Short Comparative Interrupted Time Series Using Aggregate School-Level Data in Education Research

Kelly Hallberg[1], Ryan Williams[2], Andrew Swanlund[2], and Jared Eno[3]

Short comparative interrupted times series (CITS) designs are increasingly being used in education research to assess the effectiveness of school-level interventions. These designs can be implemented relatively inexpensively, often drawing on publicly available data on aggregate school performance. However, the validity of this approach hinges on a variety of assumptions and design decisions that are not clearly outlined in the literature. This article aims to serve as a practice guide for applied researchers when deciding how and whether to use this approach. We begin by providing an overview of the assumptions needed to estimate causal effects using school-level data, common threats to validity faced in practice and what effects can and cannot be estimated using school-level data. We then examine two analytic decisions researchers face in practice when implementing the design: correctly modeling the pretreatment functional form, which is modeling the preintervention trend, and selecting comparison cases. We then illustrate the use of this design in practice drawing on data from the implementation of the school improvement grant (SIG) program in Ohio. We conclude with advice for applied researchers implementing this design.

Keywords: achievement; evaluation; policy; quasi-experimental analysis; research methodology; school/teacher effectiveness

Many programs and policies in education are implemented at the school level. For example, schools adopt new curricula, extend the school day, or introduce professional learning communities to improve teaching and learning. Education researchers are often charged with estimating the effects of these kinds of school-level interventions on students' academic outcomes. Cluster randomized trials, in which schools are randomly assigned to treatment conditions, are the best way to assess the effectiveness of school-level interventions. However, cluster randomized trials may not be feasible, either because random assignment is not ethically, politically, or financially tenable or because researchers are interested in retrospectively examining the effects of interventions that have already been implemented. The increasing accessibility of publically available, longitudinal, aggregate school-level data provides an alternative when an experimental study is not possible: the comparative interrupted time series.

In its simplest form, interrupted time series (ITS) measures the same outcome for a treatment group multiple times before and after the introduction of an intervention, adjusting for any trend in the preintervention data. The effect of the intervention is estimated by examining the difference in outcomes before and after implementation. Adding comparison schools to this simple version of the design, to reduce potential threats to internal validity, converts the ITS design to a comparative ITS or CITS (sometimes know in the economics literature as a difference-in-difference design).

The use of ITS and CITS designs in evaluations to investigate the effects of education programs and policies, especially those implemented at the school level, has increased in recent years. Longitudinal data at both the student and school levels are increasingly available to education researchers, making the design easier to implement. In 2010, only 15% of i3-awarded projects proposed to use an ITS or CITS design to evaluate their intervention; in 2011 and 2012, roughly 40% of projects proposed to use this method (U.S. Department of Education, 2013). The design has been used to study a wide range of education policies and

[1]University of Chicago, Chicago, IL
[2]American Institutes for Research, Chicago, IL
[3]University of Michigan, Ann Arbor, MI

Educational Researcher, Vol. 47 No. 5, pp. 295–306
DOI: 10.3102/0013189X18769302

programs, including school turnaround in Chicago (de la Torre et al., 2012), comprehensive school reform (Miller & Mittleman, 2012), universal class size reduction (Chingos, 2012), guaranteed tuition policies (Delany & Kearney, 2015), zero tolerance disciplinary policies (Curran, 2016), structured transfer pathways in community colleges, and the effect of the No Child Left Behind Act (NCLB) on academic outcomes (Dee, Jacob, & Schwartz, 2013; Wong, Cook, & Steiner, 2015), and childhood obesity (Anderson, Butcher, & Schanzenbach, 2017).

However, the validity of school-level CITS designs hinge on a variety of assumptions and design decisions that are not clearly outlined in the literature. This paper aims to serve as a practice guide for applied researchers when deciding how and whether to use this approach. We will clarify the assumptions underlying the CITS design and the conditions under which it is valid and provide an overview of key decisions researchers face in practice when implementing the design. We begin by providing an overview of the assumptions needed to estimate causal effects using school-level data, common threats to validity faced in practice and what effects can and cannot be estimated using school-level data. We then examine two analytic decisions researchers face in practice when implementing the design: selecting comparison cases and correctly modeling the pretreatment functional form. We focus on the short CITS designs that are frequently implemented in education research for which there are between three and 20 pretreatment measures of the outcome (Bloom, 2003). The penultimate section of the paper illustrates the use of this design in practice drawing on data from the implementation of the school improvement grant (SIG) program in Ohio. We conclude with advice for applied researchers implementing this design. Sample R code for each of the matching and modeling approaches described in this paper can be found in Appendix A.

CITS in the Potential Outcomes Framework

The potential outcomes framework, or Rubin's causal model, which characterizes causal effects as unit-specific differences between outcomes achieved under different treatment conditions, provides a clear theoretical rationale for the conditions under which school-level CITS designs produce unbiased estimates of causal effects (Heckman, 1979; Holland, 1986; Neyman, 1935; Rubin, 1978). Each school can be characterized by a set of variables $(Y_{tj}(1), Y_{tj}(0), T_t, x_j, x_{tj})$. $Y_{tj}(1)$ and $Y_{tj}(0)$ are the potential outcomes for unit j at time t under the treatment and control conditions, respectively. T_j is an indicator of whether school j is ever treated. x_j and x_{tj} are time-invariant (e.g., grade levels served) and time-varying (e.g., demographic composition) school characteristics, respectively. Following Wong, Wing, Steiner, Wong, and Cook (2012), we assume that the potential outcomes are a function of time (t) such that $Y_{tj}(0) = f_0(t) + \varepsilon_t$ and $Y_{tj}(1) = f_1(t) + \vartheta_t$, where $f_1(t)$ is the time trend under the treated condition and $f_0(t)$ is the time trend not under treatment. The treatment effect is the difference in potential outcomes: $\tau(t) = f_1(t) - f_0(t) \mid t \geq t_c$, where t_c is the introduction of the intervention (i.e., the interruption). However, in practice, we see $f_0(t)$ only in the pretreatment period and $f_1(t)$ in the posttreatment period. In the simple ITS (that include only schools in which $T_j = 1$ are available), $f_0(t)$ is unobserved in the posttreatment period and therefore is estimated as an extrapolation based on the functional form of the preintervention time series. This is depicted visually in Figure 1. The lack of a counterfactual in the postintervention period is denoted with the dotted line.

