Article

Evaluation Review
2016, Vol. 40(3) 167-198

DOI: 10.1177/0193841X16663414
erx.sagepub.com

The Validity of the Comparative Interrupted Time Series Design for Evaluating the Effect of School-Level Interventions

Robin Jacob[1], Marie-Andree Somers[2], Pei Zhu[2], and Howard Bloom[2]

Abstract

Objective: In this article, we examine whether a well-executed comparative interrupted time series (CITS) design can produce valid inferences about the effectiveness of a school-level intervention. This article also explores the trade-off between bias reduction and precision loss across different methods of selecting comparison groups for the CITS design and assesses whether choosing matched comparison schools based only on preintervention test scores is sufficient to produce internally valid impact estimates. **Research Design:** We conduct a validation study of the CITS design based on the federal Reading First program as implemented in one state using results from a regression discontinuity design as a causal benchmark. **Results:** Our results contribute to the growing base of evidence

[1] Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
[2] MDRC, New York, NY, USA

Corresponding Author:
Robin Jacob, Institute for Social Research, University of Michigan, 426 Thompson Street, Perry Rm 2338, Ann Arbor, MI 48104, USA.
Email: rjacob@umich.edu

regarding the validity of nonexperimental designs. We demonstrate that the CITS design can, in our example, produce internally valid estimates of program impacts when multiple years of preintervention outcome data (test scores in the present case) are available and when a set of reasonable criteria are used to select comparison organizations (schools in the present case).

Keywords
comparative interrupted time series analysis, validation study, quasi-experimental methods, within-study comparison, design replication

In this article, we examine the validity of the comparative interrupted time series (CITS) design in the context of evaluating the effectiveness of a school-level intervention. In a CITS design, program impacts are evaluated by looking at whether the treatment group deviates from its *baseline trend* by a greater amount than the comparison group (see Bloom, 2003; Cook, Shadish, & Wong, 2008, for detailed discussions of this design). The design is more rigorous in theory than many other nonexperimental designs (NXD) because it implicitly controls for differences in both the baseline mean and baseline trends between the treatment and comparison group, while most NXDs are only able to control for differences in the baseline mean. There are a variety of contexts in which the CITS design may be easily applied (Cook, 2012). The design has been widely used in program evaluations (e.g., Bloom & Riccio, 2005; Campbell & Ross, 1968; Mulford, Ledolter, & Fitzgerald, 1992) and has been used in education research to evaluate the impact of No Child Left Behind (Dee & Jacob, 2011; Wong, Cook, & Steiner, 2011). Yet to our knowledge, there has only been one other validation study of the CITS design in education research (St. Clair, Cook, & Hallberg, 2014). That study showed that the CITS design can reduce almost all the initial bias found in an unadjusted quasi-experimental result. Several studies in the field of medicine have also attempted to validate interrupted time series (ITS) or CITS designs (e.g., Fretheim et al., 2013; Schneeweiss, Maclure, Carleton, Glynn, & Avorn, 2004). The findings from these studies suggest that the ITS can credibly replicate the results of a randomized trial, but each study suffers from methodological complications that make its findings inconclusive.

The present study is intended to add to the body of literature about the validity of the CITS design. We conduct a validation study of the CITS

design based on an evaluation of the federal Reading First (RF) program, as it was implemented in one state. The RF program was established under the No Child Left Behind Act of 2001 and is predicated on findings that high-quality reading instruction in the primary grades significantly reduces the number of students who experience difficulties in later years. Nationwide, the program distributed over US$900 million annually to state and local education agencies (LEAs) for use in low-performing schools with well-conceived plans for improving the quality of reading instruction. This federal funding was intended to be used for reading curricula and teacher professional development activities that are consistent with scientifically based reading research (Gamse, Jacob, Horst, Boulay, & Unlu, 2008).

A federal evaluation of the RF program found that it had a positive impact on teachers' instructional practice, including a positive and statistically significant increase in the amount of time spent on reading instruction, but on average, the program appeared to have no statistically significant impact on students' reading comprehension scores. The federal evaluation also found a small, statistically significant impact of RF on first-grade students' decoding skills in the final year of the study (Gamse, Bloom, Kemple, & Jacob, 2008; Gamse, Jacob, et al., 2008).

The state used in this article is unique in that RF funds were allocated statewide based on a numeric rating system for each candidate school. This means that the school-level impact of RF can be estimated using a regression discontinuity (RD) design. RD designs can be used in situations where candidates are selected for treatment (or not) based on whether their "score" on a numeric rating exceeds a designated threshold or cut point. Candidates scoring above (or below) a certain threshold are selected for inclusion in the treatment group, while candidates on the other side of the threshold constitute a comparison group. When the assumptions of a strong RD design are met, one can account for any unobserved differences between the treatment and comparison group by controlling for the value of the rating variable in a regression analysis, and the design produces an unbiased estimate of the effect of the intervention for candidates scoring close to the cutoff. As will be argued later, these conditions are met in the present study. In addition, as will be described later, the particular RD design used for the present analysis is capable of providing impact findings that generalize beyond the cutoff in the data.

We use the estimated impacts from the RD design as a "causal benchmark" (CB) and compare them to the corresponding findings from a CITS design. The CITS can also be used to evaluate the intervention because school-level test scores on state assessments are available for multiple

years, both before and after RF was implemented in the state. Since the state is relatively large, there is also a large pool of elementary schools from which to choose a comparison group.

We also conduct exploratory analyses of several other questions related to the CITS design. First, we examine whether a CITS design can produce valid inferences about the effectiveness of a school-level intervention in situations where it is not feasible to choose comparison schools in the same districts as the treatment schools (which is often recommended in the matching literature (e.g., Cook et al., 2008; Glazerman, Levy, & Meyers, 2003). We also explore the trade-off between bias reduction and precision loss across different methods of selecting comparison groups for the CITS design in our sample. Finally, we examine whether choosing matched comparison schools based only on preintervention test scores is sufficient, in this example, to produce internally valid impact estimates or whether bias can be further reduced by matching on additional school characteristics.

This article proceeds as follows. The second section provides an overview of the CITS design and the third section reviews the relevant literature. The fourth section describes the data set and measures that are used to estimate the impact of RF on test scores and describes our validation methods. The fifth section presents the estimated impacts of RF based on the CITS design and compares these results to our CB estimates. And the sixth section concludes with a discussion of the results.

Overview of the CITS Design

The rigor of any NXD hinges on whether its comparison group provides a valid estimate of the counterfactual outcome for the treatment group. For example, in a simple difference-in-differences (DiD) design, the estimated counterfactual is *the comparison group's change from its baseline mean*. In the DiD design, the underlying identifying assumption is that in the absence of the intervention, the treatment group would have made the same average gains (or losses) as the comparison group.[1] An important (and credible) threat to this assumption is that treatment and comparison schools may have different "maturation" rates. That is, any observed differences could actually be due to a preexisting difference in the growth rates between the two groups rather than the impact of the intervention.

The CITS design addresses these concerns by making use of multiple years of pretest data. In a CITS design, program impacts are evaluated by looking at the extent to which, during the follow-up period, the treatment

group deviated from its baseline trend (baseline mean *and* slope) by an amount that is greater than its comparison group counterpart.

If the comparison group for a study is well chosen, then the CITS design can also account for the effect of policy shocks or other events that co-occur with the intervention being evaluated. These co-occurring events are sometimes referred to as historical bias (Shadish, Cook, & Campbell, 2002). Ideally, the comparison group's deviation from trend would capture the effect of such confounding events. However, if the treatment and comparison groups are not subject to the same co-occurring events, the comparison group's deviation from its trend will not represent the right counterfactual outcome. For example, if the schools in the treatment group (but not the control group) experience massive staff turnover that is unrelated to the intervention, or if the control schools (but not the treatment schools) are subject to a district-initiated school reform initiative, then the estimated impact of the program will be biased. Comparison groups, therefore, should be selected to minimize these potential confounds. If the concern, for example, is with regard to the district level policies implemented during the same period of time as the intervention being evaluated, then comparison schools should, whenever possible, be chosen from within the same district.

