all Grade 3 students in our analysis sample in that school. The analysis file includes one observation for each school and year for which Grade 3 test scores are available. The number of years

TABLE 1
Population Data Used to Estimate Impacts of Reading First

State	Number of districts	Number of Reading First schools	Number of non–Reading First schools	First year of data available	Number of years of data before implementation of Reading First	Number of years of data after implementation of Reading First
Arizona	227	156	452	1999–2000	4	5
Colorado	116	86	308	2002–2003	1	5
Delaware	21	15	66	1998–1999	5	5
Florida	68	589	825	2000–2001	3	5
Georgia	157	128	614	2002–2003	1	4
Hawaii	1	53	95	2001–2002	2	5
North Carolina	131	92	874	2001–2002	3	4
South Carolina	87	51	496	1999–2000	5	4
Texas	1,056	691	2,545	2002–2003	1	5
All nine states	1,864	1,861	6,275			

Note. For some schools, the numbers of years of data before and after Reading First implementation are less than indicated in this table due to missing data.

of data available for each state is indicated in Table 1.

Data to Identify Sites Included in Studies With Nonrandom Site Selection

As noted earlier, nearly all large-scale educational studies based on rigorous impact designs are conducted in nonrandom samples of sites since scientifically valid procedures for identifying intervention impacts require local agency willingness to gather data and enroll participants in particular ways (e.g., experimental and regression discontinuity impact evaluation designs). In education studies, the "sites" are typically school districts or schools. For our analysis, we define sites to be school districts because in the evaluations from which we obtain our data the identity of participating schools is protected by federal statute while that of districts is not. The potential for external validity bias due to nonrepresentative sample inclusion is driven in large part by district willingness to participate in research; even evaluations seeking to sample individual schools must obtain approval from the cognizant school districts before any schools enter the study. Thus, if we find that impacts differ between included and excluded districts as a whole, our findings apply directly to evaluations that include entire districts and have strong salience for evaluations that recruit individual schools *through districts.*

To identify school districts that were included in one or more nonrandom samples of sites, we contacted the project directors of the 23 impact studies that had been initiated by the National Center for Education Evaluation and Regional Assistance (NCEE) in the Institute of Education Sciences (IES) from its founding in 2001 through 2011 (excluding studies that had not yet selected sites). All of these studies were "prospective" studies that required active site-level cooperation to include sites in the study; most of the studies were randomized experiments. For 13 of these studies, we were able to obtain lists of the school districts that participated in the studies.[8] The other 10 studies were unable to provide the list of participating school districts, generally because of agreements with school districts that prohibited a study from identifying the district, or because identifying a district would effectively identify the schools that participated in the study (which would be a violation of the federal law under which the IES was established).

TABLE 2

Comparison of Student and School Characteristics of Nine Included States to the United States (50 States and DC)

State	% Urban (2006–2007)	% FRPL (2006–2007)	% Black (2006–2007)	% Hispanic (2006–2007)	Total expenditures/ pupil (2006–2007)	Total Reading First grant/ student served (2004–2005)
					US$	US$
The United States	30.4	42.4	17.1	20.5	9,683	593
Arizona	53.1	42.8	5.4	41.0	7,338	868
Colorado	33.1	34.2	6.0	27.6	8,286	615
Delaware	5.3	37.1	33.0	9.8	11,760	545
Florida	16.1	45.2	23.9	25.0	8,567	234
Georgia	17.3	50.3	39.2	9.5	9,102	803
Hawaii	0	41.0	2.4	4.5	11,060	265
North Carolina	37.0	43.9	29.2	9.6	7,878	1,016
South Carolina	12.4	51.3	39.8	4.6	8,566	1,081
Texas	42.6	47.6	14.4	46.3	7,850	407

Sources. Local Education Agency Universe Survey of the Common Core of Data for % Urban, % FRPL, % Hispanic (see Hoffman, 2009; U.S. Department of Education, 2008). Common Core of Data National Public Education Financial Survey for Total expenditures/pupil (Zhou, 2009). U.S. Department of Education (2007) for Total Reading First grant/student served.
Note. The United States includes the 50 states and the District of Columbia. % FRPL refers to the percentage of students who are eligible for Free or Reduced Price Lunch. FRPL = Free or Reduced Price Lunch.

Two of these 13 studies were excluded from the analysis presented in this article. One study was excluded because it was conducted entirely in a single state that is not one of the nine states for which we can estimate the impacts of Reading First, given the data available. Another study was excluded because for all of the states included in the study sample, we had only 1 year of pre-Reading First data. This meant that we could not estimate the effects of Reading First with our preferred estimation model, the CITS model described in the next section, because CITS models require multiple years of preintervention data.

The 11 remaining evaluations included in the analyses varied in the interventions on which they focused. In line with the focus of the U.S. Department of Education at the time, almost half of the studies were focused on literacy. But a nontrivial minority was focused on teacher quality. For the full range of topics and studies funded by NCEE, see the NCEE website (ies.ed.gov/ncee/).[9] Note that the included evaluations were not selected because of the topic areas on which they focused; they were selected because they were conducted in purposive samples of school districts.

Table 3 presents the number of school districts in each of the 11 individual evaluations

TABLE 3

Nonrandom Samples for 11 Evaluations Used in the Analysis

Nonrandom sample	Districts included in nonrandom sample: Total number of districts	Districts included in nonrandom sample: Number of districts located in the nine states included in the analysis
1	3	2
2	11	4
3	15	5
4	4	3
5	16	6
6	15	6
7	44	13
8	47	10
9	16	6
10	10	6
11	43	16
Pooled sample[a]	183	60

[a]The number of districts in the pooled sample is less than the sum of the number of districts included in each of the 11 samples because some districts were included in multiple samples—that is, some districts participated in more than one of the 11 evaluations based on nonrandom samples.