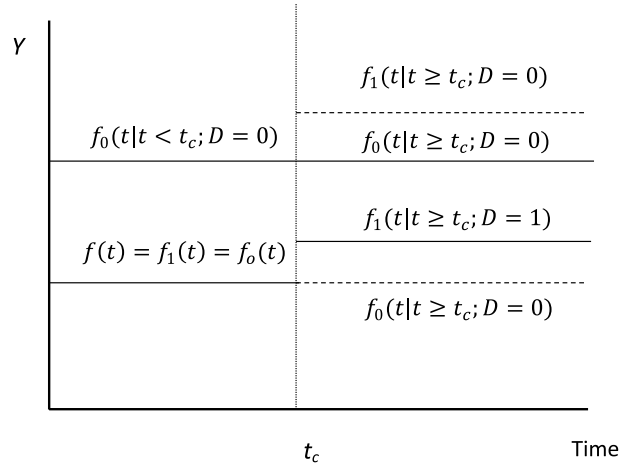

FIGURE 1. *Stylized depiction of the potential outcomes in a CITS design.*

In the simple ITS context, unbiased estimates of the treatment effect can be achieved only if (1) the pretreatment trend is correctly specified, and (2) its projection into the postintervention period is an accurate estimate of the counterfactual (Wong et al., 2013). That is, the past is an accurate predictor of what the future would have been without the intervention. This could not be the case for a variety of reasons, often referred to as threats to validity in the ITS literature:

- **History.** History threats can occur if changes from the pretreatment trend $(f(t) \mid t < t_c;$ the time points prior to the introduction of treatment) occur as a result of other, unrelated changes that act on the treatment units and occur simultaneously to the implementation of the program of focus. For example, if a district implements two mathematics interventions simultaneously, it will be difficult to disentangle which program caused a change in district mathematics achievement.
- **Selection.** In school-level ITS designs, selection can be a threat if students who attend treated schools in the pretreatment period are inherently different from those who attend in the posttreatment period. For example, if a school undergoes a composition change simultaneous to the interruption, the effect estimate could be biased.
- **Instrumentation.** This validity threat refers to changes in the outcome of interest that occur at or near the time of the interruption. For example, if a state changes its testing program, scores may not be comparable from one year to the next or may measure different content (if content standards are changed significantly) (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002).

In an attempt to address these threats, a comparison group often is added to the simple ITS, turning it into a comparative ITS

or CITS. With a comparison group, potential threats must operate differentially across groups to threaten inference (Cook & Campbell, 1979; Shadish et al., 2002). For example, if a state changed their annual assessment, the only way an instrumentation threat would occur is if the change in assessment difficulty was greater for one group than the other. The CITS approach is a generalized version of what the econometrics literature refers to as a difference-in-differences design, in which both the pretreatment outcomes of the treated group and the change in outcomes in the comparison group contribute to estimating the counterfactual outcome for the treatment group (Ashenfelter, 1978; Ashenfelter & Card, 1985; Athey & Imbens, 2006; Imbens, 2009).[1] Returning to the notation used in Figure 1, the treatment effect estimator in the CITS framework can be formulated as follows:

$$\widehat{\tau(t)} = \left[\left(f_1(t)|T_j = 1, t > t_c\right) - \left(f(t)|T_j = 1, t < t_c\right)\right] - \left[\left(f_0(t)|T_j = 0, t > t_c\right) - \left(f(t)\,|\,T_j = 0, t < t_c\right)\right],$$

where $f(t)$ is the preintervention functional form of the outcome time series. This formulation makes clear how CITS designs are an improvement over simple ITS designs. The first term is analogous to the treatment effect in a simple ITS design. The second term differences out secular trends (changes over time that are independent of treatment). Rewriting this equation illustrates how CITS is an improvement over simple matching designs that rely on the similarity of matched groups on observable characteristics to identify causal effects:

$$\widehat{\tau(t)} = \left[\left(f_1(t)|T_j = 1, t > t_c\right) - \left(f_0(t)|T_j = 0\right), t > t_c\right] - \left[\left(f(t)|T_j = 1, t < t_c\right) - \left(f(t)\,|\,T_j = 0, t < t_c\right)\right]$$

In this formulation, the first term can be seen as the estimator in a simple matching design, in which the difference between matched treated and untreated schools serves as the treatment effect estimate. Such an approach relies on the assumption that matching on *observable* characteristics is sufficient for meeting the strong ignorability assumption to estimate unbiased causal effects (Rosenbaum & Rubin, 1983). The second term in the equation differences out time-invariant *observable* and *unobservable* differences between treatment and comparison cases. As such, to estimate unbiased treatment effects, the CITS design requires proper model specification and selection of a comparison group that (1) does not differ from the treatment group in ways that vary over time that are related to the outcome of interest and (2) is exposed to the same history or instrumentation threats as the treatment group.[2] For example, if the preintervention slopes in the treatment and comparison group differ and these differences are not properly accounted for in the modeling approach employed, the estimated effects will be biased.

Meeting These Assumptions in Practice

Results from several recent within study comparisons (WSCs) provide some reason for optimism about the performance of CITS in education. WSC studies empirically estimate the extent to which a given observational (nonrandomized) study reproduces the result of an RCT when both share the same treatment group. In principle, in studies like these, the only thing that differs between the RCT and observational study is how the comparison group is formed—at random versus systematically (for example, see Cook, Shadish, & Wong, 2008; Glazerman, Levy, & Myers, 2003; Lalonde, 1986). WSCs both within and outside of education have shown that CITS can produce results that are very similar to those from an RCT (Fretheim, Soumerai, Zhang, Oxman, & Ross-Degnan, 2013; Jacob, Somers, Zhu, & Bloom, 2016; Schneeweiss, Maclure, Carleton, Glynn, & Avorn, 2004; St. Clair, Cook, & Hallberg, 2014; St. Clair, Hallberg, & Cook, 2016). However, in some cases, this correspondence is dependent on modeling choices made by the researcher as well as the stability of the pretreatment trend (St. Clair et al., 2014 , 2016). Hallberg, Williams, and Swanlund (2017) explicitly examine the performance of CITS implemented with school-level data in three WSCs. They show that while the design is robust to a variety of modeling and comparison group choices 1 year postintervention, substantial bias can be seen when examining outcomes further out from the implementation of the treatment.

Implications of Using School-Level Data in CITS

The growing availability of aggregate school-level data on state websites has led to an increased interest in conducting analyses at the school level. Jacob, Goddard, and Kim (2014), for example, argue that the use of aggregate school-level data reduces the costs associated with pulling student-level records. Stuart (2007) extended individual case matching to matching schools using aggregate data. The availability of multiple years of longitudinal aggregate data on many states' department of education website makes CITS using school-level data a natural extension in this direction. CITS analyses using publically available data can be done relatively inexpensively with minimal burden to already overextended state and district research offices.

However, when conducting school-level analyses, it is important to be clear about exactly what effect one is estimating (i.e., the estimand). School-level CITS estimates the difference in school performance under treatment and comparison conditions by comparing cohorts of students that attend treatment and comparison schools over time. It does not provide an estimate of what would have happened to individual students or groups of students under the two treatment conditions. Misinterpretation of estimates from school-level analyses as student-level effects can lead to an ecological fallacy or Simpson's paradox (Freedman, 2001), where outcomes at the student and school levels may operate differentially.

