



## The Internal Validity of the School-Level Comparative Interrupted Time Series Design: Evidence From Four New Within-Study Comparisons

Sam Sims, Jake Anders & Laura Zieger

To cite this article: Sam Sims, Jake Anders & Laura Zieger (2022): The Internal Validity of the School-Level Comparative Interrupted Time Series Design: Evidence From Four New Within-Study Comparisons, Journal of Research on Educational Effectiveness, DOI: [10.1080/19345747.2022.2051652](https://doi.org/10.1080/19345747.2022.2051652)

To link to this article: <https://doi.org/10.1080/19345747.2022.2051652>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC



Published online: 15 Apr 2022.



Submit your article to this journal [↗](#)



Article views: 597



View related articles [↗](#)



View Crossmark data [↗](#)

# The Internal Validity of the School-Level Comparative Interrupted Time Series Design: Evidence From Four New Within-Study Comparisons

Sam Sims<sup>a</sup> , Jake Anders<sup>a</sup>  and Laura Zieger<sup>a</sup>

Centre for Education Policy and Equalising Opportunities, UCL, London, UK

## ABSTRACT

Comparative interrupted time series (CITS) designs evaluate impact by modeling the relative deviation from trends among a treatment and comparison group after an intervention. The broad applicability of the design means it is widely used in education research. Like all non-experimental evaluation methods however, the internal validity of a given CITS evaluation depends on assumptions that cannot be directly verified. We provide an empirical test of the internal validity of CITS by conducting four within-study comparisons of school-level interventions previously evaluated using randomized controlled trials. Our estimate of bias across these four studies is 0.03 school-level (or 0.01 pupil-level) standard deviations. The results suggest well-conducted CITS evaluations of similar school-level education interventions are likely to display limited bias.

## ARTICLE HISTORY



Received 2 November 2020  
Revised 15 December 2021  
Accepted 26 February 2022

## KEYWORDS

Comparative interrupted time series; randomised controlled trials; within-study comparisons; internal validity

## Introduction

In education, many interventions of interest to school leaders and policymakers are implemented at the school-level, without random allocation. Consequently, researchers are required to adopt a non-experimental evaluation method. Comparative interrupted time series (CITS) (Cook & Campbell, 1979) methods constitute one such design, premised on projecting outcomes from the pre- into the post-treatment period. Perhaps the best-known variant of the CITS is the conventional two-period difference-in-difference (Card & Krueger, 1993), in which the change in the outcome in the comparison group is used to project the pretreatment level of the outcome in the treatment group into the post-treatment period. Under the parallel trends assumption, this acts as a proxy for the untreated outcomes among the treated units. However, the family of CITS methods also includes variants that focus on quantifying the relative deviation from trends, rather than levels, of the outcome. These alternative approaches are valuable in settings where selection into treatment occurs based on pretreatment values of the outcome variable, rendering the common trends assumption implausible (St. Clair & Cook, 2015). The

**CONTACT** Sam Sims  [s.sims@ucl.ac.uk](mailto:s.sims@ucl.ac.uk)  Centre for Education Policy and Equalising Opportunities, UCL, Gower St., Bloomsbury, London, UK.

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

deviation from pretreatment trends in the comparison group is then used to adjust the projection of the pretreatment trend in the treatment group into the post-treatment period, providing the necessary proxy for untreated outcomes among the treated units.

The various CITS specifications were first deployed to evaluate a school improvement programme (Bloom et al., 2001). Since then, CITS have been used in evaluations of No Child Left Behind (Dee et al., 2013; Grissom et al., 2014; Lee & Reeves, 2012; Wong et al., 2015), curricular reforms (Bell et al., 2016; Jacob et al., 2017), school wraparound programmes (Gandhi et al., 2018), and school turnaround programmes (Henry et al., 2020; Strunk et al., 2016). Indeed, the approach is in many ways well-suited to education settings. Conducting a CITS evaluation requires a well-measured outcome variable recorded consistently across multiple periods. In many countries, high-stakes school examination results are recorded and made publicly available for each academic year, meaning the data requirements for CITS are often easy to fulfill for education researchers. Moreover, CITS' emphasis on modeling and then projecting the outcome variable, rather than focusing on adjusting for or matching on covariates, means that the method can often be implemented without access to sensitive individual-level data on pupil characteristics (Jacob et al., 2014). In addition, the key identifying assumption of CITS—that unobserved time-varying confounders are shared between the treatment and comparison groups—are often more plausible in education, where schools are nested within common regional and national administrative units.

Nevertheless, any violation of the CITS identifying assumptions could undermine internal validity, yielding biased impact estimates. Since the identifying assumption relates to a counterfactual, it cannot be directly empirically tested in any given evaluation. However, an alternative method for assessing bias is available in the form of within-study comparisons (WSC). These involve comparing the impact estimates from a benchmark study—often a randomized controlled trial (RCT)—with those of an alternative evaluation of the same intervention using the same outcomes, but employing a separate, non-experimental comparison group (LaLonde, 1986; Wong et al., 2018). Thus, between-study comparisons quantify differences in effect size estimates for a given intervention. RCTs in education tend to provide noisy estimates of the true impact (Lortie-Forgues & Inglis, 2019; Spybrook et al., 2016) which means that comparing the effect sizes from a single within-study comparison provides a noisy estimate of bias. However, the accumulation of multiple WSCs for a given non-experimental design does allow bias to be estimated across WSCs (Glazerman et al., 2003).

### ***Within-Study Comparisons of the CITS Design***

Wong et al. (2018) identified twenty WSCs of interrupted time series (ITS) or CITS designs with experimental benchmarks across environmental policy (Ferraro & Miranda, 2014); epidemiology (Fretheim et al., 2013); health (Schneeweiss et al., 2004); welfare-to-work (Bloom et al., 2002) and immigration (McKenzie et al., 2010). Coopersmith et al. (2022) provide the most recent meta-analytic summary, identifying 12 unique CITS WSCs. The average deviation of the CITS estimates from the benchmark estimates was 0.03 SD. At the level of individual WSCs, all but one of the twelve deviated from the benchmark estimate by less than 0.1 SD.

Within education, we identified six WSCs that contrasted CITS impact estimates with those from RCTs. St. Clair et al. (2016) report WSCs from three whole-school interventions: two in Indiana and one in Florida. In the two trials in which the pre-trends were stable and thus easier to model, the non-experimental estimates came within 0.02 school-level SD of the experimental benchmark. In the third trial, where the pre-trends were highly unstable and thus harder to model, the non-experimental estimates deviated from the experimental estimates substantially. This highlights the importance of the stability of pre-trends for the validity of the CITS design (Kim & Steiner, 2016). Hallberg et al. (2020) report WSCs from a further three whole-school interventions—a college readiness programme, an early literacy programme, and an online mathematics programme—and find an average absolute bias of between 0.05 and 0.08 school-level SD in the immediate post-intervention period. However, the bias was found to be strongly influenced by the specific modeling approach adopted and the authors recommend deciding which of the specific CITS specifications to adopt based on careful inspection of the pre-trends.

### ***Aims and Scope***

The current paper aims to substantially increase the number of such WSCs in the education literature by conducting four additional WSCs of CITS. As previously discussed, contrasts of impact estimates from single WSCs provide limited information about bias because the typical RCT impact estimate in education is noisy. Accumulating four new WSCs is therefore particularly valuable in that it allows estimates of bias across studies, netting out the noise from the individual RCTs. By synthesizing the results across four new WSCs, we aim to contribute to the empirical evidence on the validity of a widely used evaluation design in the education literature. The next section describes the set of interventions that we reevaluate and the school-level data that we employ in this analysis. The ‘Methods’ section then describes our approach to conducting each of the CITS evaluations and the techniques by which these are subsequently synthesized. The ‘Results’ section presents our findings for each of the individual WSCs, followed by the within-study comparisons, and the ‘Discussion’ section concludes by reflecting on the limitations and implications of these findings.