TABLE 4
Data Used to Estimate Impacts of Reading First in 11 Nonrandom Samples of Districts

State	Number of included districts	Number of Reading First schools	Number of non–Reading First schools	Number of years of data before implementation of Reading First	Number of years of data after implementation of Reading First
Arizona	7	22	50	4	5
Colorado	3	10	88	1	5
Delaware	0	0	0	NA	NA
Florida	8	346	394	3	5
Georgia	5	33	108	1	4
Hawaii	0	0	0	NA	NA
North Carolina	9	6	163	3	4
South Carolina	5	7	70	5	4
Texas	23	285	529	1	5
All nine states	60	709	1,402		

Note. In some schools, the numbers of years of data before and after Reading First implementation are less than indicated in this table due to missing data. NA = not applicable.

included in the analysis, along with the number of those districts located in the nine states in which we estimate the impacts of Reading First.[10] Table 4 indicates how the latter districts are distributed across the nine states included in the analysis; it also indicates the number of Reading First and non–Reading First schools included in these districts. An in-depth investigation of how the districts included in the 11 evaluations differ from a broader target population of interest is outside the scope of the current article, and is investigated in detail in a companion paper (Stuart, Enbesajjad, Bell, Olsen, & Orr, 2015). Briefly, that work finds that the districts that participate in rigorous evaluations have a number of differences from potential target populations, including greater size, a higher proportion of non-White students, and a higher concentration in urban areas.

For the present analysis, the school districts that participated in these 11 evaluations serve as our proxy for the school districts that might have been included in a nonrandom sample for a hypothetical evaluation of Reading First. One possible concern about this proxy might arise if there were a substantial difference between the target populations for the interventions tested in these 11 evaluations and the target population for the Reading First Program. In that instance, districts that well represent the Reading First population in an external validity sense (the point we are investigating here) might not well represent the populations served by the programs the 11 evaluations *actually examined*—or vice versa. This could lead to misleading findings from our analysis. However, our review of reports from the 11 evaluations suggests that the target populations for the 11 interventions underlying our focal studies consist generally of low-performing students and schools—which, roughly speaking, was the target population for Reading First. Furthermore, as we are defining "sites" at the district level, the key question is whether the target population for these 11 interventions led the evaluations to successfully recruit substantially different districts than would be recruited for a hypothetical evaluation of Reading First. Given that districts with a substantial number of low-performing students and schools are probably considered good candidates for evaluations of most educational interventions, including Reading First, we strongly suspect that a hypothetical evaluation of Reading First would have, to a large extent, recruited the same types of school districts that participated in the 11 evaluations.[11]

Method

In this section, we describe our methodology for measuring Reading First impacts and computing the external validity bias of evaluations based on nonrandom site selection. We first define the population of interest as all third-grade students in Reading First schools in nine states (see the previous section). We then estimate the impact of Reading First on Grade 3 reading achievement *for all Reading First schools in that population* and the impact of Reading First on Grade 3 reading achievement *separately for each of the 11 nonrandomly selected samples from that population.* From there, we construct 11 estimates of external validity bias in evaluations with nonrandom site selection, test their statistical significance and average their magnitudes to produce an estimate of the degree of external validity bias that might have resulted from conducting a single impact evaluation of Reading First in a nonrandom set of school districts.

Analysis Model

To estimate the impacts of Reading First, we use the CITS model shown in Equation 1:

$$\begin{aligned} Y_{st} &= \beta_0 + \beta_1 t + \beta_2 R_s + \beta_3 P_t + \beta_4 tR_s + \beta_5 tP_t \\ &+ \beta_6 R_s P_t + \beta_7 tR_s P_t + \sum_{k=2}^{4} \alpha_k RACE_{stk} \\ &+ \sum_{j=2}^{9} \gamma_j STATE_{sj} + \left(u_s + e_{st}\right), \end{aligned} \tag{1}$$

where

Y_{st} = the average third-grade reading test score of students in school s in year t. As Reading First serves students in kindergarten through third grade, Y measures student performance at the end of the final year of students' participation in Reading First.

t = the number of years since the implementation of Reading First in the state. This variable is set to 0 in the last year before Reading First was implemented and ranges from −5 (6 years before Reading First was implemented) to +5 (5 years after Reading First was implemented).

P_t = 0 for the years before Reading First was implemented $(t \leq 0)$ and 1 for the years after Reading First was implemented $(t > 0)$.

R_s = 1 if school s implemented Reading First and 0 if school s did not implement Reading First.

$RACE_{stk}$ for $k = 2$ to 4 is a vector of time-varying school-level covariates that reflect the racial composition of school s in year t: percent White (k = 1, the omitted category), percent Black (k = 2), percent race other than Black or White (k = 3), and percent with missing data on race (k = 4).

$STATE_{sj}$ for j = 2 to 9 is a vector of time-invariant dummy variables that indicate the state in which school s is located. (The dummy variable for the first state is omitted from the regression model.)

u_s is a school-specific random intercept, with mean 0. This term captures the correlation across observations in different years from the same school.

e_{st} is the random error term in the equation, which is assumed to be independently and identically distributed with mean 0.