School-level CITS estimates do not account for changes in the composition of students in schools over time, sometimes referred to as "stayers, leavers, and joiners" in the RCT literature (What Works Clearinghouse, 2018). As such, school-level CITS estimates can be seen as estimating the effect of the combination of two forces: the change in the composition of students in the school that results from the introduction of a new intervention as well as the change in performance of the students in the school that results from the introduction of the new intervention. For some interventions, this combined effect may be of policy interest. For example, the introduction of a new magnet program in

a school may be intended to both draw talented students to the school and improve the performance of students who attend. For other interventions, this might be less appropriate. In these cases, the substantive question is the effect of introducing the intervention to the group of students that would have attended the school whether or not the intervention had been implemented. In these cases, analysts should examine the extent to which compositional shifts have accompanied the introduction of the intervention. Statistical controls can be included in the model to account for observable shifts. For example, analysts might include the percent of students who qualify for free and reduced price lunch in each year as a control variable in the model. Such statistical controls cannot account for unobserved compositional changes, such as more motivated students moving into treatment schools. Choosing what controls to include and whether those controls are sufficient should be driven by a thorough understanding of an intervention's theory of change.

Finally, applied researchers should be careful in calculating and interpreting effect sizes in school-level CITS studies. Researchers may use the standard deviation of average school performance to calculate effect sizes, but these are not equivalent to the student-level effect sizes commonly estimated in education studies because the standard deviation of average school performance is generally much smaller than that of student performance. To estimate comparable effect sizes, researchers should use student-level standard deviations (often available in state testing reports) or estimate the student-level standard deviation by dividing the school-level standard deviation by the square root of the intraclass correlation (What Works Clearinghouse, 2018). Hedges and Hedberg (2007) provide guidance for estimating the intraclass correlation if it is not known.

Selecting a Comparison Group

A key analytic decision that analysts employing the CITS designs need to make is how to identify a comparison group. As we discussed above, the comparison group is added to the simple ITS to address potential threats to validity, such as history, instrumentation, or selection. However, until recently, little guidance has been available to guide analysts in selection of a comparison group in the context of CITS. The criteria for selecting a comparison group in this context differ a bit from those in a pure matching study. In a matching study, the analyst's goal is to match on all characteristics that are related both to treatment status and outcomes. In the CITS context, the analyst only must match on the subset of characteristics that are related to treatment status and outcomes *and* vary over time. That is, in the matching context differences in preintervention outcomes could introduce substantial bias to the impact estimate, but in the CITS context any fixed differences between treatment and comparison schools are accounted for. In this section, we consider the implications of four commonly used approaches to identifying a comparison group in CITS: using all available nontreatment schools, matching on preintervention measures of the outcome or other observable characteristics, local matching, and a hybrid approach that balances local and focal matching.

All available nontreatment schools. This approach compares treatment schools to nontreatment schools in the same district, state, or country. For example, Dee et al. (2013) and Wong et al. (2015) examined the effect of NCLB by comparing the performance of implementing states to nonimplementing or lower implementing states. No attempt was made to find cases that were similar in preintervention trend or other observable characteristics. While this approach would be inadvisable in most matching studies, where it is often referred to as the "naïve treatment effect," it can provide unbiased estimates of causal effects in the CITS context as long as time-varying confounds do not operate differentially across the treatment and comparison groups.

Matching on preintervention characteristics. Researchers also could select a subset of schools that are similar to the treatment schools on preintervention measures of the outcome or other school-level characteristics. This could be done using a variety of matching methods, including a nearest neighbor or n to 1 matching method, radius matching (Jacob et al., 2016), or synthetic matching (Abadie, Diamond, & Hainmueller, 2010). Nearest neighbor, n to 1, and radius matching approaches identify intact schools that most closely resemble the schools that opted into treatment, whereas synthetic matching approaches reweight existing available comparison cases to most closely approximate the treatment schools. Although each approach has advantages and disadvantages in terms of ease of application, transparency to a policy or practitioner audience, and coarseness of the matches that result, the matching literature generally suggests that the approach to matching is generally not as salient as the covariates on which one matches (Steiner, Cook, & Shadish, 2011).

The decision of which covariates should be included when matching in the CITS context should be informed by the threats to validity inherent in this design. While in a pure matching study, researchers are trying to identify all preintervention covariates that are associated with selection into treatment and the outcomes of interest, in a CITS context researchers are focused on covariates that are time varying or characteristics that could be associated with history threats (e.g., researchers might want to match on percent of students qualifying for free or reduced-priced lunch if they are concerned that a policy change targeting low-income students could co-occur with the introduction of the intervention of interest). In terms of the former, preintervention measures of the outcome of interest are particularly salient, both in terms of level and trends. Preintervention measures can be included in any of the matching approaches described above, either by matching on each individual preintervention measure or by matching on the preintervention mean and slope.

No matter what covariates or matching approach is employed, identifying a matched comparison group may represent a trade-off in statistical power (the sample is smaller than would be the case if all available comparison schools were used). However, this precision trade-off may be worthwhile if sufficient bias reduction results, simplifying the modeling demands on the CITS.

Local matching. Limiting the comparison pool to geographically local matches has a strong tradition, especially in the job-training literature (Bell, Orr, Blomquist, & Cain, 1995; Bifulco, 2012; Bloom, Michalopoulos, & Hill, 2005; Friedlander & Robins, 1995; Heckman, Ichimura, Smith, & Todd, 1998). The

logic behind local matching is that schools that are geographically proximal often are similar in both observable and unobservable ways. Schools within the same school district, for example, often have similar student-teacher ratios and are similar in unobserved ways, such as district policies, community perceptions of schools and the importance of schooling, and the labor markets that graduates of the public schools will enter. In the CITS framework, local matching has the added benefit of decreasing the likelihood of history confounds because schools that are geographically proximal are more likely to experience similar events than those that are not. For example, schools in the same district will be exposed to the same district-level policy changes, leadership changes, and budgetary or economic shifts.

Hybrid matching. Another approach might be to select comparison schools that are drawn from the same local area (e.g., school district) and that are similar in their preintervention trends. In practice, however, finding comparison schools that fit both criteria may not be feasible. This trade-off between local and so-called focal matching has been noted in the matching literature (Hallberg, Wong, & Cook, 2017; Stuart & Rubin, 2008). Stuart and Rubin (2008) introduced an approach to addressing this trade-off between local and focal matching. Their approach draws on two populations of students as potential matches for treated students in a given district: students within the same school districts and a nonlocal group of students located in a different state. Observable preintervention characteristics are used to calculate each student's propensity to take up the treatment. Students are matched to students within the same school district if they are within a certain caliper (maximally acceptable difference in propensity scores) of one another on observable characteristics (e.g., 0.75 standard deviations of the propensity score). If a matched student is not available within this caliper, then a student is drawn from the nonlocal group of comparison students. Hallberg et al. (2017) found that this approach could reduce bias when used in intact school matching. However, the approach has not been explored for finding a comparison group in the CITS context, nor has the relative importance of local and focal matching been studied systematically.