Thus, in addition to assessing the validity of the CITS design, this article explores whether—in the context of the RF program—some comparison group selection methods are superior to others with respect to bias reduction. The potential confounds in the evaluation of RF are similar to those that would concern others evaluating educational interventions using a CITS design (e.g., changes in district or state policy that co-occurred with the treatment, changes in the composition of the students or teachers in treatment schools over time, etc.). Therefore, the present findings about CITS designs should yield useful insights for other researchers.

Review of the Literature

The internal (causal) validity of NXDs has been systematically examined in a body of literature known as "validation studies" also called "within-study comparisons" or "design replication" studies (e.g., Cook & Steiner, 2010; Fraker & Maynard, 1987; Heckman, Ichimura, Smith, & Todd, 1998; LaLonde, 1986; Michalopoulos, Bloom, & Hill, 2004; Smith & Todd, 2005). In such studies, researchers attempt to replicate the findings of a randomized experiment by using a comparison group that has been chosen using nonexperimental methods. The estimated bias of the NXD is the

difference between the experimental impact estimate (the best existing information about the "true" impact of the program) and the nonexperimental estimate. A NXD is deemed "successful" at replicating the experimental benchmark if the estimated bias is "sufficiently small."

A variety of criteria can be used to determine what constitutes a sufficiently small bias. For example, St. Clair, Cook, and Hallberg (2014) use a benchmark of 0.20 standard deviations (*SD*s) in their study of Indiana's Diagnostic Assessment Intervention because it is the minimum detectable effect size that the Institute for Education Sciences typically requires in its cluster randomized trails. We use a benchmark of 0.10 *SD*s because effect sizes this small have been shown to be potentially educationally meaningful (Hill, Bloom, Black, & Lipsey, 2008; Schochet, 2008).

The results of previous validation studies of NXDs are mixed—in some cases, NXDs are able to replicate the experimental result (e.g., Hotz, Imbens, & Klerman, 2006), while in other studies, the NXDs produce findings that are substantially biased (e.g., Heckman et al., 1998; Michalopoulos et al., 2004; Smith & Todd, 2005). Two surveys have tried to make sense of these findings by asking not only *whether* NXDs can provide the right answer but also *under what conditions* they are most likely to do so. The first of these two syntheses, by Glazerman, Levy, and Meyers (2003), focuses on validation studies from the job training sector, while the second by Cook, Shadish, and Wong (2008) draws on recent studies from a variety of fields including education.

Both syntheses conclude that some NXDs can replicate experimental results, but that several conditions help to increase the likelihood of this occurring. First, the studies suggest that one's comparison group be chosen from a group of candidates that have been *prescreened* based on having similar motivations and incentives as the treatment group (e.g., individuals who applied for the program of interest but did not receive it). Second, the studies recommend that the comparison group be in close geographical proximity to the treatment group, for example, in the same city or region (*geographically local*). Third, the studies suggest that impact estimates are more likely to be internally valid if *pretest* scores are available for the outcome of interest.[2]

Both reviews also find that, given the design and data for a NXD, the specific statistical estimation methods used to make the treatment and comparison group more equivalent and to control for bias (e.g., regression adjustment, propensity score matching, DiD analysis, etc.) matters little with respect to producing internal validity. However, a more recent study found some situations where simple regression adjustments were not

enough to reproduce the results of an experiment, whereas some other estimation methods were capable of attaining this goal (Fortson, Verbitsky-Savitz, Kopa, & Gleason, 2012).

The primary goal of the present article is thus to determine whether the CITS design can provide internally valid estimates of the impact of a school-level educational intervention. In addition, we conduct exploratory analyses to address the following questions: (1) Are some methods for choosing a comparison group better than others in this study in producing internally valid estimates of the impact of a school-level intervention? and (2) can the precision of impact estimates from the CITS design be improved without compromising internal validity through the choice of a comparison group and/or a matching method (and their resulting sample sizes)?

Method

Data

In this article, we use several data sources to estimate the impact of RF:

State assessment scores. State third-grade reading test scores (the outcome of interest) are available at the school-level from the participating state's department of education website. The third-grade reading assessment used by the state is a nationally norm-referenced test administered each spring. Scores are scaled as normal curve equivalents (NCEs) and are available from Spring 1999 to Spring 2006. The state discontinued use of the assessment in 2007 and replaced it with another test. A different assessment was also used prior to 1999. In this study, therefore, we were able to examine the impact of RF during the first 2 years of implementation, 2005 (Year 1) and 2006 (Year 2), using baseline test scores trends for the years 1999–2004.

Common core of data (CCD) and the U.S. Census data. To describe the samples and identify matched comparison schools, we use information on the characteristics of schools and their school districts. Information on school characteristics (enrollment, student demographic characteristics, location, etc.) was obtained from the CCD at the National Center for Education Statistics for the 1998–1999 to 2005–2006 school years. We also use yearly child poverty rates by school district, for children 5–17 years of age, which were obtained from the U.S. Census Bureau's Small Area Income and Poverty Estimates. Poverty rates are available for 1999–2005. These data are

measured by calendar year not academic year. Calendar year 1999 is used for school year 1998–1999 and so on.

The information we obtained from these two data sources was used to create a panel data set for all elementary schools in the state. This data set includes test scores and demographic information for eight school years (1998–1999 to 2005–2006). The implementation of RF began in 2004–2005, so there are 6 years of preintervention data (1998–1999 to 2003–2004) and 2 years of postintervention data (2004–2005 and 2005–2006). In reporting RF effects on state test scores, we report them as standardized mean-difference effect sizes and compute these effect sizes using a *SD* of 21.06, which is the student-level *SD* that is typically used for NCEs.

We restrict the data set to elementary schools with complete test score data for all 8 years of the study period (6 baseline years and 2 follow-up years). In total, 680 schools meet this requirement and are used in the analysis. Of these schools, 69 received RF funds and have complete test score data and comprise the treatment group for the present analysis.[3]

The Causal Benchmark

In a typical validation study (such as the studies reviewed earlier), the CB for true program impacts is provided by a randomized experiment. In our validation study, however, the CB for the true impact is provided by an RD design. When properly implemented, an RD design can provide estimates of program impacts with as much internal validity as that produced by a randomized experiment, although the causal quantity estimated by an RD study is, in most cases, more narrowly defined than that being estimated by a corresponding randomized experiment. In addition, for a given sample size, the precision of an RD study can be much less than that of a corresponding randomized experiment. Within-study validation comparisons of RD designs have consistently shown that they can have strong internal validity (see, e.g., Aiken, West, Schwalm, Caroll, & Hsuing, 1998; Gleason, Resch, & Berk, 2012; Shadish, Rodolfo, Vivian, Steiner, & Cook, 2011). However, the internal validity of the RD design is not guaranteed because the design must satisfy several important conditions for its impact estimates to be internally valid.

As described in detail in Somers, Zhu, Jacob, and Bloom (2013), all of the conditions for a strong RD hold for the present example. Nothing other than treatment status is discontinuous at the cut point value of the RD rating, and the rating variable and cutoff point were determined independently of each other. This latter point is true because the rating variable for each

school was determined prior to (and thus independent of) the allocation of funding and the cut point for determining the allocation of federal funding was determined exogenously. It was based solely on the amount of available funding and the amount of funding requested for each school. After ratings were assigned to schools that applied for RF, these schools were ranked from highest to lowest rating. The amount of funding requested by each school in their application (which was based on the size of the school) was also recorded. Funding was then awarded to the highest rated schools in rank order until the available pool of funds was exhausted.

Finally, the functional form representing the relationship between the rating variable and the outcome is continuous throughout the analysis interval absent the treatment, and numerous specification tests indicate that this functional form was specified correctly.