### **Sample**

#### ***Interventions***

We set out to reevaluate interventions previously tested using RCTs by the Education Endowment Foundation (EEF). The EEF is a large research funder in England, which typically commissions RCTs of interventions judged to have previously shown evidence of promise. EEF was established in 2010 and has since commissioned over 100 RCTs, the results of which are routinely published on their website (Dawson et al., 2018). These interventions can be reevaluated using school-level CITS if five conditions hold: (1) the intervention was originally randomized at school-level, (2) the trial used all pupils in the treated cohorts’ results in standardized national tests at age 11 or age 16 as one of the outcome measures, (3) the trial is complete and results have been reported,

(4) the time series necessary to conduct the CITS does not include discontinuities in how the outcome variables were recorded, and (5) it is possible to calculate effect sizes from the original RCT that are comparable to those generated by our CITS.

The EEF supplied us with a list of all of their trials/interventions that met criteria 1–3 as of December 2019, which amounted to 19 in total.<sup>1</sup> Among these, eight did not meet criterion 4) due to the relevant section of time series straddling major reforms of either age-11 examinations (implemented in 2015/16) or age 16 examinations (implemented in 2016/17).<sup>2</sup> Similarly, three interventions had to be dropped because the relevant time series included a year in which we had missing data due to e.g., a the boycott of the age-11 examinations.<sup>3</sup> A further three interventions had to be dropped because the original evaluation report did not include the information necessary to calculate an effect size measure that would be comparable with our school-level CITS evaluations.<sup>4</sup> Finally, one intervention had to be dropped because of discrepancies in the number of observations in the EEF data archive and the published evaluation report, which could not be resolved by the EEF's data scientists.

This left four interventions: two in primary schools and two in secondary schools (see Table 1). The interventions are diverse in nature, ranging from teacher professional development (PD) to the use of educational technology to adapt instruction. All four trials measured outcomes for all pupils in the relevant cohorts in participating schools and conducted analysis on an intention-to-treat basis. Attrition at both the school and pupil level was below the What Works Clearinghouse conservative standards across all four trials (What Works Clearinghouse, 2017). The estimated effect sizes reported in the original evaluation reports range from 0.00–0.26 school-level standard deviations (SD) or 0.00–0.09 pupil-level SD. None of these estimates were statistically significant at conventional levels, which is perhaps unsurprising given the lack of statistical power in many educational trials (Lortie-Forgues & Inglis, 2019; Spybrook et al., 2016). In line with EEF analysis guidelines, all four trials used pupil prior achievement as the only covariate (besides strata and cluster variables) in their main analytical models.

## Data

We utilize school-level data derived from the UK Department for Education's National Pupil Database, an administrative census containing information on all publicly funded schools in England since 1991/92.<sup>5</sup> Each school in the data has a unique identifier, which allows them to be linked across years. However, a large proportion of schools have either converted to academy status or undergone other changes that cause their identifier to change between years. Accordingly, we utilize an additional database of

---

<sup>1</sup>EEF originally sent us 21 interventions but two of these (Affordable Maths Tuition and Engage in Education) did not use school-level outcomes.

<sup>2</sup>Changing Mindsets (regrant), Grammar for Writing (regrant), Lesson Study, Philosophy for Children, Research Learning Communities, Scratch Maths, Affordable Tuition (regrant) and Embedding Formative Assessment.

<sup>3</sup>Philosophy for Children; Thinking, Doing, Talking Science; Chess in Primary Schools.

<sup>4</sup>Children's University, Pupil Motivation Financial, Pupil Motivation Non-financial. The first of these did not report any effect size measure. The latter two did report an effect size but the mean difference was divided by the population-level standard deviation, which not included in the report.

<sup>5</sup>Publicly accessible school-level data can be found here: <https://www.compare-school-performance.service.gov.uk/download-data>

**Table 1.** The interventions.

Reference	Phase	RCT results in pupil-level SD	RCT results in school-level SD	Sample size	Attrition	Brief description
Flipped learning (FLIP) Rudd et al. (2017)	Prim.	ES = 0.09 SE = 0.19 <i>p</i> = 0.62	ES = 0.26 SE = 0.55 <i>p</i> = 0.62	Schools: 24 Pupils: 1,214	Schools: 0% Pupils: 7%	Pupils study new math at home, then consolidate in class time
Learner response (LERS) <sup>a</sup> Wiggins et al. (2017)	Prim.	ES = 0.00 <sup>b</sup> SE = 0.07 <i>p</i> = 0.96	ES = 0.00 SE = 0.84 <i>p</i> = 0.96	Schools: 97 Pupils: 3,213	Schools: 3.1% Pupils: 11.7%	Immediate in-class pupil feedback from teachers via electronic devices
The RISE project (RISE) <sup>a</sup> Wiggins et al. (2019)	Sec.	ES = 0.09 SE = 0.06 <i>p</i> = 0.15	ES = 0.19 SE = 0.13 <i>p</i> = 0.15	Schools: 40 Pupils: 7,468	Schools: 2.5% Pupils: 8.8%	Teacher “research leads” supporting teachers’ use of research
Teacher observation (TEOB) <sup>a</sup> Worth et al. (2017)	Sec.	ES = 0.06 SE = 0.04 <i>p</i> = 0.13	ES = 0.15 SE = 0.10 <i>p</i> = 0.13	Schools: 82 Pupils: 7,366	Schools: 0% Pupils: 1.4%	Teacher PD based on programme of structured peer lesson observations

Prim: primary; Sec: secondary; PD: professional development; ES: effect size derived from intention-to-treat analysis based on all pupils in the relevant cohorts within the schools. SE: standard error; All standard errors are derived by dividing the widths of the 95% confidence intervals reported in the EEF evaluation reports by 3.92. Sample size is the number of schools/pupils initially randomized to generate the effect size in the “RCT results” column.

<sup>a</sup>For these trials we can only use the first treatment cohort because the relevant time series for the second treatment cohort straddles a discontinuity in the time series.

<sup>b</sup>The results was reported as “0.00” in the evaluation report, rather than to a given number of significant figures.

**Table 2.** Characteristics of the RCT treatment group, RCT control group, and CITS comparison group averaged across the four pretreatment years.

Intervention	Covariate	RCT treat	RCT cont	CITS comp	Test of difference on RCT cont and CITS comp
FLIP	# Schools	12	12	4169	–
	%FSM	36.1	40.5	23.9	**
	%EAL	31.2	41.7	9.8	**
	%SEN	9.9	9.5	10.2	–
	Prior ach.	14.7	14.4	15.3	**
LERS	# Schools	48	45	7228	–
	%FSM	58.3	59.2	28.9	**
	%EAL	9.8	18.8	17.7	–
	%SEN	16.2	16.4	10.6	**
	Prior ach.	13.7	13.7	15.2	**
RISE	# Schools	20	19	2269	–
	%FSM	23.6	27.6	27.9	–
	%EAL	4.9	15.9	14.4	–
	%SEN	6.9	6.1	7.0	–
	%PLO	14.4	16.0	15.7	–
TEOB	Prior ach.	27.6	27.3	27.4	–
	# Schools	41	40	2877	–
	%FSM	33.3	27.5	23.5	*
	%EAL	16.3	11.6	12.8	–
	%SEN	9.7	9.0	8.5	–
	%PLO	19.3	19.0	16.0	*
	Prior ach.	26.6	26.7	27.3	*

See Table 1 for intervention acronyms. FSM: free school meals; EAL: English as an additional language; SEN: special educational needs; Prior Ach: prior achievement; For the primary school interventions (FLIP and LERS) prior achievement is the average Key Stage 1 (age 7) SATs score of the cohort. For the secondary school interventions (RISE and TEOB) prior achievement is the average Key Stage 2 (age 11) SATs score. PLO: pupils with low prior attainment at end of KS2; Variables in brackets: information not available for all years (not included in CITS modeling and conditional plots).