Equation 1 is depicted graphically in Figure 1. It assumes a linear time trend in average third-grade test scores during the preimplementation period for both non–Reading First schools (blue lines in the figure) and Reading First schools (green lines) and allows the intercept in time $t = 0$ to differ between the two groups: β_0 for non–Reading First schools versus $\beta_0 + \beta_2$ for Reading First schools. The time slope between $t = -5$ and $t = 0$ also differs between the two groups: β_1 for non–Reading First schools and $\beta_1 + \beta_4$ for Reading First schools. In addition, separately for each group, the model allows the intercept and slope to make a discrete change in the year Reading First was first implemented. For the intercept, the discrete change is β_3 for non–Reading First schools and $\beta_3 + \beta_6$ for Reading First schools; for the slope, the discrete change is β_5 for non–Reading First schools and $\beta_5 + \beta_7$ for Reading First schools. We allow slope and intercept for non–Reading First schools to change at the time Reading First is implemented to control for non–Reading First factors that may affect reading achievement and that might have begun to influence outcomes at the same time Reading First was implemented. We allow the change in the slope and intercept to differ between Reading First and non–Reading First schools to capture the impacts of Reading First.

Equation 1 implies a counterfactual trend for what student outcomes would have been in the Reading First schools in the post-Reading First period in the absence of the intervention. This line is shown with green dashes in Figure 1. It assumes that in the absence of the intervention, Reading First schools would have been the same as non–Reading First schools in two respects. First, the jump in the level of student test scores

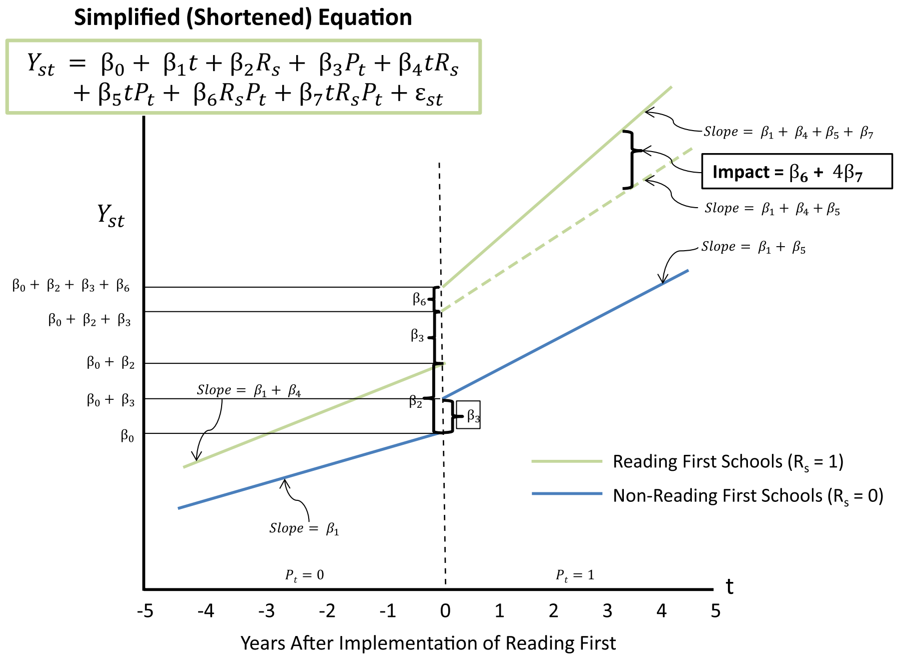
Simplified (Shortened) Equation
$Y_{st} = \beta_0 + \beta_1 t + \beta_2 R_s + \beta_3 P_t + \beta_4 t R_s + \beta_5 t P_t + \beta_6 R_s P_t + \beta_7 t R_s P_t + \varepsilon_{st}$
Y_{st}
$\beta_0 + \beta_2 + \beta_3 + \beta_6$
$\beta_0 + \beta_2 + \beta_3$
$\beta_0 + \beta_2$
$\beta_0 + \beta_3$
β_0
β_6
β_3
β_2
β_3
Slope = $\beta_1 + \beta_4 + \beta_5 + \beta_7$
Impact = $\beta_6 + 4\beta_7$
Slope = $\beta_1 + \beta_4 + \beta_5$
Slope = $\beta_1 + \beta_5$
Slope = $\beta_1 + \beta_4$
Slope = β_1
Reading First Schools (R_s = 1)
Non-Reading First Schools (R_s = 0)
$P_t = 0$
$P_t = 1$
t
-5
-4
-3
-2
-1
0
1
2
3
4
5
Years After Implementation of Reading First

FIGURE 1. *Comparative interrupted time series model of third-grade impact ($\beta_6 + 4\beta_7$).*

when Reading First was first implemented (i.e., in period $t = 1$) would have been the same in both groups in the absence of Reading First (β_3). Second, the shift in the slope of the time trend on student test scores when Reading First was first implemented would have been the same in both groups in the absence of Reading First (β_5).

To understand whether these assumptions are reasonable, we need to consider factors other than the program that could lead to *differential* changes between the two sets of schools in levels or trends in student achievement *starting simultaneously with the implementation of Reading First*. These could include, for example, changes in culture, student birth cohort characteristics, or non–Reading First educational policies or programs that occur simultaneously with the implementation of Reading First. The CITS model produces unbiased impact estimates as long as the combined effect of these other changes is the same for Reading First and non–Reading First schools. Note that state fixed effects are included in the model to make this identifying assumption more plausible. It is more plausible that the effect of factors unrelated to Reading First on the level of or growth in reading achievement would be the same for both Reading First and non–Reading First schools if those schools were operating in the same state and facing the same state educational policies, than if they were operating in different states.

The assumptions described above allow us to identify the impact of Reading First in the model. The impact in year $t > 0$ equals $\beta_6 + t\beta_7$ and is linearly related to the number of years with Reading First in place. We would expect the impacts of Reading First to be associated with the number of years since program implementation (i.e., $\beta_7 > 0$) for two reasons:

1. **Dosage.** As years pass following the initial implementation of Reading First, successive cohorts of third graders are exposed to the intervention for longer periods of time.
2. **Maturity of the program.** In each successive year, later cohorts of third graders would have been exposed, at any given grade level, to a potentially more mature version of the program than earlier cohorts.

