Education
Endowment
Foundation

Can we replicate the findings of EEF trials using school-level comparative interrupted time series evaluations? Non-technical report

June 2021

Authors:

Sam Sims, Jake Anders, and Laura Zieger.

UCL Institute of Education

CENTRE FOR EDUCATION POLICY
& EQUALISING OPPORTUNITIES

Can we replicate the findings of EEF trials using school-level comparative interrupted time series evaluations? Non-technical report

Sam Sims, Jake Anders, Laura Zieger

Centre for Education Policy and Equalising Opportunities
UCL Institute of Education
20 Bedford Way
London
WC1H 0AL

Executive Summary

- A school-level comparative interrupted time series (CITS) is a non-experimental evaluation design, which can often be conducted with publicly available data.
- Across four EEF-funded interventions, we found that impact estimates from CITS were on average 0.01 pupil-level standard deviations away from impact estimates from previous experimental evaluations of the same intervention.
- In line with our expectations, where outcomes followed a parallel trend across the treatment and control groups during the pre-treatment period, CITS designs came closer to reproducing the results of experimental evaluations.
- It is important to note that regular changes in the way that school-level average attainment is measured and reported means that many EEF-funded interventions cannot be evaluated using CITS methods.
- Having said that, when the relevant data are available, CITS are a reasonable alternative to experimental evaluations where it is either infeasible or unethical to withhold treatment randomly.
- In addition, data permitting, it may be worthwhile running CITS evaluations alongside experimental evaluations where recruitment falls short of target, or where certain schools are known not to be complying with their random allocation to treatment or control. In such circumstances, a complementary CITS evaluation is a useful way to check whether the shortcomings of the original experimental evaluation have introduced error in the findings.

Table of Contents

1. Introduction

Why discuss non-experimental evaluation designs?

The EEF generally funds experimental (randomised controlled trial) evaluations to test whether various interventions improve pupil achievement. Most often, this involves dividing recruited schools into two groups at random, with the first implementing an intervention (the treatment group) and the second not implementing the intervention (the control group). In the same way that randomly flipping a fair coin many times over would be expected to yield an equal number of heads and tails, randomly allocating schools to treatment and control groups would be expected to balance the characteristics of the schools across the two groups. For example, the proportion of pupils eligible for free school meals will be very similar in treatment in and control group schools. Crucially, even unmeasured pupil characteristics will be very similar across the two groups. This similarity between the two groups means that any differences in outcomes are likely to be the result of the one remaining systematic difference: receipt of the intervention.

Where changes are introduced into schools without using randomisation, the schools that do receive the intervention will likely differ from those schools that do not. For example, the leadership of the schools that do receive the intervention might be generally more proactive with regards to school improvement. Since treated and untreated schools differ even before the intervention takes places, it is difficult to attribute any subsequent differences in outcomes to the intervention. There is therefore a strong case for the default position being to use experimental studies, where possible (Anders et al, 2017).

In practice, however, there are a number of reasons why an experimental design may not be optimal:

- Case 1: it may be infeasible to use experimental methods. For example, if the intervention has a high cost, then randomly allocating a *large enough* sample of schools to receive the intervention may not be affordable. Studying schools that are already using the intervention would be more affordable.
- Case 2: it may be unethical to deny some schools the intervention randomly. For example, if previous research provides strong evidence that a reading intervention helps disadvantaged pupils to catch up, then it may be considered morally unacceptable to prevent some pupils from benefiting.

In Case 1 and 2, it may be better for researchers to adopt a non-experimental evaluation from the outset.