\*\* $p < 0.01$ . \* $p < 0.05$ .

consistent school identifiers to ensure that we maintain the integrity of the time series when linking schools across waves. Since these school-level data are technically averages of repeated cross-sections of pupils, it is important that we control for compositional changes (Hallberg et al., 2018; Pohl et al., 2009). We do this using the following covariates: the percentage of pupils that are eligible for free school meals (FSM), the percentage of pupils with English as an additional language (EAL), percentage of pupils with special educational needs (SEN), school size (number of pupils) and school type.<sup>6</sup> Summary statistics are available in Table 2.

For both of our primary school interventions, the original RCTs used as their outcome measure pupils' scores in SATs (Statutory Attainment Tests), which are standardized, high-stakes tests taken by all pupils in state schools in England at age 11. For both of our secondary interventions, the original RCTs used as their outcome measure pupils' scores in their math GCSE (General Certificate of Secondary Education), which are standardized, high-stakes tests taken by all pupils in state schools in England at age 16. In line with this, we use school average math SATs scores (primary) and school average math GCSE score (secondary) as our outcome measure in the within-study comparisons. For each of our four interventions, we then create a time-series of the relevant outcome measure, spanning the four years prior to the intervention being

<sup>6</sup>State-funded schools in England operate under a range of legal frameworks (see Eyles & Machin, 2019). We code the school type variables as having five categories: academy (including free schools and city technology colleges), community, foundation, voluntary aided or voluntary controlled.

introduced ( $t-4$ ,  $t-3$ ,  $t-2$  and  $t-1$ ) and a single period after the intervention ( $t + 1$ ), which is always the same year in which the original RCT estimated the impact.<sup>7</sup> We also utilize prior achievement measures: end of Key Stage 1 (age 7) SATs math scores for the primary school interventions, and end of Key Stage 2 (age 11) math SATs scores for secondary.

## Methods

### Models for the within-Study Comparisons

The conventional two-period difference-in-difference model (also referred to as the “baseline mean” model) can be written as:

$$Maths_{sy} = \beta_0 + \beta_1 Treat_s + \beta_2 Post_y + \beta_3 Treat_s * Post_y + \beta_4 X_{sy} + u_s + \epsilon_{sy}$$

Where  $Maths_{sy}$  is the average math score in school  $s$  in year  $y$ ;  $Treat_s$  is a binary variable indicating whether school  $s$  is in the treatment group or not;  $Post_y$  is a binary variable capturing whether the treatment is “on” or “off” in year  $y$  (our models only include one post-treatment period);  $X_{sy}$  is a vector of school-level covariates (FSM, EAL, SEN and cohort prior achievement);  $u_s$  is a school-specific random error term and  $\epsilon_{sy}$  is a year-specific random error term.  $\beta_1$  captures the difference in the level of the outcome variables between the treatment and comparison groups in the pretreatment period and  $\beta_2$  captures the level change between the pre- and post-treatment periods in the comparison group.  $\beta_3$  is the coefficient of interest since it captures the level change between the pre- and post-treatment period in the treatment group, over and above that in the comparison group. The key threats to the validity of the baseline mean model in our setting are changes in the composition of students attending the schools from one year to the next in a way that is related to the outcome, or a lack of common trends in treatment and comparison schools (Cunningham, 2020).

In the absence of common trends, the baseline-linear trends model (Bloom, 2003) can be adopted instead:

$$Maths_{sy} = \beta_0 + \beta_1 Year_y + \beta_2 Treat_s + \beta_3 Year_y * Treat_s + \beta_4 Post_y + \beta_5 Post_y * Treat_s + \beta_6 X_{sy} + u_s + \epsilon_{sy}$$

Where  $Year_y$  is a variable capturing the periods  $t-4$ , ...,  $t + 1$  and all other variables are defined as in the previous model.  $\beta_0$  and  $\beta_2$  now capture the intercepts for the comparison and treatment groups, respectively.  $\beta_1$  now captures the gradient of the comparison group time series in the pretreatment period,  $\beta_1 + \beta_3$  does the same for the treatment group, and  $\beta_4$  captures the deviation from trend in the post treatment period for the comparison group.  $\beta_5$  is now the coefficient of interest since it captures the deviation from trend in the treatment group, over and above that in the comparison group. The key threats to the validity of the baseline linear-trends model in our setting are compositional changes, time-varying confounders that cause the treatment group to deviate from trend in the post-treatment period differently to those in the comparison group, or misspecification of the model capturing the pretreatment trends such that

<sup>7</sup>We utilise just one period of post-treatment data in order to minimise the number of time series straddling a change in the way that outcomes are recorded and because previous empirical work suggest that adding more post-treatment periods can increase bias (Hallberg et al., 2020).



they would be incorrectly projected into the post-treatment period (Hallberg et al., 2018; St. Clair & Cook, 2015).

Misspecification of the model due to non-linear pretreatment trends showing a single turning point (quadratic functional form) can be accommodated through the addition of quadratic year terms in the baseline nonlinear-trends model:

$$\begin{aligned} Maths_{sy} = & \beta_0 + \beta_1 Year_y + \beta_2 Year_y^2 + \beta_3 Treat_s + \beta_4 Year_s * Treat_s + \beta_5 Year_y^2 * Treat_s + \beta_6 Post_y \\ & + \beta_7 Post_y * Treat_s + \beta_8 X_{sy} + u_s + \epsilon_{sy} \end{aligned}$$

In this model,  $\beta_1$  and  $\beta_2$  capture the linear and quadratic components of the pretreatment trends in the comparison group and  $\beta_4$  and  $\beta_5$  captures the linear and quadratic components of the way in which the pretreatment trend in the treatment group deviates from that of the comparison group.  $\beta_7$  is the coefficient of interest, since it captures the deviation from trend in the treatment group, over and above that in the comparison group.

The choice between the various CITS specifications should be based on the functional form of the pretreatment trends (St. Clair et al., 2016) and it is therefore common practice for CITS papers to begin by graphing the time series. Technically, however, the relevant consideration is the pretreatment trends conditional on observables. We therefore use the `binsreg` command (Cattaneo et al., 2019) to plot the mean values of the outcome variables for each year and each intervention, conditional on the covariates. These can be found in [Appendix Figure A1](#). Based on visual inspection of these figures, we categorized each of the interventions as showing either: parallel trends, in which case we analyzed it using the baseline mean model; non-parallel linear trends, in which case we analyzed it using the linear baseline-trends model; non-parallel trends displaying one turning point, in which case we analyzed it using the non-linear baseline-trend model (Hallberg et al., 2020).