There is no way to separately identify the effects of additional dosage from the effects of a more mature program because each successive year brings both a more mature program and a larger dose of the program for third graders in Reading First schools. However, we are not interested in distinguishing between them: We are only interested in estimating the average impact of Reading First for students who received a "full" dose of the program. The CITS model allows us to measure the impact of Reading First *for the first cohort of third graders who participated in Reading First in kindergarten through third grade—the entire grade span served by the program*. These are the students who attended kindergarten in year $t = 1$ and third grade in year $t = 4$. As the impact of the program under the CITS model equals $\beta_6 + t\beta_7$, impact on third-grade reading achievement for these students is $\beta_6 + 4\beta_7$. This parameter captures the average impact of Reading First for the first cohort of students who received a full dose of the program and hence is the parameter of interest in our analysis.

Estimation of External Validity Bias

To estimate the external validity bias from nonrandom site selection, we first estimate the impact of Reading First as described above for the full population of sites in our nine-state population of interest. This involves estimating Equation 1 to measure $\beta_6 + 4\beta_7$ for the full population of Title I schools serving Grade 3 students in the nine states selected for this exercise. This estimate serves as our "population benchmark" for estimating the external validity bias for most of our nonrandom samples. Then, we estimate the impact of Reading First in each of the 11 such samples.

For the 11 nonrandom samples, we make one enhancement to the model: We allow the treatment effect coefficient of interest, $\beta_6 + 4\beta_7$, to vary randomly across sites (school districts) in the population. We include this random component to estimate the standard error of the treatment effect under the null hypothesis that the 11 study samples were selected randomly. If sites were in fact selected randomly, variation in impacts across sites would generate sampling variability in the measure of the population impact measure. We know that none of the 11 studies actually selected a simple random sample of school districts. However, estimating the model under this assumption ensures that we will only reject the null hypothesis if the differences are too large to be attributed to "the luck of the draw" in the site selection process, given the sampling variation in the impacts that would occur if impacts varied across sites and sites were selected randomly.

In calculating comparable impact estimates for the population and each of the nonrandom samples, we face an estimation challenge. The CITS model used to estimate impacts includes state fixed effects to make the identifying assumptions of the model more plausible (see the discussion in the previous section). However, in particular states, for some of the 11 individual evaluation samples, there is no variation in assignment to Reading First: Either all of the schools in the sample were Reading First schools or all of the schools were non–Reading First schools. States with this configuration of data cannot contribute to the identification of the Reading First program's estimated effect. This problem arises only in some of the 11 individual studies: It does *not* arise in the larger nine-state population (as the Title I schools in each of these states include a mix of Reading First and non–Reading First schools).

To address this challenge, we exclude from the analysis states in which there is no variation in assignment to Reading First for a given nonrandom sample.[12] This exclusion is applied not only to the analysis of the nonrandom sample involved but also in *estimating the population benchmark to which the nonrandom sample is compared*. By restricting the analysis to a common set of states in both the population and the sample, we can isolate the differences in impacts that are attributable to nonrandom site selection.

The analysis yields estimates of the impact of Reading First and the standard error of the impact estimate for each of the 11 nonrandom samples. Finally, by subtracting the impact measure for the population ($\hat{\beta}_{6j}^{pop} + 4\hat{\beta}_{7j}^{pop}$) from the impact estimate for each of the 11 samples ($\hat{\beta}_{6j}^{sample} + 4\hat{\beta}_{7j}^{sample}$), we produce 11 estimates of external validity bias, one for each nonrandom sample:

$$\hat{\Delta}_j = \left(\hat{\beta}_{6j}^{sample} + 4\hat{\beta}_{7j}^{sample}\right) - \left(\hat{\beta}_{6j}^{pop} + 4\hat{\beta}_{7j}^{pop}\right) \qquad (2)$$

$$\text{for j} = 1,\ldots,11$$

Each of these estimates has known sampling variability needed for hypothesis testing.

We recognize that the CITS model is not guaranteed to produce unbiased impact estimates for the studied schools, as would random assignment of schools or districts to Reading First. We do not believe that this is a major problem, for several reasons. First, CITS is arguably the strongest nonexperimental model available in this context, as witness its use by several first-rank evaluation teams to evaluate the No Child Left Behind initiative, where random assignment was not possible (Dee & Jacob, 2011; Wong, Cook, & Steiner, 2009). Second, and more fundamentally, for the present analysis we do not need unbiased estimates; we simply need estimates that have comparable bias for the purposive site samples and for the larger population, as our estimate of external validity bias is the difference between these two estimates. As we use the same estimation procedure at both these levels, it seems likely that any bias in the two sets of estimates is, if not identical, at least quite similar.

Hypothesis Test for the Existence of External Validity Bias

One of the main goals of this article is to provide an overall assessment of whether the analysis we have conducted, taken as a whole, provides convincing scientific evidence that an evaluation of Reading First would produce impact estimates with external validity bias if conducted in one or more of the 11 nonrandom samples of sites. To this end, we use a fixed-effects meta-analysis framework to test the following hypotheses:

Null Hypothesis ($\mathbf{H_0}$): The average external validity bias across the 11 samples equals 0.

Hypothesis 1 ($\mathbf{H_1}$): The average external validity bias across the 11 samples does not equal 0.