The causal quantity estimated by the RD and the CITS designs. In most instances, the RD produces estimates of program impact that apply to candidates with ratings that are very near the cutoff and thus cannot be readily generalized to the entire sample of candidates. This is because the mean counterfactual outcome for the relevant portion of the treatment group in an RD (those candidates that just exceed the cutoff) is represented by the predicted outcomes of the comparison group who just miss the cutoff point. Hence, RD impact estimates typically represent the effect of the program for participants near *the cutoff only*. In contrast, the CITS design provides an estimate of the average impact *for all RF schools* (the average treatment effect).

Therefore, in order to use the RD as a benchmark, the RD estimates must be generalizable to all RF schools in the CITS sample. Evidence suggests that this is the case for the present analysis. As shown in Figure 1, in the case of RF in the state we are studying, there is almost no relationship between the ratings and the outcome measure during the baseline period (i.e., the baseline trend is virtually horizontal).[4] Since there appears to be almost no relationship between the ratings and test scores (school outcomes), it seems unlikely that there would be a relationship between ratings and the magnitude of school impact. As described in more detail in Somers et al. (2013), a number of specification tests support this conclusion. For example, statistical tests show that the ratings–test score slope is the same on either side of the cutoff. If RF had a larger than average impact on RF schools further from the cutoff, then an increase in these schools' test scores would make the slope for RF schools different (steeper) than the slope for non-RF schools. Similarly, the estimated impact for *all RF schools* does not differ

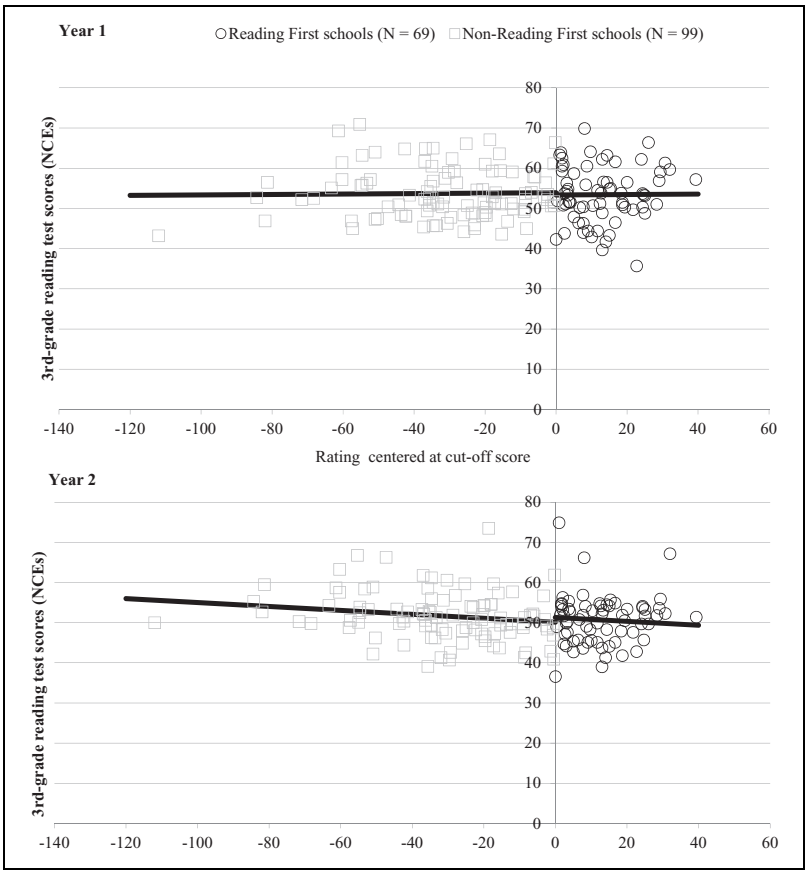

Figure 1. Relationship between reading scores and ratings.

from the impact for schools around the cutoff. This suggests that, in this case, the estimated impacts from the RD design represent the average treatment effect of RF for all of the RF schools in our study sample, which is the same causal quantity that will be estimated from the CITS designs.

Estimating the CB. We begin by estimating the impact of RF in follow-up Years 1 and 2 using a standard RD design. Specifically, we estimated the following model:

$$Y_j = \pi_0 + \psi_0 \mathrm{TREAT}_j + \rho_0 \mathrm{RATINGC}_j + \varepsilon_j, \quad (1)$$

where RATINGC_j is a continuous variable for the rating assigned to each schools' application centered at the cutoff (= 0), and all other variables are as defined above. Based on this analysis, the estimated impact of RF on reading scores is −0.026 in Year 1 (p value = .725) and 0.057 in Year 2 (p value = .434). Note that the lack of program impacts does not invalidate or weaken the validation study. What matters is the difference between the benchmark estimate and the CITS estimate of the same causal quantity.

Unfortunately, the CB estimates that are obtained from the standard RD analysis are much less precise than the estimates that can be obtained from the CITS making it difficult to assess whether the two estimates differ. However, in cases where there is no relationship between the rating and the outcome, as is the case in the present example, it can be argued that it is not strictly necessary to control for the rating variable in the RD analysis, in which case the RD model reduces to the model for a randomized experiment. Given this, we *also* analyze the data *as if it were* an randomized control trial (RCT), dropping the rating variable from our impact estimation model, in order to maximize the statistical power for our comparison with the CITS design.[5] The following model was therefore used to estimate these benchmark impacts:

$$Y_j = \alpha_0 + \beta_0 \text{TREAT}_j + X_j + \varepsilon_j, \quad (2)$$

where j denotes schools, TREAT is a dichotomous indicator for whether school, j, is a treatment school (1 = *treatment* and 0 = *control*), and X includes a vector of school-level prior year test scores which are included to increase precision.

Based on this analysis, we find that the estimated impact on reading scores in the first year of RF is −0.015 *SD*s (p value = .722). The estimated impact on reading scores in the second year of RF is −0.026 (p value = .529). These two estimates are also not statistically different from the estimates obtained based on Equation 1 (Year 1: p value = .943 and Year 2: p value = .149).

CITS Design

The CITS analysis estimates the average trend in third-grade test scores for each of the two groups of schools (treatment and control) during the baseline period and then estimates the amount by which these schools' test scores deviate from their baseline trend during each of the 2 years in the follow-up period. Average deviations from trend are obtained for both RF schools and comparison schools, and the impact of the intervention is

estimated as the *difference* between these two deviations from trend. If the program is effective, then this difference will be positive. Specifically, we estimate the following two-level model from data for all program and comparison schools during 6 baseline years and the 2 follow-up years for our analysis sample:

Level 1 (school years within schools):

$$Y_{jt} = \alpha_{0j} + \beta_{0j}\text{TREAT}_j + \phi_{0j}\text{RELYEAR}_t + \lambda_{0j}\text{RELYEAR}_t \times \text{TREAT}_j + \alpha_1\text{YR1}_t + \beta_1\text{TREAT}_j \times \text{YR1}_t + \alpha_2\text{YR2}_t + \beta_2\text{TREAT}_j \times \text{YR2}_t + \varepsilon_{jt}.$$

Level 2 (schools):

$$\alpha_{0j} = \alpha_0 + u_j,$$

$$\phi_{0j} = \beta_0 + \tau_j,$$

where j denotes schools and time t spans all 6 baseline years (1999–2004) and 2 follow-up years (2005 and 2006). The parameters in the model are defined as follows: TREAT is defined as before, YR1 is a dichotomous variable that indicates whether or not the observation represents the first intervention year (= 1 if 2005 and 0 otherwise), YR2 is a dichotomous indicator which indicates whether or not the observation represents the second intervention year (= 1 if 2006 and 0 otherwise), RELYEAR is a discrete variable for school year centered at the last baseline year (= 0 in 2004), ε_{jt} is the random variation in test scores across time within schools (within-school variation), u_j is the between-school random variation in the intercepts of the baseline trends (centered at the last baseline year), and τ_j is the between-school random variation in the slopes of the baseline trends.