In addition to the above, there are cases in which it may have been reasonable to proceed initially with an experimental evaluation, but it subsequently turns out that a non-experimental evaluation may be worthwhile:

- Case 3: recruitment proves harder than expected, which means achieving a suitable sample size for the experimental evaluation is not possible. A fair coin tends toward returning an equal number of heads and tails *the more times that it is flipped*. Relying on the coin-flipping logic to balance characteristics of treatment and control groups therefore requires the random allocation to treatment and control groups of a *sufficiently large* number of schools.
- Case 4: survey (or other) data suggest that schools in the control group have managed to access the intervention through other routes. For example, the control group schools purchase a close-substitute intervention on the open market (e.g. Jacob et al., 2015). This contamination means that comparing the outcomes across the original treatment and control groups no longer provides a contrast between receipt and non-receipt of the intervention.
- Case 5: schools do not provide outcome data and the reasons for this differ between treatment and control group schools. For example, control group schools might lose motivation to participate once they are allocated to the control group. This makes it impossible to include these schools in the analysis, which undermines the initial similarity between the treatment and control groups.

In Cases 3-5, it may be useful for researchers to complement the experimental evaluation with a separate non-experimental evaluation, or in more extreme cases to abandon the experiment and switch to a non-experimental design entirely.

Since non-experimental evaluation methods do not randomly allocate the intervention, we cannot rely on the simple coin flipping logic to balance the characteristics of schools in the treatment and control groups. Using a non-experimental approach to infer that any differences in outcomes are the result of receiving the treatment (rather than pre-existing differences) therefore requires additional assumptions. Such 'identifying' assumptions by their nature cannot be directly verified using the data in any given evaluation. However, it is possible to tests whether non-experimental methods can reproduce the results from experimental evaluations *of the same intervention*. This 'within-study comparison' approach was first introduced by Lalonde (1986) and has since been applied in various settings. For example, Chaplin et al. (2017) test whether regression discontinuity designs[1] can reproduce the results of experimental evaluations and Weidmann and Miratrix (2020) test whether propensity score matching designs[2] can do the same for EEF-funded experimental interventions.

[1] Regression discontinuity designs compare units (e.g. schools) that are assigned to receive a certain intervention because they 'score' just above a certain value on a given variable to other units that are not assigned to receive that same intervention because they score just below a certain value. For a non-technical comparison of regression discontinuity designs and experimental designs, see Hallberg, Wing, Wong and Cook (2013).

[2] In a simple experimental design, all units (e.g. schools) have the same probability of being assigned to treatment and control. Using our coin flip analogy, this probability is equal to 0.5. In a non-experimental study, schools have unequal probability of receiving a treatment. Propensity score matching designs estimate the probability of each treated school receiving a treatment and then compares their outcomes to non-treated schools with the same estimated probability of being assigned to treatment. For a non-technical comparison of propensity score matching designs and experimental designs, see Rosenbaum (2017).

Why discuss comparative interrupted time series?

This report focuses on whether one particular non-experimental method can reproduce the results from experimental evaluations: the comparative interrupted time series (CITS) design. We will discuss the CITS design in further detail below. For now, the basic idea is to compare the way in which outcomes in the treatment group deviate from trend after an intervention is introduced, relative to the way in which outcomes in the control group[3] deviate from trend at the same point in time. Under certain assumptions, the difference between these deviations can be interpreted as the effect of the intervention.

We have chosen to focus on CITS in particular because this design is in many ways well suited to evaluating education interventions. First, conducting a CITS evaluation requires a well-measured outcome variable, recorded consistently across multiple time periods (a time series). In England, high-stakes school examination results are recorded and made publicly available in (almost) every academic year. Second, the emphasis on modelling and then projecting the outcome variable in CITS, 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, Goddard, & Kim, 2014). For example, summary measures of pupil achievement for each school in England are publicly available via the UK Department for Education (DfE)'s *Compare School Performance* website. This minimises the use of individuals' personal data and, more practically, avoids the need to go through a lengthy application process for pupil-level data from the DfE's *National Pupil Database*. Finally, as we will explain in subsequent sections, the assumptions of the CITS design are often quite plausible in education settings.

Aims and structure of this report

This report provides a non-technical discussion of a research project that compared the results from four CITS evaluations of EEF-funded interventions with the original experimental evaluations of those same interventions. We have intentionally used non-technical language throughout, in order to make the contents accessible to non-specialists. Those looking for a more detailed treatment are referred to the technical report (Sims, Anders, & Zieger, 2021).