To date, evidence on the value of various approaches to selecting comparison cases is mixed. Betts et al. (2010) conducted a simulation study for the Institute for Education Sciences and found that matching schools to all other schools in the same district performed better than matching on preintervention covariates, but the simulated setup of their analysis did not attempt to recreate selection processes as they would occur in practice, instead using random draws to create null treatment effects. Jacob et al. (2016) and St. Clair et al. (2014) examined the performance of using all other schools in the state as a comparison group versus matching on preintervention characteristics and found that both approaches performed comparably in terms of bias reduction. St. Clair et al. (2016) found some evidence that matching on preintervention measures and demographic characteristics of the outcome led to reduced sensitivity to modeling choice and closer correspondence to the benchmark than simply using all available comparison cases. Little work has been done to examine the relative performance of different approaches for matching on preintervention outcomes (e.g., nearest neighbor, radial, or synthetic matching) or on the trade-off between local and focal matching in the context of CITS.

Approaches to Modeling CITS

As described above, CITS draws on data from both the preintervention outcomes in the treatment schools and the pre-post treatment outcomes in the comparison schools to estimate the counterfactual outcomes for the treatment schools in the posttreatment period. Correctly modeling the preintervention functional form is key to the validity of this design. Several different modeling approaches have been suggested to estimate effects in CITS. Bloom (2003) outlined three main modeling approaches: the baseline mean model, the linear baseline trend model, and the nonlinear baseline trend model. Another approach commonly used in the econometrics literature, the year and school fixed effects model, is also frequently used.

Baseline mean model. The baseline mean model is the simplest of the modeling approaches and closely resembles the simple difference-in-differences approach commonly used in the econometrics literature. This modeling approach assumes the differences between treatment and comparison cases are fixed (i.e., that in the absence of treatment, the distance between the treatment group slope and the comparison group slope would be constant across preintervention and posttreatment periods). In practice, we assess the validity of this assumption by examining preintervention trends. Preintervention slopes need not be flat (i.e., zero) so long as they are parallel between treatment and comparison groups. The average preintervention performance is projected into the posttreatment period as the best estimate of performance in the absence of treatment. The difference between the average preintervention and postintervention performance in the treatment schools, less this same difference in the comparison schools, serves as the estimate of treatment effects. The baseline mean model can be formulated as follows:

$$Y_{jt} = \beta_0 + \beta_1 Z_{jt} + \beta_2 post_t + \beta_3 trt_j + \beta_4 post_t trt_j + v_j + u_{jt} \quad (1)$$

Where Y_{jt} is the outcome for school j at time t; β_0 is a constant term showing average achievement in comparison schools before the intervention; Z_{jt} is a vector of school characteristics at time t; β_1 is a vector of coefficients associated with each of those covariates showing the association of each school-level characteristic and the outcome; $post_t$ is a vector of indicators for each postintervention time period t. Alternatively, an indicator variable of whether a given year was in the posttreatment period could be included. This would provide one average estimate of the effect of the intervention in the posttreatment period rather than an estimate of the program effect for each postintervention year as would be the case in this model. β_2 is a vector showing the difference in average outcomes between the preintervention time period and each postintervention time period t for comparison schools; trt_j is an indicator for whether a school received the intervention of interest; β_3 shows the average difference in performance between treatment and comparison schools in the preintervention time period; β_4 is a vector showing the change in

the difference in average performance between treatment schools and comparison schools at each time t after the intervention was implemented (i.e., the treatment effect in each of the postintervention years); v_j is a school-level random error term, with an assumed normal distribution with mean zero and variance φ^2; and u_{jt} is a year-level random error term, with an assumed normal distribution with mean zero and variance τ^2. For this and the other modeling approaches described here, we employ random effects models rather than more formal time series modeling approaches, such as ARIMA models, because the number of time points generally available in education research are normally not sufficient to support these kinds of modeling approaches.

Baseline linear-trend model. The linear baseline trend model accounts for differences in preintervention trends by including a linear term for time ($\beta_1 time_t$) as well as an interaction of this term with the treatment indicator ($\beta_5 time_t trt_j$), as shown in Model 2:

$$Y_{jt} = \beta_0 + \beta_1 time_t + \beta_2 Z_{jt} + \beta_3 post_t + \beta_4 trt_j + \beta_5 time_t trt_j + \beta_6 post_t trt_j + v_j + u_{jt} \quad (2)$$

β_1 is now the preintervention slope in the comparison group, and $\beta_1 + \beta_5$ is the preintervention slope in the treatment group. The difference in the actual posttreatment performance from the projected posttreatment performance in the treatment schools, less this same difference in the comparison schools, serves as the estimate of treatment effects (β_6). This formulation assumes that all treatment schools share the same trend and all comparison schools share the same trend (though possibly different from the treatment trend). However, this assumption could be relaxed by modeling the trends as random effects. Relaxing this assumption would mean that each school has its own trend adding more variance to the estimating equation, and that schools with lower variance in their time series will carry more weight in estimating the relationship between time and outcomes. Further, if for substantive reasons investigators were interested in whether the slope of performance changed after the introduction of the intervention, the researchers could code T_t as a dichotomous variable that takes the value of 0 in the preintervention period and 1 in the posttreatment period (rather than as a vector of indicators for each postintervention time period t). The change in slope could then be estimated adding a three-way interaction between the treatment indicator (trt_j), the postintervention indicator ($post_t$), and the linear time trend ($time_t$). In this formulation, $time_t$ should be coded to be centered on the introduction of the intervention, so that β_6 can be interpreted at the immediate shift in outcomes following the introduction of treatment. Inclusion of this interaction could be warranted if treatment effects are expected to grow or decline over time (e.g., when a school selects a new curriculum and full implementation and program impacts take a while to take hold). However, one should be cautious of interpreting a change in slope because estimating this change relies on differences between the treatment and control groups further out from implementation, which as we note above, tend to be estimated with more bias.

Baseline nonlinear-trend model. The nonlinear baseline trend model is an extension of the linear baseline trend model that addresses a more complicated functional form in the relationship between time and the outcome of interest. This approach is implemented in a way that parallels the linear time trend in Equation 2 by replacing the linear time parameter with a function of time ($f(time_t)$, such as a first degree polynomial if change in school outcomes is expressed as a quadratic form. Because both the linear and nonlinear baseline trend models involve the inclusion of additional parameters and thus the use of additional degrees of freedom, these approaches have greater data requirements than the baseline mean model. In particular, while the baseline mean model may be estimated with as little as 1 preintervention year of data (becoming a difference-in-differences design) and the baseline trend model may be estimated with 2 years of preintervention data if one assumes that all schools share the same trend, the nonlinear baseline trend model will require 3 or more years of data, depending on the complexity of the nonlinearity.[3] In the short CITS designs, modeling higher order polynomials is infrequently feasible and can exaggerate bias due to overfitting.