Rejecting H_0 in favor of H_1 would provide evidence that on average, the 11 nonrandom samples yield external validity bias in estimating the average impact of Reading First. Of course, rejecting H_0 would not indicate that nonrandom site selection *always* yields biased impact estimates.

To test the null hypothesis, we use a fixed-effects meta-analysis framework. Let the external validity bias from sample *j* be given by the following expression: $\Delta_j = B_j^{sample} - B_j^{pop}$, where B_j^{sample} is the true impact in the *j*th nonrandom sample and B_j^{pop} is the true impact in the corresponding population. Formally, we test whether the precision-weighted average of the sample-specific bias estimates equals 0. The expression for the precision-weighted average is shown in Equation 2:

$$M = \frac{\sum_{j=1}^{11} \frac{1}{V_j} \hat{\Delta}_j}{\sum_{j=1}^{11} \frac{1}{V_j}}, \tag{3}$$

where *M* is the precision-weighted average bias estimate, $\hat{\Delta}_j$ is the estimated external validity bias in sample *j* defined above, V_j is the estimated variance of the estimated external validity bias in sample *j*.

Borenstein and Higgins (2013) refer to this analysis as the "fixed-effects (plural) model" and discuss its value in testing whether the mean effect (bias, in our case) equals 0 for the studies (in our case, nonrandom samples) included in an analysis, without trying to generalize beyond them. In this fixed-effects (plural) framework, there is no assumption that the true effect size (bias, in our case) is constant across studies (samples). We use this precision-weighted average rather than a simple average so that more informative studies—those with more precise estimates of the external validity bias—receive greater weight in the statistical test.

This test requires a statistical independence assumption common to most meta-analyses. In particular, the test assumes that the 11 study samples are statistically independent of each other, meaning that a district's probability of inclusion in a given sample does not depend on whether the district was included in any of the other samples. As some districts appear in more than one sample, the independence assumption might appear to be violated. However, overlap in the samples could easily result from two statistically *independent* sampling processes that both favor certain types of sites. For example, if large urban districts are more likely to be included in purposive samples than other types of districts, we might expect to see some of the same large urban districts in more than one study, even though the sampling process for each study was statistically independent of the sampling process for each of the other studies. Therefore, overlap in the samples does not run counter to the assumption of statistical independence among samples on which our test relies.

Findings

This section presents the results of the empirical analysis and hypothesis test described in previous sections. Using the methods described earlier, we calculate the impact of Reading First in the full nine-state population as 0.10 standard deviations on the Reading Test Score scale. This suggests that, in the full nine-state population, Reading First increased reading achievement of third graders who had spent 4 years in the program by one tenth of a standard deviation of the distribution of student test scores for the state reading assessment, relative to how the students would have scored if they had spent the same 4 years in the same school but without Reading First.

Using the same methods, we calculate the impact estimate that would have been produced by an evaluation of Reading First in each of the 11 nonrandom samples of sites. Note that each of these samples is a subset of the schools included in the full nine-state population sample. Figure 2 provides a forest plot of the external validity bias estimates in effect size units (and their 95% confidence intervals) for each of the 11 samples along with the average external validity bias estimate across the samples. For nine of the 11 study-specific samples, the estimated impact is smaller than the population benchmark, as indicated by nine of the 11 squares being to the left of the dashed vertical line representing 0 external validity bias.

To formally test whether on average across the 11 samples the external validity bias is nonzero, we conducted the fixed-effects meta-analysis test described in the previous section. This yielded a precision-weighted average external validity bias estimate of −0.10, with a *p* value of .003. Thus, we can reject the null hypothesis of zero external validity bias on average across the 11 samples at the conventional 5% level (and even at the 1% level). At the same time, only one of the 11 individual sample estimates of external validity bias is significantly different from 0 (as evidenced by the fact that all but one of the 95% confidence intervals in the forest plot overlap 0). Notwithstanding this point, the evidence from the 11 samples combined shows that the average external validity bias from these 11 nonrandom samples when estimating the impact of Reading First is not 0.

Given this evidence of external validity bias, the seriousness of the problem is revealed by the estimated magnitude of bias. A bias of -0.10 is large enough to yield an impact estimate of 0 for an intervention whose true effect we estimate as 0.10 standard deviations. It completely negates evidence of effectiveness for Reading First—a billion dollar program—in the nine states examined. We want to stress, however, that this bias arises in 11 samples that are, by necessity, a convenience "sample of samples." We cannot claim that they, and their degree of bias, are representative of the broader population of nonrandom samples that have been selected in prior studies or could be selected for future studies.

We can also translate the −0.10 estimate of external validity bias in standard deviation units into a more intuitive metric. Bloom, Hill, Black, and Lipsey (2008) analyze data from several standardized tests in reading and mathematics to estimate how much students in U.S. public schools learn each year—that is, how much their standardized test scores increase between annual spring testing periods. Using their estimates, we are able to translate the average absolute bias estimate of 0.10 standard deviations into the number of months that on average is required for students to increase their reading scores by 0.10 standard deviations. Our computations suggest that *this magnitude of bias corresponds to the increase in reading achievement expected for the average third grader over 1.5 months*.[13]

It may also be useful to compare our estimate of external validity bias due to nonrandom site selection with estimates of the *internal* validity bias that might result from two different impact estimation methods that are less rigorous than the CITS design that we used in our analysis, and hence more subject to selection bias: a "naïve" model that does not control for baseline differences in outcomes and a difference-in-differences model that does. We view the naive model as setting an informal upper bound on internal validity bias, as it is arguably one of the weakest impact models available. The difference-in-differences