In this model, β_1 represents the mean estimated impact in follow-up Year 1—the deviation from trend for treatment schools minus the deviation from trend for comparison schools for that year. The estimated impact in follow-up Year 2 is β_2.

Selection of Comparison Schools

As noted above, the choice of comparison group can help ensure the validity of the design. The choice of comparison group can also have an impact on the precision of the estimates. Many different strategies and methods for choosing comparison schools exist, and some may provide a less-biased representation of the mean counterfactual than others. At the same time,

some selection strategies yield larger comparison groups than others and therefore may produce more precise impact estimates (greater precision). Therefore, we use several methods for selecting comparison groups with the goal of comparing both their bias reduction and precision gain.

We begin by using all 611 non-RF schools in the state as a comparison group. While this provides a large sample size with which to work, as already noted, some existing literature (Cook et al., 2008; Glazerman et al., 2003) indicates that comparison groups are more likely to reflect the right counterfactual when they are somehow "prescreened" for program participation, geography proximity, or demographic similarity. This is because, for example, there may have been specific state or district policies, demographic shifts, or contextual factors that co-occurred with the treatment and affected the test scores of students in schools that were granted RF funds. A convincing comparison group might therefore meet the geographical or needs-based conditions for participating in the intervention or might have taken the further step of applying for consideration, thus increasing the likelihood that they were exposed to the same external factors. Narrowing the comparison pool based on ability, motivation, geography, or some other known selection criterion is a way of increasing the likelihood that the comparison group outcome will represent a valid counterfactual.

Therefore, in addition to using all non-RF schools in the state as a comparison group, we consider a comparison group consisting only of schools that are located in districts eligible for RF funds. To be eligible, a LEA had to have at least one school with more than 50% of students reading below proficiency in fourth grade and had to be within one of the three prespecified categories for school improvement.[6] A total of 419 schools spread across 79 eligible districts were included in this comparison group. Note that we did not further restrict this comparison group to schools in the *same set of districts* as the RF schools, which affords us an opportunity to examine whether in this example, using geographically local comparisons is a necessary condition for the internal validity of the CITS designs.

As a third alternative comparison group, we use the 99 schools that actually *applied* for but did not receive RF funds. By definition, these schools met all RF eligibility criteria, but in addition, they had the motivation and resources to apply for RF, which may make them more similar to the schools who received RF funds. These nonwinning applicants also constitute the comparison group used to estimate the impact of RF in the RD analysis.

Finally, we use statistical matching methods to identify schools among the 419 schools in districts that were eligible to apply for RF funds—that are similar to the RF schools based on baseline test performance and other school-level eligibility criteria (such as the percentage of low-income children and Title I status).[7] Selecting schools which look most similar to the RF schools (in terms of demographics and prior test scores) may provide the best estimate of the counterfactual outcome for RF schools. If so, then matching techniques may be the best option for creating a credible comparison group.

The primary characteristic used for matching RF schools to comparison schools are the mean third-grade reading scores for schools during the preintervention period. Prior test scores are strong predictors of test score outcomes in the follow-up period, so matching on school-level pretest scores (and where possible baseline trends) is likely to increase the probability that comparison school outcomes will represent valid counterfactual outcomes for treatment schools. However, we also examine whether there is any benefit to matching on additional school characteristics. Specifically, we tried matching on test scores *plus* the following 12 school characteristics: the location of the school (rural or urban), total school enrollment, third-grade enrollment, the percentage of students who receive free or reduced price lunch, the racial–ethnic composition of the school (percentage of students who are White, Black, Hispanic, Asian, or Other), the percentage of third-grade students who are girls, the pupil–teacher ratio, and child poverty rates for the district. These characteristics were chosen because they have been used in the past to predict test scores. Matching on these characteristics may improve the comparability of the treatment and comparison schools, and further reduce bias, because schools' eligibility for RF funds was partly based on the characteristics like their percentage of low-income students. On the other hand, in some applications, it may not always be possible to match well on both test scores and demographic characteristics (if, e.g., the matching pool is limited), so that matching on demographic characteristics could result in a poorer match on test scores.

There are numerous approaches to statistical matching. In previous work (Somers, Zhu, Jacob, & Bloom, 2013), we compared three different matching methods (nearest neighbor, nearest neighbor without replacement, and the radius method) and demonstrated that the best matching method in this sample and context (i.e., the one that introduced the least bias and had the greatest precision) was the radius method. For this reason, we use the radius method here.

For the radius method, each treatment school is matched to all "suitable" comparison schools defined as all schools within a given statistical distance (or radius) of the treatment school as measured by a propensity score. The propensity score is based on the matching characteristics. It is the estimated probability of being a treatment school, given the schools' characteristics. Because several characteristics and multiple years of data are used for matching, we needed to collapse these variables into an overall index of "similarity" to make the process of matching more tractable. Matching is conducted with replacement (a comparison school can be matched to more than one treatment school).[8] The advantage of the radius method, which allows more than one comparison school per treatment school, is that the size of the comparison group is larger than for one-to-one matching, thus impact estimates may be more precise. However, if the radius is too wide, then greater precision will come at the cost of less suitable comparison schools, which could introduce bias into the impact estimate. See Somers et al. (2013) for a detailed description of how the optimal radius was chosen.[9]

Thus, we created sets of comparison schools in two different ways. First, we use a set of prescreened comparison schools that were not matched but resemble the RF schools with respect to either geography (all non-RF schools in the state), eligibility (all non-RF schools in eligible districts), or motivation (schools that applied for RF funds but did not receive them). Second, we use "matched" comparison schools that were created by matching on a propensity score calculated from 6 years of baseline data; one set matches on test scores alone and the other also matches on demographic characteristics.

Assessing Bias

The key question in this article is whether the CITS design can produce internally valid estimates of education program impacts. To answer this question, we estimate the *bias* for each CITS estimate defined as the difference between the CITS impact estimate and the CB:

$$\widehat{\text{BIAS}}_{\text{CITS}} = \widehat{I_{\text{CITS}}} - \widehat{I_{\text{CB}}},$$

where $\widehat{I_{\text{CB}}}$ is the CB estimated from the RD design and $\widehat{I_{\text{CITS}}}$ is the estimated impact from the CITS design.

The bias is assessed based on these two impact estimates each of which contains estimation error. Therefore, what we observe is the *estimated* bias, which is also estimated with error. This error must be taken into account

when interpreting the magnitude of the estimated bias and, in particular, whether the confidence interval around the estimated bias includes 0. If it does, then there is no evidence that the CITS impact estimates are biased.

To conduct hypothesis testing on the estimated bias, we need to determine its standard error. Yet estimating a correct standard error is tricky because the impact estimates being compared ($\hat{I_{CB}}$ and $\hat{I_{CITS}}$) are not independent of each other. This is because the treatment group is the same across impact estimates, and there is also overlap in the comparison groups used for each impact estimate. For example, some of the non-RF schools used in the CB analysis are also comparison schools in the CITS analyses. In order to make correct inferences about the size of the bias, the standard error of the estimated bias must account for this dependence. If we were to incorrectly assume that the impact estimates are independent, then the standard error of the estimated bias would be too large, and we could mistakenly conclude that the estimated bias is not statistically significant when in fact it is.

We therefore use nonparametric bootstrapping to estimate standard errors for the estimated bias. Bootstrapped standard errors account for the dependence between impact estimates and can be used to test whether the estimated bias for a given CITS impact estimate is statistically different from 0 (Efron & Tibshirani, 1993).[10] Importantly, bootstrapping also accounts for uncertainty in the propensity score matching process. In addition, bootstrapping is used to test whether bias estimates *differ* across different comparison group selection methods.[11]

Finally, we also compare the *standard error* of impact estimates from the CITS design as a means of gauging their relative precision. Precision is especially relevant for the choice of the comparison group selection method. As noted earlier, some matching methods produce comparison groups that are larger than those produced by other matching methods or unmatched comparison groups and thus may produce impact estimates that are more precise. Assuming that two methods have similar bias, then the method whose estimates are more precise is preferred because it increases the likelihood of detecting policy-relevant impacts.