School and year fixed effects model. An approach that was not mentioned in the Bloom (2003) paper but is often used in the econometrics literature is a model that does not explicitly model the preintervention trend but includes school and year fixed effects. This can be seen as a more flexible modeling approach because it does not impose any functional form assumptions on the relationship of student achievement over time. Unlike the baseline mean model, this approach uses only the variation within schools to estimate treatment effects. Both models, however, assume parallel preintervention time series. The year fixed effects serve to account for year-to-year deviations across schools. This can be seen as an extension of the baseline mean model which accounts for the year to year deviations in school performance. The treatment effect estimate is the difference-in-difference demeaned to account for overall performance over the study period. The model can be formulated as follows:

$$Y_{jt} = \sum_{t=0}^{T} \beta_t year_t + \beta_{trt} trt\_post_{jt} + \sum_{k=0}^{N} \beta_{sk} s_k + u_{jt} \quad (3)$$

Where Y_{jt} is the outcome for school j at time t; $year_t$ is a vector of indicator variables for each year in the study period (year fixed effects for both pre- and postintervention years); β_t is a vector of coefficients associated with each of the year fixed effects; trt_post_{jt} is an interaction of two indicator variables, one signifying whether a school is in the treatment condition and the other whether the year is in the postintervention period. This variable is always 0 for comparison cases, 0 for treatment cases in the preintervention period, and 1 for treatment cases in the postintervention period; β_{trt} is the difference in average performance between treatment schools and comparison schools in the period after the intervention was implemented, net of school and year fixed effects; S_k is a vector of school indicator variables (school fixed effects); and β_{sk} is a vector of coefficients associated with each of the school fixed effects.

Choosing among models. Little guidance is available to applied researchers regarding selecting from among these modeling

approaches in a particular case. Bloom (2003) argues that the baseline mean model is the "least risky" of the models he considers because it avoids large errors associated with incorrectly specifying the slope. However, the baseline mean model itself assumes that year-to-year variations in the preintervention trend are essentially random variation around the school's mean performance and not evidence of a consistent increase or decrease, essentially constraining the slope to be zero. If this assumption does not hold, the baseline mean model can lead to bias. In fact, St. Clair et al. (2014) found that for at least one of the outcomes they examined in a WSC, use of the baseline mean model, led to biased results. Notably, the bias increased as additional years of preintervention data were included in the model. This result showed evidence of a difference in preintervention trend across treatment condition, as theory would predict. Somers, Zhu, Jacob, & Bloom (2013), however, found no difference in performance between the baseline mean model and the linear baseline trend model in another empirical application.

To our knowledge, St. Clair et al. (2016) provide the most thorough examination of the implications of modeling decisions in short CITS to date. They examined the performance of the baseline mean and the baseline trend modeling approaches. The data in each of the three WSCs exhibited a different pattern of preintervention outcomes. In the first dataset, preintervention outcomes in the treatment and comparison cases were relatively flat and parallel over time. In the second datasets, inspection of the preintervention data revealed evidence of differential slopes; the treatment group was at a different rate than the comparison group. In the third dataset, the pretreatment outcomes were characterized by unclear functional forms in which performance fluctuated from year to year without displaying a clear pattern. The authors found that in the first two datasets, employing the modeling approach suggested by visually inspecting the pretreatment data—the baseline mean model for dataset 1 and the baseline trend model for dataset 2—led to very close correspondence between the CITS and the RCT. In the case of the third dataset, the degree of correspondence was more mixed. In this dataset, the closest correspondence to the RCT was found when modeling approaches were combined with matches on pretreatment measures of the outcome and demographic characteristics.

These results suggest that, as theory predicts, modeling choices matter in CITS. Applied education researchers should closely inspect pretreatment data to select the modeling approach that best fits their data. Further, in cases in where there is no clear functional form in the pretreatment period, for example when only a very small number of pretreatment data points are available, analysts should proceed with caution.

Advice for Applied Researchers

Table 1 below provides a summary of guidance around each phase of implementing a CITS design with school-level data. It is intended to be a resource to help researchers navigate decision points during the design and analysis for a CITS study. The following section provides on example of how these decisions might be made in practice.

An Example of CITS in Practice: The Effect of Receiving a School Improvement Grant in Ohio

To demonstrate how CITS is implemented in practice, we turn to an applied example: studying the effect of receiving a School Improvement Grant (SIG) in Ohio. The federal SIG program provided states with resources to make competitive subgrants to local education agencies (LEAs) to improve the performance of their lowest performing schools. SIG schools could use the resources to implement one of four approved school improvement. In 2010–2011, the first cohort of SIGs was distributed among 41 Ohio schools in 11 LEAs.

Evaluating the effectiveness of the SIG program provides a useful example of implementing CITS in practice using aggregate data. The effects of interest are at the school level, and necessary data are publically available. Our team collected, from the Ohio Department of Education website,[4] school-level academic achievement data from 2004 to 2014. Academic achievement data, including aggregate (mean) scale scores on the state standardized test, were disaggregated by grade, subject, and year. That is, for each year, for each grade, we had the average scale score in reading and mathematics[5] for all students tested. We also collected school-level demographic data from the Elementary and Secondary Information System (ElSi; formerly known as the Common Core of Data). Demographic data on school composition included the following variables: free or reduced priced lunch; race, gender, school level, Title I eligibility, school size, and urbanicity. The outcomes of interest were standardized by grade, year, and subject to account for any changes in the scale scores over time.

Is school-level CITS the right design?. The first major decision in conducting a school-level CITS is confirming that it is the right design. SIG was implemented as a larger policy initiative, outside of an experimental research setting. The 41 Ohio schools that received funds demonstrated a need for the funds and were historically low performing. Given the nature of the program, inferences about school-level effectiveness are especially relevant. CITS is a promising quasi-experimental design option for evaluating SIG, especially because historical school-level performance data are available for all public schools in Ohio.