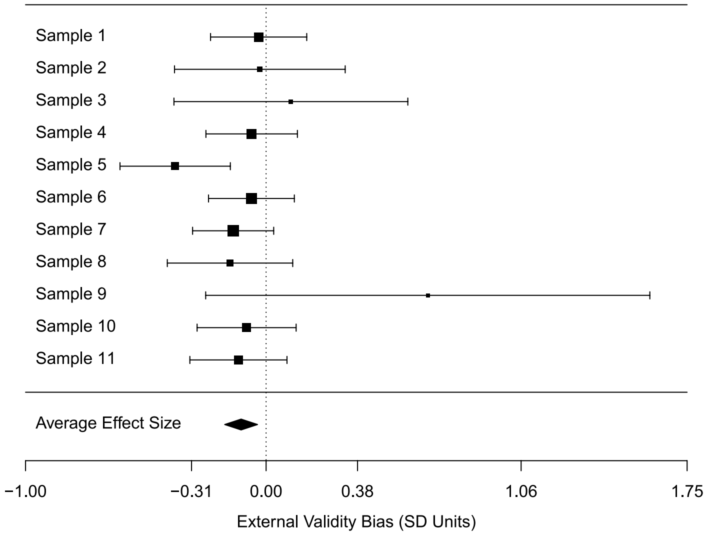

FIGURE 2. *Effect size estimates of external validity bias in 11 nonrandomly selected samples of sites and on average across the samples.*

Note. The graph displays (a) a dashed line to signify zero external validity bias in the population, (b) squares to indicate the estimated external validity bias in each of the 11 nonrandom samples, (c) bars to represent the 95% confidence interval around the estimated external validity bias in each sample, and (d) a diamond to indicate the average estimated external validity bias across the 11 samples (using fixed-effects meta-analysis, as described earlier in the article).

model provides an estimate of internal validity bias using one of the most common nonexperimental methods—one that eliminates the internal validity bias that results from preexisting, time-invariant differences in the *level* of the outcome between the treatment and comparison groups, but does not eliminate the bias from transitory preexisting differences or differences in the outcome *trend* between the two groups. We provide these benchmarks simply to help the reader assess the magnitude of estimated external validity bias; we do not suggest that either should be regarded as an absolute standard for bias. To preview, we find that the external validity bias from nonrandom site selection is half as large as the internal validity bias from using a naïve regression specification that does not control for baseline differences and twice as large as the bias from using a difference-in-differences model.

The naïve regression specification for third grade reading score as the dependent variable includes as explanatory variables an indicator variable for Reading First schools, and indicator variables that control for the racial composition of the student body.[14] It will produce an internally biased estimate of the effect of Reading First if a school's racial composition does not fully explain selection into the Reading First program. Estimating the model in the nine-state population yields an impact estimate of −0.119, with internal validity bias of −0.219. This estimate of internal validity bias is roughly twice as large, in absolute value, as our estimate of the external validity bias from nonrandom site selection.

The difference-in-differences model removes sources of selection bias that do not vary over time, including those that are not captured by the racial composition variables included in the naïve model.[15] This model will produce an internally biased estimate of the effect of Reading First if Reading First and non–Reading First schools were on different outcome "trajectories," as the model attributes differences between the two groups in outcome changes over time to the program. For the nine-state population, this approach yields an impact estimate of 0.054, with estimated internal validity bias of −0.046. Our estimate of the external validity bias from nonrandom site selection is roughly twice this size.

Our estimate of the external validity bias from nonrandom site selection is also larger than the bias tolerance level for education studies established by the U.S. Department of Education's What Works Clearinghouse (WWC). Randomized controlled trials with missing data rates that suggest attrition bias of 0.05 standard deviations or more cannot meet the WWC evidence standards. Our estimate of the external validity bias from conducting an evaluation of Reading First in the 11 selected samples exceeds this threshold by a factor of 2. While thresholds for acceptable bias are inherently arbitrary, the WWC evidence standards suggest that bias of the magnitude found in this article is generally viewed as cause for concern.

Conclusion, Limitations, and Future Research

The analysis in this paper suggests that nonrandom site selection in the context of large multisite impact evaluations of educational interventions can result in important external validity bias. In our analysis sample, the estimated impacts of the Reading First program are less favorable for students in the districts that participated in 11 rigorous multisite impact evaluations of educational interventions than they are for the broader population of districts from which those study samples were drawn.

An important caveat is that our work was not designed to yield findings that are broadly generalizable to other programs and other nonrandom samples. It does not speak to whether purposive sampling conducted as it was in the 11 studies would yield external validity bias for estimating the impacts of programs *other than* Reading First. In addition, there is no single way to select a purposive sample: Some purposive sampling approaches may favor other types of districts than those favored in the 11 studies used here. Our analysis cannot speak to whether other ways of selecting purposive samples would produce external validity bias when estimating the effect of Reading First.

At the same time, our results raise the question of whether the most commonly used approach to obtaining a sample of sites for multisite impact evaluations should be trusted to yield impact estimates that are unbiased for the broader populations for which policy decisions are made.

In addition, the estimated magnitude of external validity bias from nonrandom site selection found, 0.10 standard deviations in absolute value, is about twice as large as the internal validity bias produced by a simple difference-in-differences impact estimator and twice as large as the bias tolerance level for education studies established by the Department of Education's WWC.

The empirical findings of the article should be interpreted in light of the limitations of the analysis. Three important limitations should be borne in mind. First, the types of nonrandom samples included in rigorous educational impact evaluations may differ across educational programs. If so, the samples we used in our analysis, which were selected for evaluations of programs other than Reading First, may at best approximate the type of nonrandom samples that would potentially be selected for an evaluation of Reading First. Therefore, the results from this analysis may simply approximate the external validity bias that would result from a rigorous evaluation of Reading First if the study were conducted in a nonrandom sample of school districts.