Previous validation studies have opted for criteria other than bias and precision to compare impact estimates across designs. There are a number of reasons why we do not use these other criteria in our analysis. The first criterion is the statistical significance of impact estimates—that is, whether inferences about program effectiveness (based on p values) are the same across study designs (see, e.g., Cook et al., 2008). In our study, we do not use this criterion for two reasons. First, in a validation study, the key

question is not whether the levels of statistical significance differ but whether the magnitudes of the impacts differ. Two estimates could both yield statistically significant impacts, but the magnitudes might differ substantially. These would lead to very different interpretations regarding the effectiveness of the program. Accordingly, the relevant hypothesis test in a validation study is whether differences between impact estimates are statistically significant not whether impact estimates themselves are statistically significant. Second, the impact estimates in our analysis (and in many analyses) differ in terms of their precision due to differences in the study design and the size of the comparison group. When precision differs across two estimates, these estimates may exhibit different patterns of statistical significance even when both estimate the same impact. In other words, bias and precision are confounded. We consider bias and precision separately, since bias is the most important consideration in a validation study.

Previous studies have also used the mean squared error (MSE) as a criterion for comparing alternative ways of estimating impacts (e.g., Orr, Bell, & Kornfeld, 2004). We do not use the MSE as a criterion for comparing impact estimates in this article because it also, by definition, combines the bias and precision of an estimated impact into one measure, which makes it difficult to interpret and compare the MSE of different impact estimates. We argue that for the present analysis, it is more useful to consider bias and precision separately as outlined in our approach.

Results

We begin by exploring the characteristics of the five comparison groups relative to the characteristics of the RF schools (the treatment group). Table 1 presents the characteristics of the comparison groups used in the analysis. In this table, statistical tests of the difference between RF schools and other groups are not shown for two reasons. First, the precision of the estimated difference varies across comparison groups—for a difference of given magnitude, comparison groups with more schools are more likely to be deemed statistically different from RF schools. Second, our goal is to assess the relative similarity of groups, so the statistical significance of differences is less relevant than the *size* of the observed differences and ultimately the size of the estimated bias. As a rule of thumb in propensity score matching, it has been suggested that treatment and comparison groups should differ by not more than 0.25 *SD* on key characteristics (Ho, Imai, King, & Stuart, 2007), so values greater than this threshold are flagged in the table ("†").[12] The table also shows (in parentheses) the standard error of the differences.

Table 1. Characteristics of Reading First Schools and Comparison Groups (for Impacts on Reading Scores).

		Comparison Groups				
School Characteristic	RF Schools	State	Eligible	Applicants	Radius	Radius With Demographics
Baseline reading test scores						
Predicted score in last baseline year	52.75	57.69 (0.25)*	56.51 (0.3)*	53.07 (0.54)	55.20 (0.3)*	55.13 (0.34)
Slope of baseline trend (6 years)	1.24	1.26 (0.05)	1.28 (0.06)	1.23 (0.12)	1.20 (0.06)	1.22 (0.07)
Demographic characteristics (last baseline year)						
Percentage of schools that are urban	37.68	34.26 (6.15)	35.80 (6.3)	22.22 (7.22)*	41.02 (7.05)	30.07 (8.46)
Enrollment	382.61	409.56 (17.31)	400.13 (17.87)	362.55 (22.77)	376.88 (20.32)	371.14 (29.15)
Free/reduced-price lunch (%)	65.64	53.96 (2.36)*	57.97 (2.41)*	70.73 (2.67)	65.18 (2.76)	67.28 (3.79)
Racial/ethnic composition						
White (%)	81.35	88.31 (2.59)*	85.73 (2.66)	88.36 (3.11)*	83.31 (3.06)	83.78 (4.12)
Hispanic (%)	2.50	1.60 (0.51)	1.68 (0.53)	1.35 (0.56)*	1.94 (0.59)	1.57 (0.55)
Black (%)	15.17	9.16 (2.2)*	11.54 (2.26)	9.70 (2.67)*	13.83 (2.59)	13.91 (3.87)
Other (%)	2.50	1.60 (0.51)	1.68 (0.53)	1.35 (0.56)*	1.94 (0.59)	1.57 (0.55)
Number of third grade students	59.97	62.89 (3.7)	60.59 (3.77)	52.04 (4.59)*	56.29 (4.27)	56.46 (5.82)
Third graders who are female (%)	47.91	47.48 (0.58)	47.56 (0.6)	46.69 (0.79)*	47.60 (0.78)	48.49 (1.08)
Children in poverty in district (%)	22.00	20.66 (0.89)	22.45 (0.9)	25.75 (1.12)*	23.18 (0.17)	22.69 (0.1)
Pupil–teacher ratio	14.47	15.57 (0.32)*	15.40 (0.33)*	14.32 (0.37)	15.17 (0.35)*	14.62 (0.37)
Number of schools	69	611	419	99	369	324

Note. Values shown in parentheses are the standard error of the difference between Reading First (RF) and comparison schools. Differences greater than 0.25 standard deviation in effect size are indicated with an "*."

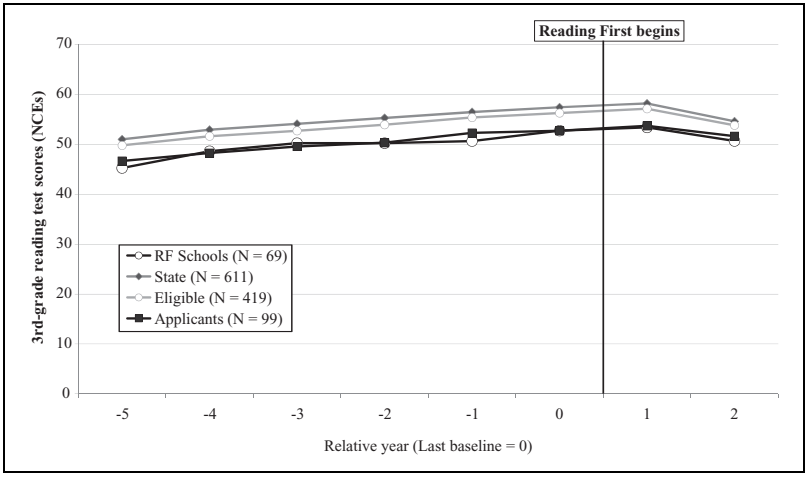

Figure 2. Reading test score trends for Reading First schools and prescreened comparison groups.

The results in Table 1 show that the RF schools are much lower performing compared to all other schools in the state, which is to be expected given the eligibility requirements. RF schools are also lower performing than the other schools in districts that met the eligibility criteria, which indicates that schools that were motivated to apply for RF funds had lower-than-average scores among those schools in eligible districts. In terms of their reading achievement, RF schools are most similar to the other schools that applied for RF funds. On the other hand, RF schools and the other RF applicants are somewhat dissimilar with respect to demographic characteristics; effect size differences in racial–ethnic composition, enrollment, and poverty are all larger than 0.25.