This document addresses three research questions (RQs). The first, and most important, relates to whether the assumptions underpinning CITS methods are likely to hold in evaluations typical of those commissioned by EEF:

[3] Ideally, we would refer to the untreated group in non-experimental designs such as CITS as the 'comparison group'. However, since this paper is written for a non-technical audience, we have kept to the 'control group' language in order to minimise jargon.

> **RQ1**: How closely can CITS reproduce the impact estimates from experimental evaluations of EEF-funded interventions?

As well as looking at the average result across the four studies, we also investigate the conditions under which particular CITS gets closer to the corresponding RCT estimates. To this end, we empirically test whether the theory underpinning CITS can help us understand when CITS designs are likely to be valid:

> **RQ2**: How does the closeness of the CITS and RCT estimates vary depending on the shape of the trends in the outcome variables in the pre-treatment period?

This research is closely related to prior work by Weidmann and Miratrix (2020), which uses similar data to test whether an alternative non-experimental design – propensity score matching – can reproduce the results of experimental evaluations of EEF-funded interventions. This project complements their work by testing a different non-experimental evaluation method and by exploring which of the two might work better, and under what circumstances:

> **RQ3**: Does CITS or matching get closer to reproducing the impact estimates from the EEF RCTs?

In initial planning, we considered further research questions unpacking these points. However, because of the relatively limited number of cases in which it was ultimately feasible to carry out CITS and obtain impact estimates that would be directly comparable with those reported in the original RCT, we have judged that it was not possible to carry these out in a meaningful way. Likewise, our answers to RQ2 and RQ3 should also interpreted with caution, given the smaller number of studies on which they are based.

The remainder of the report is structured as follows. Section 2 provides the intuition behind the CITS design, provides a non-technical description of our methods for comparing the CITS and RCT estimates, and sets out our findings against the three research questions listed above. Section 3 then discusses the potential of CITS designs to assist in the evaluation work that EEF funds in the future. In particular, we set out how and when CITS can be used to address the different scenarios (Cases 1-5 above) in which experimental evaluations may not be optimal.

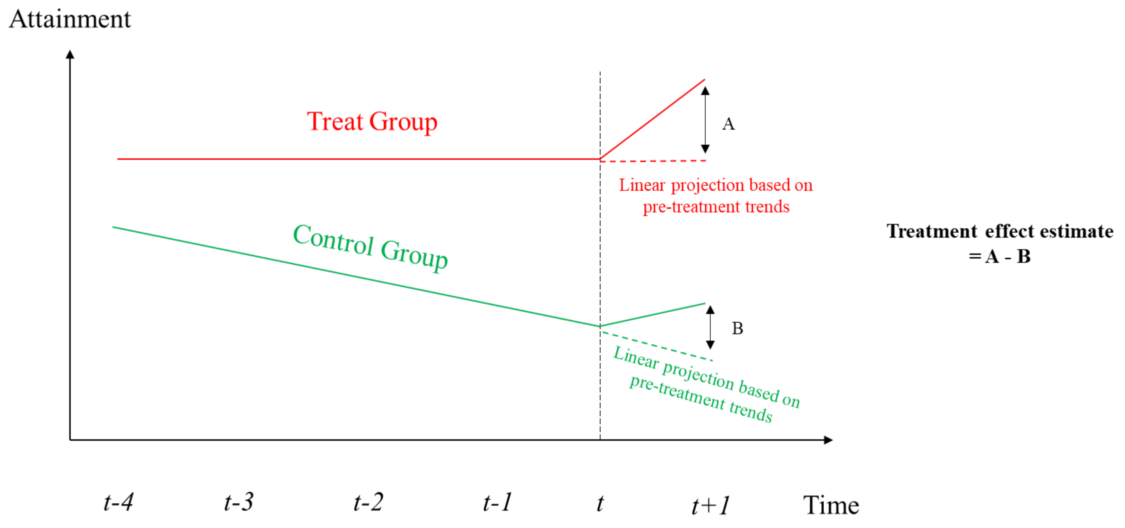
2. Non-technical summary of the project

What is a CITS evaluation?