In light of this limitation, one way to view these results is that they reflect the external validity bias due to factors that are common across different types of interventions. For example, reluctance to participate in a randomized trial would probably be as applicable to an evaluation of Reading First as to the interventions for which these sites were actually selected. Factors more specific to the interventions tested—such as a district's interest in the intervention or perceived value of the evaluation to the district—are likely to vary among interventions but may well average out across 11 studies. If so, the bias measured in this study is primarily due to factors that are common across these 11 interventions, which is arguably the component of most general interest to the education research community. In this sense, the use of purposive samples from evaluations of interventions or programs other than Reading First is both a limitation and a strength of the current analysis.

Second, in conducting our analysis, we could not estimate the external validity bias from nonrandom site selection in the full population of all states from which the 11 individual study samples were selected. Due to data limitations, we had to constrain our analysis to a subgroup of states—specifically, the nine states for which we could estimate the impact of Reading First on the population of all schools. If we had been able to include all 50 states in the analysis, our measure of the average external validity bias could have been either larger or smaller than the estimate reported here.

Third, we cannot rule out the possibility that our measure of external validity bias is skewed by internal validity bias in the impact estimate for the population, the nonrandom study-specific samples, or both. Given that all of our impact estimates are based on a non-experimental CITS design, they will biased under specialized circumstances discussed above. However, if the degree of internal validity bias is the same for both the population and the nonrandom samples, it nets out in forming our measure of external validity bias. If the internal validity bias in our impact estimates differs between the population and the nonrandom samples of districts used in the analysis, our measure of external validity bias will itself be biased. However, we would expect this bias to be small: We cannot think of any particular reason to expect the assumptions of the CITS model would be violated to a greater or lesser extent in the nonrandom samples selected for other evaluations than in the population.

These limitations suggest some possible fruitful directions for further research on this topic. Just as the evaluation field has benefited from the growing body of evidence on the internal validity bias caused by selection in nonrandomized impact evaluations (for reviews, see Cook, Shadish, & Wong, 2008, and Glazerman, Levy, & Myers, 2003), we believe that the field would benefit from additional empirical evidence on external validity bias. Finding data that allow researchers to produce credible impact estimates for both a broad population and a nonrandomly selected sample of sites may be challenging. However, additional evidence is necessary to assess the likely magnitude of external validity bias that results from conducting a rigorous impact study in a nonrandom sample of sites. If an evidence base continues to suggest that the magnitude of this bias is high, the field should seriously consider ways of addressing this problem (e.g., through alternative site selection approaches, better site recruiting techniques, or statistical methods to reduce the bias at the analysis stage).

Acknowledgment

The authors would like to thank the U.S. Department of Education for allowing us to use data collected for an evaluation of the Reading First Program. We would also like to thank the project directors of 11 evaluations for sharing lists of the school districts that participated in those evaluations. Finally, we would like to thank Edward Bein, David Judkins, Jacob Klerman, and Laura Peck for their feedback on an earlier draft of the paper and Yeqin (Heather) He for the statistical programming required for the analysis.

Declaration of Conflicting Interests

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

Funding

The author(s) received financial support for the research, authorship, and/or publication of this article from the Institute of Education Sciences, U.S. Department of Education, through grant R305D100041 to Abt Associates Inc., and the National Science Foundation, through grant DRL-1335843 to Johns Hopkins University. Dr. Stuart's time was partially supported by National Institute of Mental Health grant K25MH083846.

Notes

1. In our experience, site cooperation is an issue in all prospective studies that require active cooperation from sites for sample identification and data collection, including cooperation with data collection efforts, though it may be a greater issue in social experiments if sites particularly object to random assignment.

2. In related research, we are exploring the extent to which external validity bias in non-representative samples can be reduced with standard statistical methods.

3. This study differs from the more widely known Reading First Impact Study (Gamse, Jacob, Horst, Boulay, & Unlu, 2008). No impact findings have previously been published from the data source used here.

4. To be suitable for this article, an experiment must have been conducted in a population of sites; in addition, selective samples of those sites must have been used to estimate the impacts of some program. The National Job Corps Study is the only experiment of which we are aware that was conducted in the full population of all program sites. However, to the best of our knowledge, no rigorous impact study has been conducted in a naturally occurring, nonrandom sample of Job Corps sites.

5. Reading First funding was restricted to schools that received funding to operate a "schoolwide program" through Title I of the Elementary and Secondary Education Act. For a school to be eligible for this funding, the percentage of students who come from low-income families must be at least 40%.

6. Two of these states, Delaware and Hawaii, contribute no schools or students to any of the 11 Institute of Education Sciences (IES) studies examined. Because those states would be as relevant as any other in, for example, a national study, we include them in the population of interest here when calculating the Reading First impact estimate for the full population. This parallels the complete omission of certain states from many rigorous impact evaluations based on nonrandomly selected sites in education and other fields.

7. We recognize the possibility that the external validity bias from nonrandom site selection could differ between the students we excluded due to data limitations—students who moved between Reading First schools or between non–Reading First schools—and the students included in the analysis because they did not change schools. However, we cannot think of any reason why this would be the case.

8. Without naming individuals, we gratefully acknowledge the information provided by the study directors for these projects. In some cases, they had to contact included school districts to obtain permission to release district names. All of their efforts are a testament to their collegiality and commitment to the importance of the current research and are greatly appreciated by the authors of this paper.

9. A list of the topics and studies under each topic can be found at ies.ed.gov/ncee/projects/evaluation/index.asp.

10. We excluded the sites from one study that were not school districts. While most sites in that study were school districts, some were community-based organizations for which Reading First impacts could not be calculated using our methodology.