Table 1 also shows that, in terms of baseline test scores, the matched comparison groups are less similar to RF schools than the ''applicant'' group but are more similar to RF schools than the ''eligible'' group. These similarities and differences can also be seen in Figure 2. In addition, both radius matching methods produce few large differences in terms of observable demographic characteristics, as opposed to the applicant pool, which produced many differences that were greater than 0.25 *SD*s. Interestingly, matching on demographics—in addition to pretest scores—does not appreciably improve the comparison group's similarity to RF schools with respect to either demographics or pretest scores. In this case, matching only on pretest scores is sufficient for achieving comparability with respect to

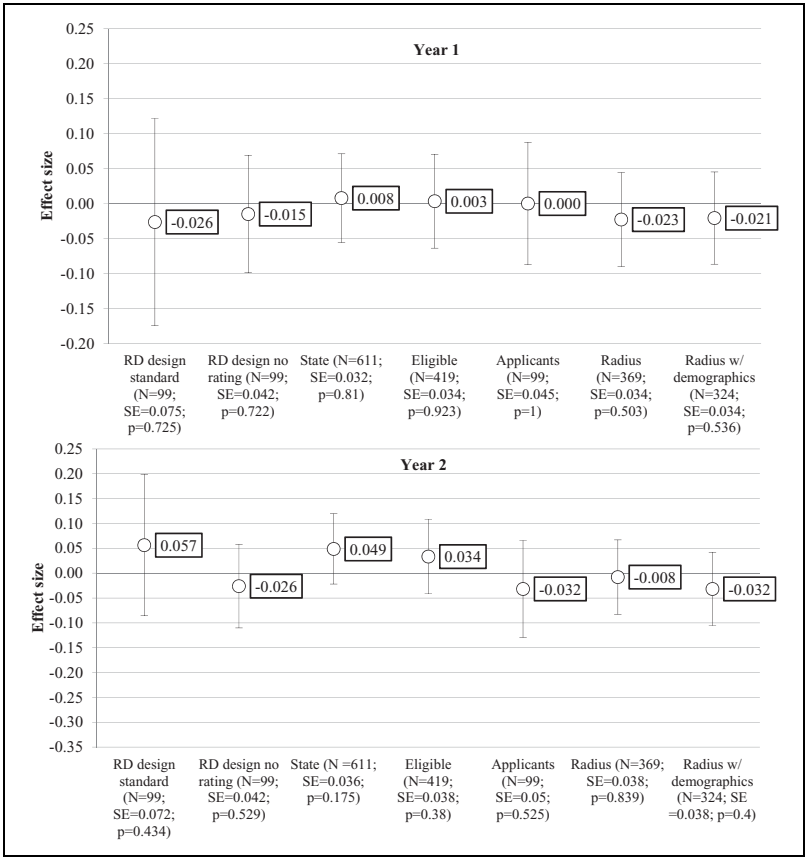

Figure 3. Estimated impact on reading scores by comparison group, the comparative interrupted time series design. *N* = number of comparison schools, *SE* = standard error, and *p* = *p* value.

test scores *and* demographic characteristics, even though the latter are not included in the matching process. This suggests that in the present data set, there is a strong relationship between baseline test scores and student demographics. This same relationship might not exist in other data sets.

Estimated Bias of the CITS Design

Next we explore the estimated bias of the CITS design. Figure 3 presents impact estimates and 95% confidence intervals for the CITS design in the

first and second year of the intervention for each of the five comparison groups. This figure also includes the CB impact estimate and confidence interval as a reference point.[13] In general, we see that all impact estimates (including the CB) hover around 0 and that there is no discernible pattern of bias. The estimated impacts are, for the most part, well below 0.10 *SD*s and thus are not likely to be educational meaningful. There is also substantial overlap in the confidence intervals for the CB estimate and the intervals for other estimates, which suggests that the CITS estimates are not statistically different from the CB. As noted earlier, however, the impact estimates are correlated and so strictly speaking, the confidence intervals cannot be directly compared.

Table 2 presents formal tests of whether estimated bias for each CITS estimate is statistically significant. The estimated bias is defined as the difference between the CITS estimate and the casual benchmark estimate, which here are scaled as effect sizes. Bias estimates are small in magnitude, ranging from −0.01 to 0.075 (all below the specified criteria of 0.10). Based on bootstrapped standard errors—which account for the correlation among impact estimates—none of these bias estimates are statistically different from 0 at the 5% level for either intervention year (Year 1 or Year 2). Among all the bias estimates, the largest is 0.075 for the statewide comparison group in Year 2. This bias estimate also comes close to statistical significance ($p = .056$). The other impact estimates are all quite small and do not approach statistical significance, suggesting that they are internally valid (p values ranged from .129 to .969).[14]

Differences in Bias and Precision Across Comparison Groups

Next, we can compare the size of the estimated bias and the precision of impact estimates across comparison groups. Bias estimates for each group are presented in Table 2, while the standard error (which was not based on bootstrapping) of each impact estimate is shown in Figure 3. Statistical tests for the difference in bias estimates across groups and designs (based on bootstrapping) can be found in Table 3. There are some statistically significant differences in the estimated bias between the state sample and both the eligible sample and the applicant sample in Year 2 (which range from 0.015 to 0.08), again calling into question the internal validity of the result that uses the state sample.

The findings indicate that there is no evidence of bias for two of the three prescreened (unmatched) groups nor for the two matched groups. Although all produce internally valid estimates, as shown in Table 1, impact estimates

Table 2. Estimated Bias (in Effect Size) for Impact on Reading Scores by Comparison Group.

Comparison Group	Estimated Bias	Bootstrap Standard Error	Bootstrap *p* Value	Bootstrap 95% CI [Lower, Upper]
Year 1				
State	.023	.033	.460	[−.038, .088]
Eligible	.018	.032	.540	[−.042, .084]
Applicants	.015	.029	.580	[−.041, .072]
Radius	−.008	.031	.675	[−.076, .048]
Radius with demographics	−.006	.089	.969	[−.171, .179]
Year 2				
State	.075	.039	.056	[−.002, .150]
Eligible	.060	.039	.129	[−.020, .137]
Applicants	−.006	.042	.909	[−.092, .079]
Radius	.018	.035	.841	[−.061, .078]
Radius with demographics	−.006	.081	.948	[−.143, .164]

Note. The estimated bias is equal to the estimated impact based on the relevant comparison group minus the estimated impact from the causal benchmark (CB), assuming a functional RCT, using the actual data. The standard error (*SE*), *p* value, and confidence intervals (CI) for the bias are obtained using bias estimates from bootstrapped samples (1,000 iterations). The *SE* is the standard deviation of bias estimates across iterations. The *p* value is obtained by assuming that the distribution for bias is normally distributed. The CIs are the 2.5th and 97.5th percentiles of the bias estimates across iterations. All bias estimates, *SE*, and CIs are shown in effect size based on a standard deviation of 21.06, which is the student-level standard deviation for scores in normal curve equivalents. The bootstrap *SE* for the radius with demographics is substantially larger than for the other groups. Mathematically, the *SE* of the bias is a function of (1) the *SE* of the CB, (2) the *SE* of the comparative interrupted time series impact, and (3) the correlation between the two. For most of the impacts, the *SE* of the impact estimates that obtained from bootstrapping is very similar to the parametric *SE* of the impact estimate shown in Figure 3, but in the case of radius with demographics, the bootstrap *SE* is much larger (.086 vs. .034). We believe this is because the bootstrap *SE* also builds in uncertainty from the matching process, and there is more variability in the comparison group drawn when we also matching on demographics.

from the matched groups are relatively more precise. In Year 1, for example, the standard error for the radius matching method is comparable to the size of the standard error for the CITS impact estimate based on the eligible schools, even though the latter group is somewhat larger. Similarly, in Year 1, the standard error for the CITS impact estimate based on the radius method is about 75% of the size of the standard error for the impact based

Table 3. Differences in Comparative Interrupted Time Series Bias Estimates for Reading in Years 1 and 2.