We illustrate how a CITS evaluation works using the stylised diagram in Figure 1. This shows the evolution of the average academic performance of two groups of schools over time: a treatment group (red) and a control group (green). Schools have not been randomly assigned to the two groups. An intervention is introduced in treatment group schools at time t (horizontal axis). The solid lines plot average academic attainment (vertical axis) in the four periods prior to the treatment being introduced (t-4, t-3, t-2 and t-1), the period in which treatment begins (t), and one period afterwards (t+1). The CITS design generates an impact estimate using the following procedure:

1. project the pre-treatment trend in the treatment group into the post-treatment period (dashed red line);
2. compare the projected trend in the treatment group to the observed trend (solid red line to the right of time t), which gives vertical distance A;
3. project the pre-treatment trend in the control group into the post-treatment period (dashed green line);
4. compare the projected trend in the control group to the observed trend (solid green line to the right of time t), which gives vertical distance B;
5. assume that the deviation from trend in the control group (B) represents the deviation from trend that would have been observed in the treatment group if in fact the intervention had not been introduced;
6. conclude that A minus B is the causal effect of the intervention.

Figure 1: hypothetical illustration of a CITS design


Attainment
Treat Group
A
Linear projection based on pre-treatment trends
Control Group
B
Linear projection based on pre-treatment trends
Treatment effect estimate = A - B
t-4
t-3
t-2
t-1
t
t+1
Time

The intuition behind the CITS design is that the observed deviation from trend in the control group in the period after the intervention is a proxy for the deviation from trend in outcomes that would have occurred in the treatment group, had there in fact been no intervention. Where this assumption holds true, the CITS method can identify the causal effect of the intervention. This assumption is in some ways plausible in education settings, since all state-funded schools are subject to many common influences e.g. national government policy, changes to the examinations system, the current graduate job market in which they recruit new teachers, and so on. For these reasons, the control group can plausibly proxy for the treatment group had the intervention not been introduced. However, if anything influences treatment and control group schools differently between time t and time t+1, in a way that is not captured by the pre-treatment trends across the two groups, then the impact estimates from the CITS will confuse the effect of the intervention with the effect of this outside influence (St. Clair & Cook, 2015).

Unfortunately, as with all 'identifying' assumptions for non-experimental evaluation methods, we can never directly verify whether the CITS assumption holds because we never observe what would have happened to outcomes in the treatment group in period t+1 had they in fact not been treated. Nevertheless, we can assess the plausibility of the assumption via inspecting the pre-treatment trends. Parallel pre-treatment trends suggest common influences before the intervention and – by extrapolation – common influences after the intervention. This makes the CITS assumption more plausible. Non-parallel but stable, linear pre-treatment trends (as pictured in the diagram above) suggest that – while there may be different influences on attainment in treatment and control group schools – these are operating in a stable way within each group over time. Again, this makes the CITS assumption more plausible. However, if the pre-treatment trends are non-parallel and/or unstable over time, then the assumption behind the CITS becomes much less plausible. For more information on how the CITS approach can be adapted to fit the specific shape of the pre-trends, see the technical report.

What we did in this project: within-study comparisons

While inspecting the pre-treatment trends can cast doubt on whether the CITS assumptions hold, it cannot tell us when they do in fact hold. This is because, even if there are common influences operating on the treatment and control groups up until the point the intervention is introduced, there may be an event/shock that affects only the control group or only the treatment group schools at time t+1. In this situation, the outcomes in the control group do not provide a good proxy for what would have happened to outcomes in the treatment group, had they not in fact received the intervention.

So how can we ever have confidence in the CITS design? As alluded to in the introduction, we can provide empirical evidence on whether the CITS assumptions are likely to hold true in a given setting by comparing the results from an experimental evaluation of an intervention with the results from a

CITS evaluation of the same intervention. This method is known as a ‘within-study comparison’. We set out to do just this by using the CITS design to re-evaluate interventions previously tested using experimental methods in projects funded by the Education Endowment Foundation (EEF).