11. One might also be concerned about our analysis if the purposive samples were selected at a different period in time than the period over which we could estimate the effects of Reading First. However, our analysis estimates the effects of Reading First 3 years after the implementation of the program—the early to mid-2000s in all nine states—which falls within the window of time during which the purposive samples were selected.

12. For example, one of the 11 studies' nonrandom samples includes a single school district in Arizona that contains only non–Reading First schools. When calculating the population benchmark impact estimate for this sample, we exclude Arizona from the population.

13. Bloom et al. report that the average total increase in student reading scores between spring of Grade 2 and spring of Grade 3 is 0.60 standard deviations (based on the pooled standard deviation of reading scores in Grades 2 and 3). If we assume for simplicity that this entire increase occurred uniformly over the 9 months when school was in session, the average monthly increase in scores for students while in Grade 3 would be roughly 0.600 / 9 = 0.067 standard deviations. Therefore, bias of 0.10 standard deviations equals the expected increase in reading achievement for the average Grade 3 student over 1.5 months of the school year (0.10 / 0.067).

14. The naïve model is a restricted version of the comparative interrupted time series (CITS) model specified in Equation 1, where $\beta_1 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. In estimating this model, we included only the data for $t = 4$ to produce $\hat{\beta}_2$ as an estimate of the effect of Reading First for the 4th year after Reading First implementation, as we did with the CITS model.

15. The difference-in-differences model is a restricted version of the CITS model specified in Equation 1, where $\beta_1 = \beta_4 = \beta_5 = \beta_7 = 0$. In estimating this model, we included all years of preintervention data but only 1 year of post-intervention data—the data for $t = 4$—to produce $\hat{\beta}_6$ as an estimate of the effect of Reading First for the 4th year after Reading First implementation, as we did with the CITS model.

References

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, *1*, 151-178.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289-328.

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*, 134-143.

Braslow, J. T., Duan, N., Starks, S. L., Polo, A., Bromley, E., & Wells, K. B. (2005). Generalizability of studies on mental health treatments and outcomes, 1981-1996. *Psychiatric Services*, *56*, 1261-1268.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724-750.

Dee, T., & Jacob, B. (2011). The impact of no child left behind on student achievement. *Journal of Policy Analysis and Management*, *30*, 418-446.

Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053-1062.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, *22*, 194-227.

Gamse, B. C., Boulay, B., Rulf Fountain, A., Unlu, F., Maree, K., McCall, T., & McCormick, R. (2011). *Reading first implementation study 2008-09 final report*. Washington, DC: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, U.S. Department of Education.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading first impact study final report* (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, *589*, 63-93.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies*, *64*, 605-654.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, *9*, 85-110.

Hoffman, L. (2009). *Numbers and types of public elementary and secondary education agencies from the Common Core of Data: School year 2006-07* (NCES 2009-303). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009303

Humphreys, K., Weingardt, K. R., & Harris, A. H. S. (2007). Influence of subject eligibility criteria on compliance with national institutes of health guidelines for inclusion of women, minorities, and children in treatment research. *Alcoholism: Clinical and Experimental Research*, *31*, 988-995.

Imbens, G. M., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*, 5-86.

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *American Economic Review*, *76*, 604-620.

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

Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. London, England: SAGE.

Rothwell, P. M. (2005). External validity of randomized control trials: "To whom do the results of this trial apply?" *The Lancet*, *365*(9453), 82-93.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1334-1356.

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

Smith, J. C., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators. *Journal of Econometrics*, *125*, 305-353.

Stuart, E. A., Enbesajjad, C., Bell, S. H., Olsen, R. B., & Orr, L. L. (2015). *Characteristics of school districts that participate in rigorous national educational evaluations*. Manuscript submitted for publication.

U.S. Department of Education. (2007). *Reading first state data profiles: 2002-2006*. Office of Elementary and Secondary Education. Retrieved from http://www2.ed.gov/programs/readingfirst/state-data/grantee-profiles.pdf

U.S. Department of Education. (2008). *Public elementary/secondary school universe survey: 2000-01, 2004-05, 2005-06 and 2006-07*. National Center for Education Statistics. Retrieved from https://nces.ed.gov/programs/digest/d08/tables/dt08_042.asp

Wong, M., Cook, T. D., & Steiner, P. M. (2009). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series (Working Paper #WP-09-11). Institute for Policy Research, Northwestern University, Evanston, IL.

Zhou, L. (2009). *Revenues and expenditures for public elementary and secondary education: School year 2006-07 (Fiscal Year 2007)* (NCES 2009-337). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009337

Authors

STEPHEN H. BELL is a senior fellow and principal scientist at Abt Associates, Bethesda, Maryland. His work focuses on designing and conducting experimental evaluations of the effects of government social programs on disadvantaged Americans using multiarm trials and econometric methods.

ROBERT B. OLSEN is a principal scientist at Abt Associates, Bethesda, Maryland. His research is focused on randomized trials in education and approaches to improving the external validity of those evaluations.

LARRY L. ORR teaches program evaluation at Johns Hopkins University and works as an independent consultant on the design and analysis of evaluations of public programs. He has designed and/or directed a number of large-scale experiments and evaluations and has authored a number of scholarly articles and books, including the graduate-level text *Social Experiments: Evaluating Public Programs with Experimental Methods*.

ELIZABETH A. STUART is a professor in the Department of Mental Health, Biostatistics, and Health Policy and Management at Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland. Trained as a statistician, her work straddles public health and education, focusing on issues related to estimating causal effects.

Manuscript received September 16, 2014
First revision received June 28, 2015
Second revision received September 24, 2015
Accepted October 20, 2015