Study Design—Comparison Set	(1)	(2)	(3)	(4)	(5)
(1) Year 1—State		.004	.008	.031	.029
		(.497)	(.796)	(.200)	(.824)
(2) Year 1—Eligible			.003	.026	.024
			(.924)	(.259)	(.866)
(3) Year 1—Applicants				.023	.021
				(.445)	(.896)
(4) Year 1—Radius					−.002
					(.841)
(5) Year 1—Radius with demographics					
(1) Year 2—State		.015*	.081*	.057	.080
		(.032)	(.004)	(.097)	(.319)
(2) Year 2—Eligible			.065*	.041	.065
			(.018)	(.214)	(.429)
(3) Year 2—Applicants				−.024	.000
				(.770)	(.983)
(4) Year 2—Radius					.024
					(.880)
(5) Year 2—Radius with demographics					

Note. The value in the first row of each cell is the estimated difference (in effect size) between impact estimates based on the actual data. That difference is equal to the estimated impact based on the comparison group in row X minus the estimated impact based on the group in column Y. Effect sizes are calculated using a standard deviation of 21.06, which is the student-level standard deviation for scores in normal curve equivalents. The value in the second row of each cell is the *p* value for the difference between impact estimates based on bootstrapped samples (1,000 iterations). Gray shading indicates that the bias is by default "0" for these cells.

on applicants. Finally, the standard error for CITS impact estimate based on the radius method is only slightly larger than the impact estimates derived using the full state sample. Thus, in our sample, the matching confers precision that is equal to or greater than that for most of the unmatched comparison groups while still providing impact estimates that are internally valid.

At the same time, we conclude that there was no benefit to matching on demographic characteristics in addition to test scores in this example. Estimates of bias for these two approaches (matching with test scores only or

matching on demographic characteristics plus test scores) are not statistically different from each other, and their standard errors are also similar (ranging from 0.02 to 0.03). Furthermore, in the present analysis, adding demographics to baseline test trends when matching produced almost the same comparison group as matching on baseline test trends alone. Among comparison schools in the "radius" comparison group, 69% are also included in the "radius with demographics" group. In many situations, this may not be the case. However, there is often a trade-off between matching on test scores and matching on demographic characteristics, such that matching well on one produces a poor match on the other. Thus, when test scores are the outcome of interest, matching on test scores is probably preferable.

Discussion

Overall, our findings suggest that the CITS design does, in the present example, provide internally valid estimates of program impacts when the RD design is used as the CB. Statistical tests confirm that the estimated bias is not statistically significant for any of the estimated impacts, and almost all estimates of bias have small magnitudes. These results are consistent across comparison groups and follow-up years with the one exception involving all schools in the state. The finding that a CITS design can provide results that are internally valid (at least under the conditions examined by the present analysis) is important because randomized experiments at the school-level are not always politically feasible, and thus other rigorous evaluation options are needed. More evidence regarding the validity of NXDs is thus needed.

It is also reassuring that, in the present example, the comparison group did not need to be "geographically local" to obtain internally valid impact estimates. Prior literature has suggested that in many circumstances, selecting a comparison group that is geographically local reduces the bias associated with a NXD substantially and in the case of educational interventions, schools that are in the same district are more likely to have been exposed to the same outside influences that serve as a potential threat to the validity of the CITS design. However, there are situations in which it may not be appropriate (or possible) to restrict the comparison group to schools in the same districts as the treatment schools—for example, when there is spillover from treatment schools to other schools in the district, when an intervention is implemented districtwide, or when an intervention is implemented in schools within a district that are by intention quite different from other schools in that district. When feasible and appropriate,

choosing comparison schools from the same set of districts where the treatment schools are located will probably increase the likelihood that the comparison group will yield internally valid impact estimates, but this does not appear to be a necessary condition for validity in this study. More work is needed to see if these findings can be replicated in other samples.

We also demonstrate that it is possible to improve the precision of CITS impact estimates without undermining their internal validity. For example, the radius matching method that we examined produced internally valid impact estimates but also improved the precision beyond what was obtained with the applicant sample. Other matching methods likely also accrue this added benefit. Again, additional examples would help determine whether this finding is more widely generalizable.

One can also increase the sample size by using all ''untreated'' schools in the *state* or all schools *eligible* for the intervention. In our study, we find that these larger ''unmatched'' groups have higher baseline test scores, and therefore, lack face validity. While, in this example, controlling for preintervention levels and slopes was generally sufficient to obtain unbiased estimates, there are cases when this will not be sufficient. For example, if there were a major program that coincided with the intervention of interest but affected only low-income schools in a state, controlling for preintervention levels and slopes alone might not be sufficient to obtain unbiased estimates. We might also need to make sure that our comparison pool included only low-income schools. The marginally significant effect we found for the comparison group that included all schools in the state in Year 2 underscores this point. Therefore, in some instances, matching methods may be preferred because the results are more likely to reflect a valid counterfactual condition while the precision is almost equal to that of larger unmatched comparison groups.

Finally, the study finds that matching on baseline demographic characteristics did not further reduce bias beyond what could be accomplished by matching on pretests alone. Matching on pretest scores alone was sufficient to ensure that the comparison group's deviation from trend reflected the right counterfactual outcome for RF schools in the follow-up period. However, this conclusion may not be widely applicable. Other studies have found that further matching on demographic characteristics *does* substantially reduce bias (Steiner, Cook, Shadish, & Clark, 2010). In our study, pretest trends are sufficient—and demographics do not help—because baseline test trends are a powerful predictor of future test scores. This happens for two reasons. First, we use 6 years of baseline test scores for matching rather than just one, which strengthens the extent to which baseline scores

can predict scores in the follow-up period. Second, our analysis is conducted at the school level rather than at the student level. School-level test scores are more reliable than student-level scores, and by extension, baseline test scores are more predictive of future test scores at the school level. The fact that both test scores and demographics are more reliably measured at the school level also increases the correlation between these two sets of measures and therefore reduces the amount of additional information provided by demographics once test scores have been taken into account in the matching process (Somers et al., 2013).

In conclusion, our results contribute to the growing base of evidence regarding the validity of NXDs. We demonstrate that the CITS design can produce internally valid estimates of program impacts when pretest scores are available and when a set of reasonable criteria are used to select comparison schools. Our article also contributes to the literature by showing that (1) using a comparison group that is "local" (i.e., from the same set of districts as the treatment schools) is not always a necessary condition for obtaining internally valid estimates of program impacts in the case of educational interventions, (2) further matching on demographic characteristics is not always necessary in the context of the CITS design, and (3) the precision of impact estimates can, in some circumstances, be increased by matching using the radius method without compromising validity.

However, this is only one empirical example. Future studies should assess whether these findings hold in others samples as well. Furthermore, since Glazerman et al. (2003) note that studies that found no impact or indeterminate impacts were more easily replicated, future research should include examples where the impact of the program is statistically significant.

Acknowledgments

The authors thank Kristin Porter and Alexander Mayer for comments on an earlier draft of this article and Rebecca Unterman for her contributions to discussions about the analytical framework. We are also indebted to Larry Hedges for suggesting that we use bootstrapping as an additional tool in our analyses. Finally, we thank Edmond Wong, Nicholas Cummins, Ezra Fishman, Jessica Huff, and Anna Erikson for providing outstanding research assistance.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, the U.S. Department of Education, through Grant R305D090008 to MDRC. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

Notes

1. See Lechner (2011) for a technical discussion of the assumptions behind the difference-in-differences design.
2. Cook, Shadish, and Wong (2008) also recommend matching intact groups rather than individuals. Because we are matching schools rather than students, the groups are intact by definition.
3. Although 74 schools received funding, 5 schools do not have test score data for all 8 school years in the study period (because they opened more recently or were closed). Of the 116 schools that applied for but did not receive Reading First (RF) funding, 99 had complete test score data.
4. In Year 2, there is a slight relationship between the rating variable and the outcome measure, which is marginally significant at the $p = .052$ level.
5. We acknowledge that dropping the rating variable is not technically correct because (a) it does not reflect the initial design of the study (regression discontinuity [RD]) and (b) it does not account inferentially for the fact that an empirical test was used to drop the rating variable. Nonetheless, it is an approach that has precedence in the literature (see, e.g., Lee & Lemieux, 2010; Olsho et al., 2015) and estimating impacts without including the RD assignment variable yields useful insights about the validity of the comparative interrupted time series (CITS) design and so we include it here along with the standard RD estimates. Analyses that use the standard RD estimates as a benchmark yield similar results. More details of analyses that use the standard RD estimate can be found in Somers, Zhu, Jacob, and Bloom (2013).
6. The three categories are (1) the local education agency (LEA) has jurisdiction over a geographic area that includes an area designated as an empowerment zone or an enterprise community, (2) the LEA has jurisdiction over a significant number or percentage of schools that are identified for school improvement under Section 1116(b) of Title I of the Elementary and Secondary Education Act (ESEA), or (3) the LEA has the highest number or percentage of children in the state who are counted by the U.S. Department of Education under Section 1124(c) of Title I of the ESEA. In total, 103 districts in our study state were eligible for RF funds.