Selecting a comparison group. The second major decision in conducting a school-level CITS is determining how to select a comparison group. We considered each of the four approaches described above for selecting a comparison group: using all non-SIG schools in Ohio (all available nontreatment cases); creating a matched comparison from all other non-SIG schools in the state using pre-SIG achievement and demographic data (matching on preintervention characteristics); using all other non-SIG schools in districts receiving SIG funding (local matching); and an approach that balances finding a local match with finding comparison cases that are similar on pretreatment characteristics (hybrid matching). In deciding between these approaches, we carefully considered the specifics of the SIG program. First, the program was designed to target persistently low performing schools in the state. Across the state, persistently low performing

Table 1
Design Considerations and Guidance

Design Considerations	Guidance
Is school-level CITS the right design?	• Is an RCT feasible? If so, an RCT is preferable to a CITS design because it requires fewer assumptions to support causal inference. • Is your research question about school-level effects? School-level CITS estimates the difference in school performance under treatment and comparison conditions. • We recommend having at least three preintervention and one postintervention measure of the outcome. The availability of publicly available school-level data has increased the ease of implementing this design, but researchers should be aware of changes in assessments that could bias effect estimates.
How do you create a comparison group?	• What other changes could explain a change in outcome that co-occurs with the implementation of the treatment? Comparison schools should be selected to rule out potential threats to validity. • Correctly selecting comparison cases can also help simplify modeling assumptions. To take advantage of this design feature, select comparison cases with similar preintervention trends in outcomes. • A variety of approaches, including using all available nonimplementing schools, local/within district matching, focal matching on observable characteristics, and a hybrid approach that balances finding local and focal matching, are all possibilities to consider.
How do you model outcomes?	• Are the preintervention trends equivalent between the treatment and comparison schools? If so, the baseline mean model will maximize efficiency without introducing bias. However, if the preintervention trends are divergent, more flexible modeling approaches, such as the baseline trend model, should be used. Visual inspection and more formal statistical tests comparing preintervention trends can both be used to make this assessment. • Postintervention outcomes can be examined as annual differences between the treatment and control cases in the postintervention period or whether there is a shift in intercept and slope. If the latter approach is employed, time should be centered on the intervention year. • Staggered implementation provides additional opportunities to vet internal validity, but must be taken into account in modeling.
How should results be interpreted?	• Because more assumptions must be met for a CITS to be causally valid, they should be interpreted more cautiously than results from an RCT. • The longer the lag between the collection of outcome data and the implementation of the treatment, the less confidence one should have in a causal interpretation of the estimated effects. • School-level CITS only provides estimates of school level effects, not what would have happened to individual students or groups of students under the two treatment conditions. • Like all experimental and quasi-experimental studies, the effects from CITS studies only provide estimated effects for the schools that actually implement the program. This may or may not be the estimand of interest for policy or practice. Although the naturalistic setting under which CITS designs are frequently implemented may lead to greater external validity than RCTs that are limited to schools that are willing to be randomized, researchers should still think carefully before generalizing beyond the study population.

schools are likely similar in a variety of ways. In addition, several other large-scale programs or policy initiatives were introduced during the post-SIG period, including a new state funding formula that reallocates aid away from top performing districts (2011); a new teacher evaluation system that tied half of evaluation scores to student performance (2011); Race to the Top funding (2011); adoption of Common Core (2011); federal stimulus funding for Ohio public schools (2010 and 2011); and doubling the number of school vouchers available in low-performing districts (2011). Most of these initiatives were statewide initiatives, but several were specifically targeted at low-performing schools. Because these initiatives coincide with the introduction of 2010–2011 SIG funding, it becomes critically important to find matched comparisons that would mirror the effects these programs might have on SIG schools. This would suggest that matching on treatment characteristics would be particularly useful.

At the same time, Ohio, like other states, awarded SIG funding competitively to districts that in turn selected persistently low performing schools for the program. This suggests that district context could be particularly important to control for in this case. Districts that successfully sought out SIG funding might be engaging in other initiatives designed to the performance of their lowest performing schools. Moreover, SIG schools may be among the better managed schools across the state.

Ideally, one would use schools that are within district but also similar on pretreatment performance. However, in some targeted districts, all low-performing schools received funding. This is a situation for which hybrid matching is well suited because it preferences local matches when they are similar enough but draws on matches from outside the district when they are not available. To implement the hybrid approach, we first calculated propensity scores for all schools in the state, then computed propensity score differences between each of the SIG schools against each of the potential comparison schools. Using a prespecified caliper (maximally acceptable difference in propensity scores), we then identified all schools that sufficiently match each of the

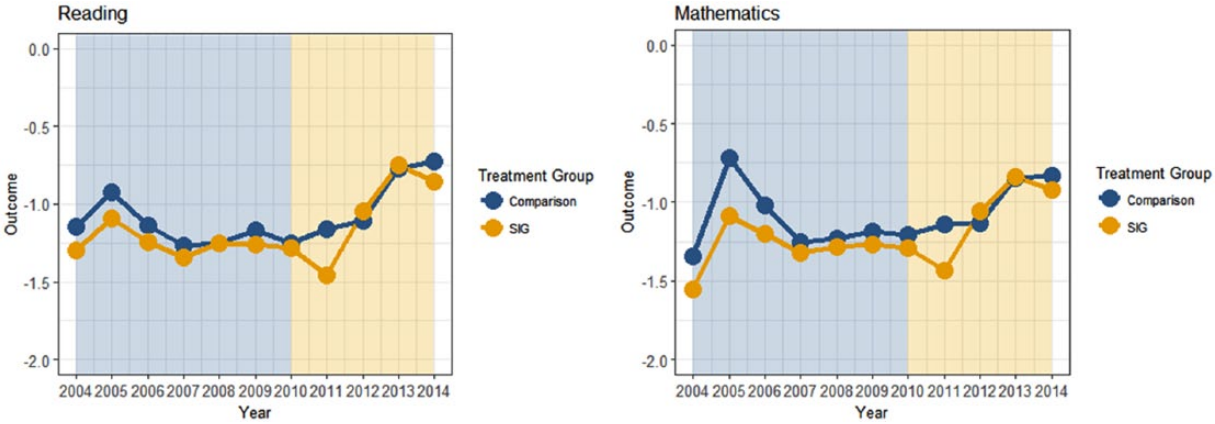

FIGURE 2. *Achievement trends in SIG schools,hybrid matched comparison schools.*

Table 2
Ohio SIG CITS Results, Baseline Mean and Hybrid Matched Comparison

Subject	Year	Estimate	SE
Mathematics	2011	–0.07	0.09
	2012	0.14	0.09
	2013	0.01	0.09
	2014	–0.01	0.10
Reading	2011	–0.09	0.09
	2012	0.11	0.10
	2013	0.00	0.10
	2014	–0.07	0.10

SIG schools, with a target ratio of two matches for each SIG school. Then, any acceptable matching that was available within a SIG district was prioritized over schools in other districts, to the extent they were available, even if there were slightly better matches elsewhere. We matched 37 of the 40 SIG schools to 57 unique comparison schools. Over half (60 percent) matched comparisons from this approach were within-district matches.

Modeling outcomes. The third major decision that needs to be made when conducting a school level CITS is how to model the outcomes (i.e., the pretreatment trends). Figure 2 illustrates the pre- and posttreatment achievement trends in the Ohio SIG schools and the comparison group identified through hybrid matching. Based on visual inspection of these figures, there appears to be little evidence of differential slopes in the treatment and comparison groups. As such, we decided to implement the baseline mean model.

Interpreting results. The results do not provide strong evidence that SIG had a positive impact on school-level achievement, with an initial drop in performance in 2011 followed by a steady increase in performance over the following 3 years, similar to the comparison schools. The effects for each postintervention year are provided in Table 2. While we suggest specifying one's preferred modeling approach before running the analysis, the other modeling approaches can be implemented as a robustness check. In this case, using the baseline trends or year and school fixed effects modeling approaches do not substantively change the results.