A final methodological decision relates to which schools to include in our comparison time series. In line with guidance based on previous within-study comparisons, we chose to use only geographically local schools, on the grounds that they are more likely be similar in terms of unobservable characteristics (Bifulco, 2012; Cook et al., 2008; Hallberg et al., 2016; Wong et al., 2017). More specifically, for each intervention we include in our comparison group all state-funded schools in the same phase (primary/secondary) in regions of England that contain at least one treatment school.<sup>8</sup> Comparisons of the characteristics of schools in the RCT treatment, RCT control and CITS comparisons groups can be found in [Table 2](#). Comparisons of the outcomes across the same three groups, after conditioning on covariates, can be found in [Figure A1](#) in the appendix. [Appendix Table A3](#) provides a summary of the design of all four within-study comparisons.

### **Estimating Overall Bias**

Once we have used the relevant models to estimate impact relative to this comparison group, we are in a position to compare the CITS and RCT results. Within-study

---

<sup>8</sup>There are nine regions in England, each containing between three million and ten million people.

comparisons generally compare both the impact estimates (in effect size units) and the precisions or  $p$  values from the two evaluations. The latter often involves comparing the  $p$  values of the benchmark (experimental) impact estimate and the  $p$  values of the comparison (non-experimental) study, which yields a simple binary metric for whether the results of the two studies agree. We avoid this approach for two reasons. First,  $p$  values have very different interpretations in experimental and observational research, calling into questions their comparability (Berk, 2004; Rosenbaum, 2017). Second, the difference between one statistically significant impact estimate and another not statistically significant impact estimate may not itself be statistically significant (Gelman & Stern, 2006).

Instead, we focus exclusively on comparisons of the impact estimates expressed as the standardized mean difference effect sizes, calculated by subtracting the mean math outcome in the untreated schools ( $T=0$ ) from that in the treated schools ( $T=1$ ) and dividing by the SD across both  $T=0$  and  $T=1$  schools pooled together. This is also known as the *standardized effect difference* (Steiner & Wong, 2018). The school-level data that we employ for our CITS evaluations leads us to employ the school (s) level SD of the outcome measures in the denominator. Where effect sizes in the original RCT reports were calculated using the pupil level SD, we recalculate these using the school-level SD in order to ensure comparability (see Appendix Table A1 for details). Using school-level SD is also consistent with existing CITS WSC, which makes our findings comparable with the existing literature (Hallberg et al., 2020; St. Clair et al., 2016). However, because most analysts are used to pupil-level effect sizes, which are much smaller, we also report these wherever possible (Kraft, 2020).

$$Effect\ Size = \frac{(\overline{Maths}_{s, T=1} - \overline{Maths}_{s, T=0})}{SD_{s, pooled}}$$

### Summary Measures of Bias

Well-implemented RCTs provide an unbiased estimator for the causal impacts of interventions. In an individual RCT with finite sample size, however, there will be imprecision due to sampling error or so-called “realized confounding” (Deaton & Cartwright, 2018). Indeed, in education, RCT impact estimates tend to be very imprecise—the mean width of 95% confidence intervals in EEF trials, for example, is 0.34 pupil-level SD (Lortie-Forgues & Inglis, 2019). For a given intervention  $i$ , differences in treatment effect estimates between an RCT ( $\widehat{ATT}_i^{RCT}$ ) and a CITS evaluation ( $\widehat{ATT}_i^{CITS}$ ) can therefore reflect either selection bias from the CITS ( $\theta_i^{CITS}$ ) or random sampling error from the RCT ( $e_i^{RCT}$ ) or the CITS ( $e_i^{CITS}$ ), which have mean 0 and variance  $v$  :

$$\begin{aligned} \widehat{ATT}_i^{RCT} - \widehat{ATT}_i^{CITS} &= \theta_i^{CITS} + e_i^{RCT} + e_i^{CITS} \\ e_i^{RCT} &\sim N(0, v_i^{RCT}) \\ e_i^{CITS} &\sim N(0, v_i^{CITS}) \end{aligned}$$

Given that the  $e_i$  are unknown, we need some method for isolating  $\theta_i^{CITS}$ . A simple way to address this is to take the mean of  $\widehat{ATT}_i^{RCT} - \widehat{ATT}_i^{CITS}$  across interventions  $i$ .

Since  $e_i^{RCT}$  and  $e_i^{CITS}$  are zero in expectation, across many WSC we would expect these terms to fall out of the equation above. An important corollary of this is that we should not consider  $\widehat{ATT}_i^{RCT} - \widehat{ATT}_i^{CITS}$  for a given intervention to provide an unbiased estimate of  $\theta_i^{CITS}$  for that intervention. For a given intervention, we have no way of knowing how much of  $\widehat{ATT}_i^{RCT} - \widehat{ATT}_i^{CITS}$  is composed of bias  $\theta_i^{CITS}$  and how much is composed of noise. We therefore only interpret the mean of  $\widehat{ATT}_i^{RCT} - \widehat{ATT}_i^{CITS}$  across our four interventions as an estimate of bias.

A more sophisticated approach to combining estimates would use random effects meta-analysis in order to account for both the noise from RCT estimate within each WSC and the variation in selection bias across the WSCs (Chaplin et al., 2018; Weidmann & Miratrix, 2021). However, to our knowledge, there is no established method for ascertaining the variance surrounding the *difference* in effect size estimates without access to micro-data. Our school-level data does not permit this. However, in Appendix Table A4, we re-estimate the impact of the original RCTs using our school-level data, bootstrap the standard errors for the effect size difference (Steiner & Wong, 2018), and then meta-analyse the results. This approach ignores the units (pupils) within each cluster (school) and the resulting standard errors should therefore be considered as an upper bound on the true standard errors. However, this allows us to provide an approximate precision-weighted meta-analytic estimate of the bias across studies, as well as providing information that may be useful for subsequent meta-analysts.

## Results

Table 3 shows the results of the CITS regression models for each of the four interventions, with all coefficients expressed in terms of school-level SD. The coefficient of

**Table 3.** Comparative interrupted time series regressions.

	FLIP	LERS	RISE	TEOB
Treat x post	0.13 (0.51)	-0.03 (0.10)	0.16* (0.08)	0.22 (0.15)
Treat	-0.05 (0.50)	0.06 (0.09)	-0.05 (0.09)	-0.35* (0.15)
Post	0.09** (0.03)	0.15** (0.01)	-0.07** (0.01)	0.16** (0.02)
Year	0.70** (0.02)	-	-	0.43** (0.02)
Year x treat	0.13 (0.44)	-	-	0.24 (0.13)
Year <sup>2</sup>	-0.09** (0.01)	-	-	-0.08** (0.01)
Year <sup>2</sup> × treat	-0.01 (0.09)	-	-	-0.04 (0.02)
Marginal $R^2$	0.442	0.338	0.707	0.738
Conditional $R^2$	0.660	0.607	0.881	0.892
$N$	4,067	7,117	2,124	2,694
Parallel trends	-	✓	✓	-
Model type	Baseline non-linear trend	Baseline mean	Baseline mean	Baseline non-linear trend

Each column is a separate OLS regression. Coefficients and standard errors are expressed in school-level standard deviations. Numbers in parentheses are standard errors.

\*\* $p < 0.01$ . \* $p < 0.05$ .