7. Matching is undertaken among the pool of schools in eligible districts (rather than among the 99 applicant schools) because some matching methods require a relatively large sample size; therefore, it is technically preferable to use a larger "eligible" pool as the group from which to select comparison schools.
8. For the radius matching weights were used to account for variation in the matching ratio across treatment schools as well as the number of times a comparison school is selected as a match.
9. To determine the optimal radius, we used the radius that minimized the mean squared error of the estimated impact in the last baseline year, since we know that the true impact in the last baseline year was 0.
10. Bootstrap data sets (1,000 in total) were created by sampling with replacement from the original data set of schools for the state. At each iteration, sampling was stratified so as to sample 69 RF schools, 99 schools from the "applicant" pool, 419 schools from the eligible pool, and 93 schools from the noneligible/applicant pool (for a total of 69 RF schools and 611 non-RF schools). For each of these 1,000 data sets, the causal benchmark (CB) was estimated, comparison schools were chosen for the CITS analysis, CITS impacts were estimated, and the difference between each pair of impact estimates was calculated. The bootstrap standard error for a given difference in impact estimates (e.g., between the CB and the CITS estimate based on the radius matching method) is the standard deviation (*SD*) of this difference across the 1,000 bootstrap data sets. The confidence intervals are the 5th and 95th percentiles of the distribution of the difference across the 1,000 data sets.
11. This approach was helpfully suggested to us by Larry Hedges.
12. We use the school-level *SD* (rather than the student-level *SD*) for two reasons. First in the matching literature, standardized mean differences are gauged based on the *SD* for the unit of observation (in this case schools). Second, we do not have student-level *SD*s for many of the outcomes presented in this table. We use the *SD* for all schools in eligible districts because it constitutes the largest relevant pool of schools. We use characteristics in the last baseline year because outcomes are not yet affected by the intervention at this point in time and matching will be based on baseline characteristics.
13. The standard errors and confidence intervals in Figure 3 are parametric not bootstrapped standard errors.
14. Based on the bootstrapped standard errors, we are able to detect differences in estimated bias that are around 0.10 *SD*s, which is also the threshold we established for differences that would be educationally meaningful. As shown in Bloom (1995), when there are enough degrees of freedom for the test statistic to approach a normal distribution (around 20 or more) then for a two-sided test

with 80% power and α equal to .05, the minimum detectable effect can be calculated by multiplying the standard error of the estimate by 2.8.

References

Aiken, L. S., West, S. G., Schwalm, D. E., Caroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, *22*, 207–244.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*, 547–556.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27*, 3–49. doi:10.1177/0193841X02239017

Bloom, H. S., & Riccio, J. A. (2005). Using place-based random assignment and comparative interrupted time-series analysis to evaluate the jobs-plus employment program for public housing residents. *Annals of the American Academy of Political and Social Science*, *599*, 19–51.

Campbell, D. T., & Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law & Society Review*, *3*, 33–54. Retrieved from http://www.jstor.org/stable/3052794

Cook, T. D. (2012). *Peter H. Rossi Award lecture*. Retrieved from http://www.welfareacademy.org/rossi/Rossi_Tom%20Cook_Acceptance%20Remarks_November%202012.pdf

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.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of covariate choice, unreliable measurement and mode of data analysis. *Psychological Methods*, *15*, 56–68.

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Fortson, K., Verbitsky-Savitz, N., Kopa, E., & Gleason, P. (2012). *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.

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

Fretheim, A., Odgaard-Jensen, J., Røttingen, J.-A., Reinar, L. M., Vangen, S., & Tanbo, T. (2013). The impact of an intervention programme employing a hands-on technique to reduce the incidence of anal sphincter tears: Interrupted time-series reanalysis. *BMJ Open*, *3*, *e003355*. doi:10.1136/bmjopen-2013-003355

Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, R. T. (2008). *Reading first impact study: Interim report* (NCEE 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, 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. *Annals of the American Academy of Political and Social Science*, *589*, 63–93.

Gleason, P. M., Resch, A. M., & Berk, J. A. (2012). *Replicating experimental impact estimates using a regression discontinuity approach*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*, 1017–1098.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177. doi:10.1111/j.1750-8606.2008.00061.x

Ho, D., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

Hotz, V. J., Imbens, G., & Klerman, J. A. (2006). Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program. *Journal of Labor Economics*, *24*, 521–566.

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

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*, 165–224.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281–355.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics*, *86*, 156–179.

Mulford, H. A., Ledolter, J., & Fitzgerald, J. L. (1992). Alcohol availability and consumption: Iowa sales data revisited. *Journal of Studies on Alcohol*, *53*, 487–494.

Olsho, L. E. W., Jacob, A. K., Lorrene, R., Patricia, W., Karen, L. W., & Susan, B. (2015). Impacts of the USDA fresh fruit and vegetable program on child fruit and vegetable intake. *Journal of the Academy of Nutrition and Dietetics*, *115*, 1283–1290.

Orr, L. L., Bell, S. H., & Kornfeld, R. (2004). *Tests of nonexperimental methods for evaluating the impact of the New Deal for Disabled People (NDDP)*. London, England: Department for Work and Pensions.

Schneeweiss, S., Maclure, M., Carleton, B., Glynn, R. J., & Avorn, J. (2004). Clinical and economic consequences of a reimbursement restriction of nebulised respiratory therapy in adults: Direct comparison of randomized and observational evaluations. *British Medical Journal*, *328*, 560.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

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

Shadish, W. R., Rodolfo, G., Vivian, C. W., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, *16*, 179–191.

Smith, J., & Todd, P. (2005). Does matching overcome Lalonde's critique of non-experimental estimators? *Journal of Econometrics*, *125*, 305–353.

Somers, M. A., Zhu, P., Jacob, R., & Bloom, H. (2013). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation*. New York City, New York: MDRC.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*, *35*, 311–327.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*, 250–267.

Wong, M., Cook, T. D., & Steiner, P. M. (2011). *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. Northwestern University Institute for Policy Research Working Paper Series, WP 09-11, Northwestern University Institute for Policy Research, Evanston, IL.

Author Biographies

Robin Jacob is a research associate professor at the Institute for Social Research and the School of Education at the University of Michigan. Her research focuses on evaluations of education interventions and evaluation methods. She has a special interest in how policies and programs can affect instructional quality and outcomes in low-income schools.

Marie-Andree Somers is a senior research associate at MDRC, who has experience and expertise evaluating the impact of educational interventions that are aimed at improving children's outcomes. She has used a comparative time series designs to evaluate several educational interventions including ninth-grade academies, a tiered whole-school model with integrated student supports, and preservice teacher training.

Pei Zhu is a senior research associate at MDRC, whose work focuses on experimental and quasi-experimental evaluation design, analyses, and related methodological issues.

Howard Bloom is the chief social scientist at MDRC. He has a PhD in political economy and government from Harvard University. His research focuses on the development of experimental and quasi-experimental methods for estimating program impacts.