As a reference guide, we have summarized the decision-making process for this CITS analysis in Table 3. We have also provided, in Appendix A, sample code for conducting the CITS modeling in R.

Conclusions

CITS designs using aggregate data offer promise to applied education researchers. The designs are fairly straightforward to implement using data that are frequently publically available. In addition, findings from within-study comparisons suggest that these designs *can* replicate the findings from RCTs. However, there are no one-size-fits-all approaches to implementing these designs. Rather in each application, researchers must assess whether the assumptions that undergird this design hold. In this case, is the past an accurate predictor of what the future would have been had the intervention not been implemented? Careful selection of both the comparison group and modeling of the pretreatment trends are critical to ensuring that this assumption holds.

Table 1 summarizes the most up-to-date guidance from the empirical and theoretical literature on how to make these decisions in practice. We believe the field would benefit from additional empirical WSCs that expand on this knowledge base. Insights from this work could serve to improve the implementation of CITS designs and the evidence they generate to inform policy and practice. For the promise of this design to be fully realized, such evidence is needed to help applied research know a priori which approaches to modeling and selection of a comparison group are most likely to yield estimates with moderate or no bias.

Table 3
Design Considerations and Guidance Applied: Ohio SIG

Design Considerations	Guidance
Is school-level CITS the right design?	• Ohio SIG funds were allocated to persistently low performing school on the basis of an application, rendering an RCT infeasible. • The policy question of interest is the effect of the SIG program on school performance. The SIG program could affect schools either through improving the performance of students in the school or changing the composition of students who attend the school. • We have school-level data for 7 years preintervention and 4 years postintervention. • For these reasons, *school-level CITS is an appropriate research design for this study.*
How do you create a comparison group?	• District-level policy changes that co-occur with the introduction of the SIG program are the most likely history threats to validity in this application as district-level policy changes more frequently than state-level policy. • However, because the policy specifically targeted the lowest performing schools in the state, matching on preintervention outcomes and trends is also important. • To balance these factors, the hybrid model was selected as the best approach to *matching.*
How do you model outcomes?	• There were no significant differences in the preintervention outcome trends, so *we employed the baseline mean model to estimate program effects.*
How do you interpret the results?	• The results from the CITS analysis suggest that the SIG program did not have significant effect of academic outcomes in the targeted schools in Ohio. • Because more assumptions must be met for a CITS to be causally valid, these results should be interpreted more cautiously than results from an RCT. In particular, the longer the lag between the collection of outcome data and the implementation of the treatment, the less confidence one should have in a causal interpretation of the estimated effects. • The estimates are school-level effects and could be influenced both by changes in performance among students who attend SIG schools or by changes in the composition of these schools. • The effects are our best estimate of the effects for the schools that opted into the SIG program in Ohio. They don't tell us what would have happened if a broader set of schools implemented the program.

NOTES

Funding for this article was provided by Institute of Education Sciences (grant no. R305D140030).

[1]Note that some scholars use the terms *difference-in-differences* and *CITS* interchangeably, whereas others use *difference-in-differences* to refer exclusively to cases in which there are only two time points (pre and post) or the subset of CITS designs that we refer to below as the baseline mean model.

[2]Lechner (2010) provides a more formal discussion of these assumptions, which he refers to as the common trends and common bias assumptions, as well as a proof that when these assumptions are met, CITS/difference-in-difference analyses can provide unbiased estimates of the average effect of treatment for the treated.

[3]This assumes more than one school and that the preintervention trend is the same for all schools. Otherwise, more data may be required.

[4]Staff at the department of education assisted with providing additional years of school-level data that were not available directly from the state website.

[5]Data are also available for writing, social studies, and science.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Anderson, P. M., Butcher, K. F., & Schanzenbach, D. W. (2017). Adequate (or adipose?) yearly progress: Assessing the effect of "No Child Left Behind" on children's obesity. *Education Finance and Policy*, 121, 54–76.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, *60*, 47–57.

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics*, *67*, 648–660.

Athey, S., & Imbens, G. W. (2006). Identification and inference in non-linear difference-in-difference models. *Econometrica*, *74*(2), 431–497.

Bell, S. H., Orr, L. L., Blomquist, J. D., & Cain, G. G. (1995). *Program applicants as a comparison group in evaluating training programs: Theory and a test*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. https://doi.org/10.17848/9780585284545

Betts, J., Levin, J., Miranda, A. P., Christenson, B., Eaton, M., & Bos, H. (2010). An evaluation of alternate matching techniques for use in comparative interrupted time series analyses: An application to elementary education. American Institutes for Research Working Paper.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management*, *31*(3), 729–751.

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

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 173–235). New York, NY: Russell Sage Foundation.

Chingos, M. M. (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review*, *31*(5), 543–562. https://doi.org/10.1016/j.econedurev.2012.03.002.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cook, T. D., Shadish, W. J., & Wong, V. C. (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management*, *27*(4), 724–750.

Curran, F. C. (2016). Estimating the effect of state zero tolerance laws on exclusionary discipline, racial discipline gaps, and student behavior. *Educational Evaluation and Policy Analysis*, *39*(4), 647–668.

Dee, T. S., Jacob, B., & Schwartz, N. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*, 252–279.

Delaney, J. A., & Kearney, T. D. (2015). The impact of guaranteed tuition policies on postsecondary tuition levels: A difference-in-difference approach. *Economics of Education Review*, *47*, 80–99. https://doi.org/10.1016/j.econedurev.2015.04.003.

de la Torre, M., Allensworth, E., Jagesic, S., Sebastian, J., Salmonowicz, M., Meyers, C., & Gerdeman, R. D. (2012). Turning around low-performing schools in Chicago: Summary report. Chicago, IL: University of Chicago Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/publications/12CCSRTurnAround-3.pdf.

Freedman, D. A. (2001). Ecological inference and the ecological fallacy. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia for the social and behavioral sciences* (Vol. 6, pp. 4027–4030). New York, NY: Elsevier.

Fretheim, A., Soumerai, S. B., Zhang, F., Oxman, A. D., & Ross-Degnan, D. (2013) Interrupted time-series analysis yielded an effect estimate concordant with the cluster randomized controlled-trial result. *Journal of Clinical Epidemiology*, *66*(8), 883–887.

Friedlander, D., & Robins, P. K. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *The American Economic Review*, *85*(4), 923–937.

Glazerman, S., Levy, D., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, *589*, 63–91.

Hallberg, K., Williams, R. T., & Swanlund, A. (March 2017). Examining the internal validity of school-level comparative interrupted time series designs using randomized experiment causal benchmarks. Paper presented at the Annual Meeting of the Society for Research on Educational Effectiveness, Washington, DC.