**Table 4.** Within-study comparison results.

Specification		ES and (SE) in school-level SD			ES in pupil-level SD* RCT-CITS
		RCT	CITS	RCT-CITS	
FLIP	Non-parallel trends, one turning point: Baseline non-linear trends	0.26 (0.55)	0.13 (0.51)	+0.13	+0.04
LEERS	Parallel trends: Baseline mean model	0.00 (0.84)	-0.03 (0.10)	+0.03	0.00
RISE	Parallel trends: Baseline mean model	0.19 (0.13)	0.16 (0.08)	+0.03	+0.01
TEOB	Non-parallel trends, one turning point: Baseline non-linear trends	0.15 (0.10)	0.22 (0.15)	-0.07	-0.03
-	-	Mean difference:		0.03	0.01
-	-	Mean absolute difference:		0.07	0.02

See Table 1 for intervention acronyms. ES: effect size; RCT: randomized controlled trial; CITS: comparative interrupted time series; RCT-CITS: RCT effect size minus the CITS effect size; To calculate the RCT-CITS effect size difference in terms of pupil-level SD we carry out a conversion using the pupil-and school-level SD appropriate to each trial, as reported in Appendix Table A1. We multiply the estimated difference by the school-level SD and divide by the pupil-level SD. This is a direct reversal of the process used to convert the RCT-estimated effect size into one expressed in terms of the school-level SD.

interest can be found in bold in the first row (Treat × Post), which shows the change in the outcome in the treatment group in the post-treatment period, relative to that in the comparison group. The LEERS and RISE models use the baseline linear trends specification. The FLIP and TEOB models use the non-linear baseline trends model, which also include quadratic terms for year interacted with treatment status, which allows for divergent non-linear trends in the treatment and comparison groups. The coefficients of interest vary from -0.03 to + 0.22 SD. It should be kept in mind that these coefficients are expressed in terms of school-level SD, which are typically 1.5–3 times larger than those expressed using pupil-level SD (Kraft, 2020). For transparency, we also report the results for each intervention using all three CITS specifications in Table A2 in the Appendix.

Our primary interest is in comparing the impact estimates from Table 3 with those from prior RCT evaluations of the same interventions. Table 4 shows the results of these four within-study comparisons. The effect size columns report the impact estimates from the original RCT and new CITS evaluations and the final two columns shows the RCT effect size minus the CITS effect size. The differences in effect sizes in the four WSCs vary from 0.03 to 0.13 school-level SD (0–0.04 pupil-level SD) in absolute magnitude. The mean absolute difference is 0.07 school-level SD (0.02 pupil-level SD).

The two interventions that showed parallel trends in the pretreatment period (LEERS and RISE) showed the smallest absolute differences between the RCT and CITS impact estimates: 0.03 school-level SD or ≤ 0.01 pupil-level SD. By contrast, the interventions showing non-linear, non-parallel trends (TEOB and FLIP) showed slightly larger absolute differences: 0.07–0.13 school-level SD or 0.03–0.04 pupil-level SD. While all of these results rely on estimates from single within-study comparisons, it is notable that the more complicated the pretreatment trends are, the further the CITS impact estimate tends to be from the RCT impact estimate. Taking the mean of the signed (rather than absolute) differences helps net out the (mean zero) noise across the RCT impact

estimates and gives us our key result on the bias across the four CITS studies in our setting: 0.03 school-level SD or 0.01 pupil-level SD.

As discussed in the methods section, we cannot calculate correct standard errors for our effect size differences (RCT-CITS) using our school-level data. However, in Appendix Table A4, we calculate upper-bound estimates, which allows us to conduct a random-effects meta-analysis across the four effect size differences. Using this method, we estimate mean bias of 0.11 school-level SD, with an upper-bound standard error of 0.1 school-level SD (Appendix Figure A2). Scaling this down by the ratio of our school-level and pupil-level SD estimates in Table 4, this is equivalent to 0.04 pupil-level SD, with a standard error of 0.03 pupil-level SD. This precision-weighted average is slightly larger than our estimate in Table 4 (0.01 pupil-level SD) but is not statistically distinguishable from zero ( $p = 0.25$ ).

## Discussion

The CITS is in an increasingly popular approach for studying the impact of education policies, in part because it is well suited to the institutional characteristics and data available in education settings. Yet like all non-experimental methods, the assumptions on which the CITS depend cannot be directly verified. Within-study comparisons provide an alternative approach to evaluating the internal validity of CITS design, by comparison with an experimental benchmark. Using a database of experimental evaluations in England, we have contributed four new within-study comparisons of the school-level CITS design. By summarizing the effect size differences across the WSCs, we were able to get a better estimate of the bias across school-level CITS evaluations, providing important new evidence on the internal validity of the design in this setting.

Across the four interventions, we estimate mean bias to be 0.03 school-level SD, or 0.01 pupil-level SD. One way of putting these findings into perspective is to compare it to the effect sizes typically found in the education literature. Median effect sizes for academic achievement in RCTs in education are between 0.07 around 0.1 pupil-level SD (Evans & Yuan, 2020; Kraft, 2020). Our estimate of bias thus represents 10–14% of the effect size that researchers might expect to find in similar research. Put another way, a recent review of effect sizes recommended deeming 0–0.05 pupil-level SD as a “small” effect, suggesting that 0.01 is also small (Kraft, 2020).

At the intervention level, the differences between the RCT impact estimates and the CITS impact estimates are all below 0.13 school-level SD or 0.04 pupil-level SD in absolute value. This suggests that our (across intervention) estimate of bias is not the result of large intervention-level bias but with random sign, canceling each other out. Rather, our small across intervention estimate of bias likely reflects similarly small bias at the intervention level.

An important caveat on our overall findings is that the choice between the baseline mean, baseline linear trends and baseline non-linear trend models does have an effect on our estimates (see Appendix Table A2). In line with previous research in this literature, this highlights the importance of choosing the correct model specifications through careful inspection of the pre-trends (St. Clair et al., 2016). However, our findings also serve as reminder that this is not entirely straightforward. Looking across the rows in

**Table A2** it is clear that we did not always choose the CITS specification that minimized the difference in impact estimates between the CITS and the original EEF RCT.

How do our results compare to those from similar education WSCs? St. Clair et al. (2016) report results from three within-study comparisons of education interventions. Two of the three interventions showed modellable pretreatment trends and in these cases the CITS impact estimates differed from the RCT impact estimates by 0.01–0.07 school-level SD in the authors' preferred specifications. The third of their three interventions showed pretreatment trends with more than one turning point and, as hypothesized, the CITS impact estimates showed much larger differences with those from the RCT. Across the three CITS within-study comparisons reported by Hallberg et al. (2020), the CITS estimates for one year post-intervention differed from the RCT estimates in absolute magnitude by 0.05–0.08 school-level SD, depending on the model specification adopted. Our estimate of 0.03 school-level SD bias is therefore quantitatively similar to those in the existing literature on CITS within-study comparisons.

Weidmann and Miratrix (2021) report within-study comparisons of 14 interventions but using propensity score matching rather than CITS as their non-experimental estimator. Their meta-analytically derived estimate of mean bias across the interventions is 0.01 pupil-level SD. Our overall estimates of bias are hence also very similar to those from within-study comparisons using propensity score matching. The set of interventions studied by Weidmann and Miratrix are drawn from the same database as ours, and two of these interventions (LERS and FLIP) overlap with those in our study. The CITS impact estimates for the LERS intervention differs from the CITS RCT impact estimate by approximately 0.01 pupil-level SD. The equivalent result in Weidmann & Miratrix is reported graphically rather than numerically but appears very similar in magnitude. Our CITS impact estimate for the FLIP intervention differs from the RCT by approximately 0.05 pupil-level SD, which again appears very similar to the equivalent result in Weidmann and Miratrix. Although based on only two interventions, the results reviewed in this paragraph suggest that CITS and PSM methods, as implemented in these two papers, perform similarly.

### **Limitations**

These findings should, of course, be interpreted in light of the limitations of this research. In particular, as we have emphasized, selection bias depends upon the setting and characteristics of the specific intervention. The treatments studied here involve school-level interventions deemed by the EEF to have shown evidence of promise in previous evaluations. All schools that participated in these trials volunteered to take part and, as can be seen from **Table 2**, these volunteer schools tended to have higher proportions of pupils eligible for free school meals than those in the surrounding areas. While volunteer schools appear to be comparable to surrounding schools in terms of their other observable characteristics, they may also differ in terms of unobservable characteristics, such as the extent to which they are open to adopting evidence-based practices. This limits the generalizability of our findings to other settings.

Relatedly, our sample of interventions is—by necessity—composed entirely of interventions that have previously been evaluated using RCTs. This highlights an important

constraint on the generalization of our findings to settings where an RCT is infeasible for ethical or logistical reasons. This is an important limitation, given that such settings are a prime candidate for the use of non-experimental designs. Having said that, the findings are still informative with respect to CITS designs used to retrospectively evaluate policies in similar settings in which random allocation was feasible but was not in fact used. Indeed, our view is that many of the interventions previously evaluated using CITS could have been randomly allocated but were not (e.g. Bell et al., 2016; Gandhi et al., 2018; Henry et al., 2020; Jacob et al., 2017; Strunk et al., 2016). In addition, as we discuss below, the results are also useful in assessing the validity of CITS designs for analysis of interventions in which attrition or noncompliance has raised doubts about the validity of an RCT.

A further limitation of the present research is that we only reevaluate four interventions; a larger sample of interventions would do a better job of purging noise from the estimates of bias. Finally, our within-study comparisons exclusively used CITS specifications that projected the counterfactual outcomes one period into the future. Other research has found that CITS that project more periods into the future show larger bias (Hallberg et al., 2020). Our study is silent on this and therefore does not provide any warrant for using such CITS specifications.

### **Implications**

Despite these limitations, we believe these findings have implications for the practice of impact evaluation in education. Hierarchies of evidence quality often privilege the use of RCTs over those from observational studies (Clarke et al., 2014) for the understandable reason that they provide impact estimates that are unbiased in expectation. However, this tends to ignore two important limitations of such studies in education. The first is that many such RCTs are underpowered, yielding findings that are either uninformative with respect to impact (Lortie-Forgues & Inglis, 2019) or yield apparently positive findings but with exaggerated effect sizes (Sims et al., 2021). A second limitation of RCTs in education is that they often suffer from high levels of attrition or noncompliance (Edovald & Nevill, 2020). This limits the internal validity of the trial by undermining the group-level balance initially achieved through randomization.

Our results suggest that CITS designs show substantively small bias when used to estimate promising education interventions among volunteer schools. Moreover, in line with theory, differences between the CITS and RCT estimates are particularly small in cases that show stable, easily-modellable and parallel pre-trends. This implies that—if used judiciously—CITS evaluations can be used to complement RCT evaluations in such settings. In particular, school-level CITS designs could be used to corroborate effect size estimates from weakly-powered trials, in order to check that these are not overly influenced by residual imbalance from the RCT. Furthermore, school-level CITS estimates can be used to recover non-experimental impact estimates from “broken” trials in which attrition or contamination has compromised the initial random allocation. In the case of a trial with high attrition, a CITS can help wherever it is possible to obtain the school-level outcome data on the full sample. In the case of trial with treatment contamination, a CITS can help wherever the contamination has been caused by control group schools

seeking the treatment *as a result* of being allocated to the control group. The CITS can address this problem by using a comparison group that does not overlap with the RCT control group. Our results suggest that these non-experimental impact estimates are likely to provide a good second-best impact estimate in a situation that would otherwise yield little useful evidence.

It should be noted that other non-experimental methods, such as careful propensity score matching designs (Weidmann & Miratrix, 2021), can also be used to complement RCTs in the same way. Indeed, our results suggest that these two methods result in very similar levels of bias in this setting. The main advantage of the school-level CITS relative to propensity score matching is that it can often be implemented using non-sensitive, publicly available, school-level data (Jacob et al., 2014). The corresponding downside, as this paper has illustrated, is the need for a consistently recorded time series, which is often not available due to exam reforms or the cancelations of examinations in certain years. The relative benefits of the two non-experimental designs are therefore likely to depend on data availability. Either way, the evidence presented here suggests CITS are a form of non-experimental design that can estimate impact with substantively small bias in similar settings.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Funding

Sam Sims, Jake Anders and Laura Zieger received funding from the Education Endowment Foundation. Laura Zieger has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant [agreement no. 765400].

## ORCID

Sam Sims  <http://orcid.org/0000-0002-5585-8202>

Jake Anders  <http://orcid.org/0000-0003-0930-2884>

## References

- Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, 38(2), 318–335. <https://doi.org/10.3102/0162373715617549>
- Berk, R. A. (2004). *Regression analysis: A constructive critique*. SAGE.
- 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. <https://doi.org/10.1002/pam.20637>
- 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(1), 3–49. <https://doi.org/10.1177/0193841X02239017>