Hallberg, K., Wong, V. C., & Cook, T. D. (2017). Evaluating methods for selecting school-level comparisons in quasi-experimental designs: Results from a within-study comparison. Manuscript submitted for publication.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrika*, *47*, 153–161.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). *Characterizing selection bias using experimental data* (No. w6699). National Bureau of Economic Research.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Imbens, G. W. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Jacob, R. T., Goddard, R. D., & Kim, E. S. (2014). Assessment of the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis*, *36*(1), 44–66.

Jacob, R., Somers, M.-A., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. *Evaluation Review*, *40*(3), 167–198. https://doi.org/10.1177/0193841X16663414

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

Lecher, M. (2010). The estimation of causal effects by difference-in-difference models. University of St. Gallen Discussion Paper no. 2010–28.

Miller, L. C., & Mittleman, J. (2012). *High Schools That Work* and college preparedness: Measuring the model's impact on mathematics and science pipeline progression. *Economics of Education Review*, *31*(6), 1116–1135. https://doi.org/10.1016/j.econed urev.2012.07.014.

Neyman, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, *2*, 107–108.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

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 randomised and observational evaluations. *BMJ*, *328*, 560.

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

Somers, M., 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. *MDRC* working paper in research methodology. New York, NY.

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*(3), 1–17.

St. Clair, T., Hallberg, K., & Cook, T. D. (2016). The validity and precision of the comparative interrupted time series design: Three within-study comparisons. *Journal of Educational and Behavioral Statistics*, *41*(3), 269–299.

Steiner, P.M., Cook, T.D., & Shadish, W.R. (2011). On the importance reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2), 213-236.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, *36*(4): 187–198.

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups and adjusting for group differences. *Journal of Educational and Behavioral Statistics*, *33*(3): 279–306.

U.S. Department of Education. (2013). *Investing in Innovation Fund (I3) awards* [Webpage]. Retrieved from http://www2.ed.gov/programs/innovation/awards.html.

What Works Clearinghouse. (2018). *Procedures and standards handbook* (Version 4.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern matching. *Journal of Research on Educational Effectiveness*, *8*(2), 245–279.

Wong, V. C., Wing, C., Steiner, P. M., Wong, M., & Cook, T. D. (2012). Research designs for program evaluation. In W. Velicer

& J. Schinka (Eds.), *Handbook of psychology: research methods in psychology* (2nd ed.). Hoboken, NJ: Wiley and Sons.

AUTHORS

KELLY HALLBERG, PhD, is the scientific director for the University of Chicago Urban Labs, 33 N LaSalle, Suite 1600, Chicago, IL 60602; *khallberg@uchicago.edu*. She oversees a portfolio of applied research projects examining innovative approaches to improving the academic and life outcomes of low-income, urban youth and specializes in methodological issues in evaluating the impacts of interventions.

RYAN WILLIAMS, PhD, is a principal researcher at the American Institutes for Research, 10 S. Riverside Plaza, Chicago, Illinois 60606; *rwilliams@air.org*. His research focuses on methods for improving experimental and quasi-experimental designs in education and on meta-analytic methods and applications.

ANDREW SWANLUND, PhD, is a principal researcher at the American Institutes for Research, 10 S. Riverside Plaza, Suite 600, Chicago, IL 60606; *aswanlund@air.org*. His research focuses on statistical methods for quasi-experimental designs, psychometrics/measurement, and the efficacy of education policies and interventions.

JARED ENO, MPP, is a graduate student in sociology and public policy at the University of Michigan, 500 S. State St., Room 3115, Ann Arbor MI 48103; *jpeno@umich.edu*. His research focuses on the history of U.S. higher education.

Manuscript received October 4, 2016
Revisions received September 26, 2017, and January 5, 2018
Accepted February 7, 2018

Appendix A

Below is sample R code for running the models discussed in this paper (baseline mean, linear trend, and school and year fixed effects). A simulated dataset (with 80 schools, 10 time points, and a simulated treatment effect of .25 standard deviations) is provided as an online supplement, available in the online journal.

```
load("simulated_cits_data.rdata")
library(lme4)
################################################
############################
#CREATE POST VARIABLE
################################################
############################
data$post1 <- "pre"
data$post1[data$time == 6] <- "post year 1"
data$post1[data$time == 7] <- "post year 2"
data$post1[data$time == 8] <- "post year 3"
data$post1[data$time == 9] <- "post year 4"
data$post1[data$time == 10] <- "post year 5"
data$post1 <- relevel(factor(data$post1), ref = "pre")
################################################
############################
#CREATE A SEPARATE TREATMENT INDICATOR (AND
TIME INDICATORS FOR THE SCHOOL/YEAR
#FE MODEL
################################################
############################
data$tx1 <- 0
data$tx1[data$tx == "comparison"] <- 0
data$tx1[data$tx == "treatment"] <- 1

data$post2 <- 0
data$post2[data$post == "post"] <- 1

data$posty0 <- 0
data$posty0[data$post1 == "pre"] <- 1
data$posty1 <- 0
data$posty1[data$post1 == "post year 1"] <- 1
data$posty2 <- 0
data$posty2[data$post1 == "post year 2"] <- 1
data$posty3 <- 0
data$posty3[data$post1 == "post year 3"] <- 1
data$posty4 <- 0
data$posty4[data$post1 == "post year 4"] <- 1
data$posty5 <- 0
data$posty5[data$post1 == "post year 5"] <- 1
################################################
############################
#MODEL OUTCOMES (OVERALL PRE-POST EFFECTS)
################################################
############################

#BASELINE MEAN MODEL
fit_bm_1 <- lmer(y1 ~ tx * post + (1 | clust), data = data)
summary(fit_bm_1)

#BASELINE TREND MODEL
fit_bt_1 <- lmer(y1 ~ tx * post + tx * time + (1 | clust), data =
data)
summary(fit_bt_1)

#SCHOOL AND YEAR FIXED EFFECTS MODEL
fit_syfe_1 <- lm(y1 ~ tx1:post2 + factor(time) + factor(clust),
data = data)
summary(fit_syfe_1)
################################################
############################
#MODEL OUTCOMES (SEPARATE EFFECTS FOR EACH
POST YEAR)
################################################
############################

#BASELINE MEAN MODEL
fit_bm_2 <- lmer(y1 ~ tx * post1 + (1 | clust), data = data)
summary(fit_bm_2)

#BASELINE TREND MODEL
fit_bt_2 <- lmer(y1 ~ tx * post1 + tx * time + (1 | clust), data =
data)
summary(fit_bt_2)

#SCHOOL AND YEAR FIXED EFFECTS MODEL
fit_syfe_2 <- lm(y1 ~ tx1:posty1 + tx1:posty2 + tx1:posty3 +
tx1:posty4 + tx1:posty5 + factor(time) + factor(clust), data =
data)
summary(fit_syfe_2)
```