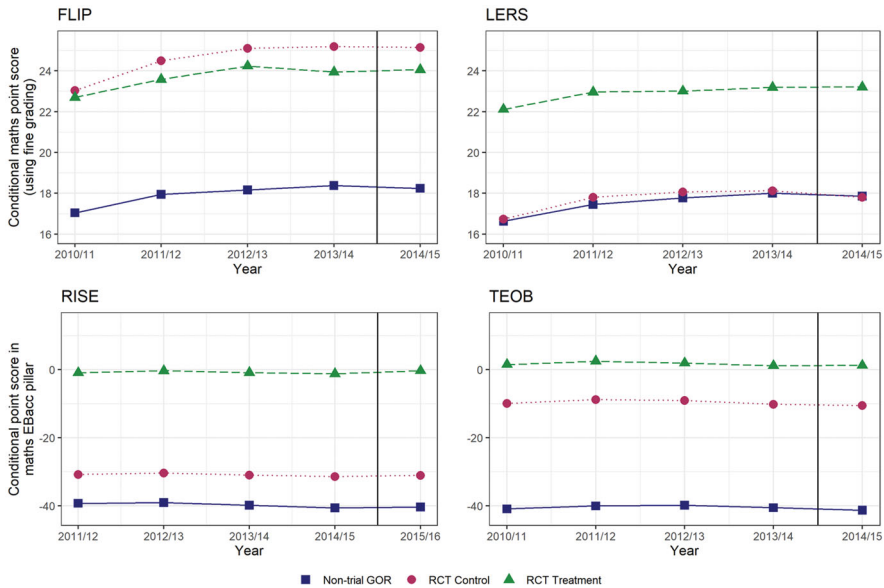


- Bloom, S., Ham, S., Melton, L., & O'Brien, J. (2001). *Evaluating the accelerated schools approach: A look at early implementation and impacts in eight elementary schools*. Manpower Demonstration Research Corporation.
- Bloom, H., Michalopoulos, C., Hill, C., & Lei, Y. (2002). *Can nonexperimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs?* (MDRC Working Papers on Research Methodology). [https://www.mdrc.org/sites/default/files/full\\_43.pdf](https://www.mdrc.org/sites/default/files/full_43.pdf)
- Card, D., & Krueger, A. B. (1993). *Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania* (No. w4509). National Bureau of Economic Research.
- Cattaneo, M. D., Crump, R. K., Farrell, M. H., & Feng, Y. (2019). On binscatter. *arXiv preprint arXiv:1902.09608*.
- Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, 33(2), 339–360. <https://doi.org/10.1007/s11245-013-9220-9>
- Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2), 403–429. <https://doi.org/10.1002/pam.22051>
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.
- 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(4), 724–750. <https://doi.org/10.1002/pam.20375>
- Coopersmith, J., Cook, T. D., Zurovac, J., Chaplin, D., & Forrow, L. V. (2022). Internal and external validity of the comparative interrupted time-series design: A meta-analysis. *Journal of Policy Analysis and Management*, 41(1), 252–277. <https://doi.org/10.1002/pam.22361>
- Cunningham, S. (2020). *Causal inference: The mixtape* (Vol. 18). Tufte-Latex.GoogleCode.com.
- Dawson, A., Yeomans, E., & Brown, E. (2018). Methodological challenges in education RCTs: Reflections from England's education endowment foundation. *Educational Research*, 60(3), 292–310. <https://doi.org/10.1080/00131881.2018.1500079>
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*, 210, 2–21. <https://doi.org/10.1016/j.socscimed.2017.12.005>
- Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, 35(2), 252–279. <https://doi.org/10.3102/0162373712467080>
- Edoald, T., & Nevill, C. (2020). Working out what works: The case of the education endowment foundation in England. *ECNU Review of Education*, 4(1), 46–64.
- Evans, D., & Yuan, F. (2020). *How big are effect sizes in international education studies?* Centre for Global Development Working Paper 545.
- Eyles, A., & Machin, S. (2019). The introduction of academy schools to England's education. *Journal of the European Economic Association*, 17(4), 1107–1146. <https://doi.org/10.1093/jeea/jvy021>
- Ferraro, P. J., & Miranda, J. J. (2014). The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark. *Journal of Economic Behavior and Organization*, 107, 344–365. <https://doi.org/10.1016/j.jebo.2014.03.008>
- 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.
- Gandhi, A. G., Slama, R., Park, S. J., Russo, P., Winner, K., Bzura, R., Jones, W., & Williamson, S. (2018). Focusing on the whole student: An evaluation of Massachusetts's wraparound zone initiative. *Journal of Research on Educational Effectiveness*, 11(2), 240–266. <https://doi.org/10.1080/19345747.2017.1413691>

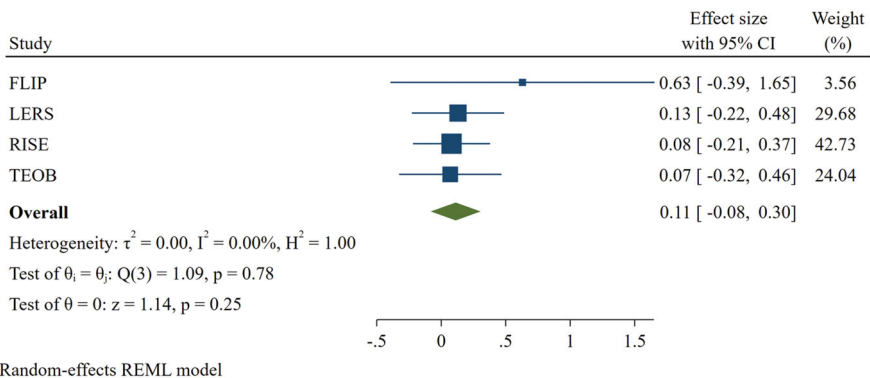
- Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, 60(4), 328–331. <https://doi.org/10.1198/000313006X152649>
- Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of “no child left behind on teachers” work environments and job attitudes. *Educational Evaluation and Policy Analysis*, 36(4), 417–436. <https://doi.org/10.3102/0162373714533817>
- Jacob, B., Dynarski, S., Frank, K., & Schneider, B. (2017). Are expectations alone enough? Estimating the effect of a mandatory college-prep curriculum in Michigan. *Educational Evaluation and Policy Analysis*, 39(2), 333–360. <https://doi.org/10.3102/0162373716685823>
- 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. <https://doi.org/10.3102/0162373713485814>
- Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychologist*, 51(3–4), 395–405.
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4), 604–620.
- Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, 34(2), 209–231. <https://doi.org/10.3102/0162373711431604>
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166. <https://doi.org/10.3102/0013189X19832850>
- Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, 589(1), 63–93. <https://doi.org/10.1177/0002716203254879>
- Hallberg, K., Wong, V. C., & Cook, T. D. (2016). *Evaluating methods for selecting school-level comparisons in quasi-experimental designs: Results from a within-study comparison*. EdPolicyWorks Working Paper Series No. 47.
- Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. *Educational Researcher*, 47(5), 295–306. <https://doi.org/10.3102/0013189X18769302>
- Hallberg, K., Williams, R., & Swanlund, A. (2020). Improving the use of aggregate longitudinal data on school performance to assess program effectiveness: Evidence from three within study comparisons. *Journal of Research on Educational Effectiveness*, 13(3), 518–545. <https://doi.org/10.1080/19345747.2019.1698088>
- Henry, G. T., Pham, L. D., Kho, A., & Zimmer, R. (2020). Peeking into the black box of school turnaround: A formal test of mediators and suppressors. *Educational Evaluation and Policy Analysis*, 42(2), 232–256. <https://doi.org/10.3102/0162373720908600>
- McKenzie, D., Stillman, S., & Gibson, J. (2010). How important is selection? Experimental vs. non-experimental measures of the income gains from migration. *Journal of the European Economic Association*, 8(4), 913–945. <https://doi.org/10.1111/j.1542-4774.2010.tb00544.x>
- Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, 31(4), 463–479. <https://doi.org/10.3102/0162373709343964>
- Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 2, pp. 295–316). Russell Sage Foundation.
- Rosenbaum, P. R. (2017). *Observation and experiment*. Harvard University Press.
- Rudd, P., Aguilera, A., Elliott, L., & Chambers, E. (2017). *Mathsflip: Flipped learning. Evaluation report and executive summary*. Education Endowment Foundation.

- 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. *British Medical Journal*, 328(7439), 560.
- Sibieta, L., Greaves, E., & Sianesi, B. (2016). *Increasing pupil motivation: Evaluation report and executive summary*. Education Endowment Foundation.
- Sims, S., Anders, J., Inglis, M., & Lortie-Forgue, H. (2021). *Quantifying 'promising trials bias' in randomized controlled trials in education*. CEPEO Working Paper Series, 20–16. UCL.
- Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and non-experimental estimates in within-study comparisons. *Evaluation Review*, 42(2), 214–247.
- Strunk, K. O., Marsh, J. A., Hashim, A. K., & Bush-Mecenas, S. (2016). Innovation and a return to the status quo: A mixed-methods study of school reconstitution. *Educational Evaluation and Policy Analysis*, 38(3), 549–577. <https://doi.org/10.3102/0162373716642517>
- Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the US institute of education sciences. *International Journal of Research and Method in Education*, 39(3), 255–267. <https://doi.org/10.1080/1743727X.2016.1150454>
- St. Clair, T., & Cook, T. D. (2015). Difference-in-differences methods in public finance. *National Tax Journal*, 68(2), 319–338. <https://doi.org/10.17310/ntj.2015.2.04>
- 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. <https://doi.org/10.3102/1076998616636854>
- Weidmann, B., & Miratrix, L. (2021). Lurking inferential monsters: Quantifying selection bias in non-experimental evaluations of school programs. *Journal of Policy Analysis and Management*, 40(3), 964–986. <https://doi.org/10.1002/pam.22236>
- What Works Clearinghouse. (2017). *Standards handbook version 4.1*. <https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf>
- Wiggins, M., Jerrim, J., Tripney, J., Khatwa, M., & Gough, D. (2019). *The RISE project: Evidence-informed school improvement*. Education Endowment Foundation.
- Wiggins, M., Sawtell, M., & Jerrim, J. (2017). *Learner response system: Evaluation report and executive summary*. Education Endowment Foundation.
- 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. <https://doi.org/10.1080/19345747.2013.878011>
- Wong, V. C., Steiner, P. M., & Anglin, K. L. (2018). What can be learned from empirical evaluations of nonexperimental methods? *Evaluation Review*, 42(2), 147–175.
- Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness*, 10(1), 207–236. <https://doi.org/10.1080/19345747.2016.1164781>
- Worth, J., Sizmur, J., Walker, J., Bradshaw, S., & Styles, B. (2017). *Teacher observation. Evaluation report and executive summary*. Education Endowment Foundation.

Appendix



**Figure A1.** Treatment and comparison group trends across the four interventions. Notes: FLIP: flipper learning- $N = 4,084$  schools; LERS: learner response system- $N = 7,165$  schools; RISE:  $N = 2,224$  schools; TEOB: teacher observation- $N = 2,829$  schools; All Ns include treatment, control and comparison group schools. Non-trial GOR: schools in the same Government Office Region of England that did not participate in the original RCT; RCT Treatment: schools that were in the treatment group in the original RCT; RCT Control: schools that were in the control group in the original RCT; The years left to the vertical line are used as pretreatment year and the year right to the vertical is the post-treatment year. The mathematics scores were conditioned upon prior achievement, % special educational needs (except for RISE), % free school meals, % English as additional language, % pupil low prior achievement at the end of KS2 (only RISE and TEOB), school type and school size.



**Figure A2.** Forest plot of the RCT-CITS for the four interventions. Notes: Figure A2 shows a forest plot of the effect size differences taken from Table A4. The meta-analytic (precision weighted) average of these effect size differences is 0.11 school-level standard deviations, with a 95% CI of -0.08–0.3 school-level standard deviations. It should be noted that these SE are upwardly biased (see the notes to Table A4) and should be considered upper bounds on the true SE.

**Table A1.** Calculating the school-level effect sizes for the RCT impact estimates.

Study	EEF RCT pupil-level effect size	EEF pupil-level pooled SD	Measure correction formula	School-level pooled SD	RCT effect size using school-level pooled SD	Notes
FLIP	0.09	5.28	None	1.82	0.26	EEF-reported effect size extracted from Table 4 in the FLIP report.
LERS	0.00	22.22	Unknown but moot due to zero effect	1.83	0.00	EEF-reported effect size for Cohort 1 extracted from Table 10 in the LERS report.
RISE	0.09	10.79	$1/6 * SD(X) = SD(Y)$ with Y the original scale of RISE and X our scale	5.06	0.19	EEF-reported effect size for cohort A extracted from Table 12 in the RISE report.
TEOB	0.06	11.98	None	4.83	0.15	EEF-reported effect size for Cohort 1 extracted from Table 13 in the TEOP report.

Study acronyms can be found in Table 1.

**Table A2.** CITS results using all three specifications.

Study	EEF RCT School-level effect size	CITS		
		Baseline mean	Linear trends	Non-linear trends
FLIP	0.26 (0.55)	0.14 (0.21)	0.04 (0.28)	<b>0.13</b> (0.51)
LERS	0.00 (0.84)	<b>-0.03</b> (0.10)	0.17 (0.13)	0.40 (0.25)
RISE	0.19 (0.13)	<b>0.16</b> (0.08)	0.06 (0.11)	0.16 (0.20)
TEOB	0.15 (0.10)	0.08 (0.06)	0.01 (0.08)	<b>0.22</b> (0.15)

Study acronyms can be found in Table 1. Bolded numbers represent the specification adopted in our preferred estimates. Coefficients and standard errors are shown in school-level SD. Numbers in parentheses are standard errors.

**Table A3.** Summary table for the within-study comparisons.

Intervention		Original EEF RCT										Within-study comparison CITS		
Name	Brief description	Estimand	No. of schools	Location of schools	Control variables	Attrition	Estimand	No. of schools	Periods pre/post intervention	Preferred CITS specification	Control variables	Location of schools		
FLIP: flipped learning	Pupils study new math at home, then consolidate through work in class.	Intention to treat	24	SE, SW, WM	KS1 SATs prior test scores	Schools: 0% pupils; 7%	Intention to treat	4,181	Pre: 4 Post: 1	Baseline non-linear trends	KS1 SATs, %FSM, %SEN, school type and size.	SE, SW, WM		
LEERS: learner response system	Immediate in-class pupil feedback from teachers via electronic devices.	Intention to treat	97	EE, LO, NW, YH	KS1 SATs prior test scores	Schools: 3.1% pupils; 11.7%	Intention to treat	7,276	Pre: 4 Post: 1	Baseline mean	KS1 SATs, %FSM, %EAL, %SEN, school type and size.	EE, LO, NW, YH		
RISE: rise programme	Teacher "Research Leads" supporting teachers' use of research.	Intention to treat	40	LO, NE, NW, SE, SW, WM, YH	KS2 SATs prior test scores	Schools: 2.5% pupils; 8.8%	Intention to treat	2,289	Pre: 4 Post: 1	Baseline mean	KS2 SATs, %FSM, %EAL, %PLO, school type and size.	LO, NE, NW, SE, SW, WM, YH		
TEOB: teacher observation	Teacher PD based on programme of structured peer lesson observations.	Intention to treat	82	All of England	KS2 SATs prior test scores	Schools: 0% pupils; 1.4%	Intention to treat	2,918	Pre: 4 Post: 1	Baseline non-linear trends	KS2 SATs, %FSM, %EAL, %SEN, %PLO, school type and size.	All of England		

EEF: education endowment foundation; RCT: randomized controlled trial; CITS: comparative interrupted time series; KS1: key stage 1 (age 5–7); KS2: key stage 2 (age 8–11); SATs: statutory attainment tests; FSM: free school meals; EAL: English as an additional language; SEN: special educational needs; EE: East of England; LO: London; NE: North East; NW: North West; SE: South East; SW: South West; WM: West Midlands; YH: Yorkshire and the Humber.

**Table A4.** Summary table for the within-study comparisons (expressed in school-level SD).

	RCT		CITS		RCT-CITS	
	ES	SE	ES	SE	ES	SE
FLIP	0.76	0.41	0.13	0.43	0.63	0.52
LERS	0.12	0.19	-0.02	0.12	0.13	0.18
RISE	0.23	0.16	0.15	0.07	0.08	0.15
TEOB	0.27	0.17	0.21	0.10	0.07	0.20

Table A4 summarizes the inputs to our meta-analysis of the effect size differences across our four within-study comparisons. All effect sizes (ES) and standard errors (SE) are expressed in terms of school-level standard deviations. The two RCT columns show the results of re-estimating the impact of the four interventions using the experimental treatment and control group schools, using school-level data. In each case, the models use the same outcomes, measured at the same timepoint, for the same treatment cohorts, using the same covariates (albeit measured at school-level), as the original EEF RCT. The equivalent results derived from pupil-level data, as reported in the original EEF evaluation reports, can be found in the "RCT" column of Table 4. The ES differ from those in Table 4 because we do not have access to the underlying micro data, meaning that we cannot control for prior attainment at the pupil-level. In addition, the SE are upwardly biased because they do not take account of the units (pupils) within each cluster (schools). The SE should therefore be considered upper bounds on the true SE. The CITS columns contain the results from Table 3. The ES in the RCT-CITS columns are simply the ES from the RCT column minus the ES from the CITS columns. The SE in the RCT-CITS columns are derived using the procedure from footnote 3 of Steiner and Wong (2018) and are upwardly biased for the same reason that the SE in the RCT columns are upwardly biased. The results in the RCT-CITS columns form the bases for our meta-analysis, as reported in Figure A2.