These interventions can be re-evaluated using school-level CITS if five conditions hold: 1) the intervention was originally randomised at school-level, 2) the trial used all pupils in the treated cohorts’ results in standardised 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 the trials/interventions that they have funded that met criteria 1-3 as of December 2019, which amounted to 19 in total.[4] 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).[5] Similarly, three interventions had to be dropped because the relevant time series included a year in which we had missing data due to, for example, a widespread boycott of the age 11 examinations.[6] 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 (did not meet criterion 5).[7] In addition, one trial had to be dropped because of an important discrepancy between the original EEF evaluation report and the EEF database.[8] This left four interventions: two in primary schools and two in secondary schools (see Table 1). They are diverse, ranging from teacher professional development (PD) to initiatives encouraging evidence-based teaching, to the use of educational technology to adapt instruction. The effect sizes estimated in the original experimental evaluations range from -0.01 to 0.09. We re-evaluated all of these interventions using the CITS design (see technical report for details).

[4] EEF originally sent us 21 interventions but two of these (Affordable Maths Tuition and Engage in Education) did not use school-level outcomes.

[5] Changing Mindsets (regrant), Grammar for Writing (regrant), Lesson Study, Philosophy for Children, Research Learning Communities, Scratch Maths, Affordable Tuition (regrant) and Embedding Formative Assessment.

[6] Philosophy for Children; Thinking, Doing, Talking Science; Chess in Primary Schools.

[7] Children’s University, Pupil Motivation Financial, Pupil Motivation Non-financial. The first of these calculated effect sizes using a gain scores approach. The latter two reported an effect size based on dividing by the population-level standard deviation, but without reporting this population-level standard deviation in the report.

[8] Teacher Effectiveness Enhancement Programme.

Table 1: The interventions

Intervention	Reference	Phase	Original Results	Brief description
Flipped Learning	Rudd et al. (2017)	Prim.	ES=0.09 (*p*=0.62)	Pupils study new maths at home, then consolidate through work in class
Learner Response System*	Wiggins et al. (2017)	Prim.	ES=0.00 (*p*=0.96)	Immediate in-class pupil feedback from teachers via electronic devices
RISE*	Wiggins et al. (2016)	Sec.	ES=0.04 (*p*=0.57)	Teacher 'Research Leads' supporting teachers' use of research
Teacher Observation*	Worth et al. (2017)	Sec.	ES=-0.01 (*p*=0.80)	Teacher PD based on programme of structured peer lesson observations

Notes: "Prim" = Primary. "Sec" = Secondary. All of the primary phase interventions use age 11 math scores as their primary outcome and all of the secondary phase interventions use age 16 math scores as their primary outcome. "PD" = professional development. "ES" = Cohen's D effect size (expressed in pupil-level standard deviations). "*p*" = p value. *for these trials we can only use the first treatment cohort since the relevant time series for the second treatment cohort straddles a discontinuity in the time series.

Answering RQ1: How closely can CITS reproduce the impact estimates from EEF RCTs?

Table 2 shows the results of the four within-study comparisons. The two effect size columns report the impact estimates from the original RCT and new CITS evaluations. The final column shows the difference between the two. The differences in effect sizes in the four WSCs vary from 0.03 to 0.13 in absolute magnitude. It should be noted that these effect sizes are not directly comparable to those usually reported in EEF-funded trials because they are expressed in school-level, rather than pupil-level standard deviations.[9]

For reasons we explain in the technical report, the best test of the CITS assumptions in this setting comes from looking at the mean differences across the four within-study comparisons, rather than each in isolation. As can be seen from the final row of Table 2, the CITS and RCTs diverge by 0.03 school-level standard deviations on average across these four interventions (approximately equivalent to 0.01 pupil-level standard deviations). Should this be considered as a large divergence or not? One way of putting this 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 and

[9] EEF-funded evaluations use pupil-level data and EEF (2018) statistical analysis guidance requires evaluators to use total variance (rather than within-school pupil-level variance or between-school i.e. school-level variance). However, use of school-level standard deviations are more appropriate in the context of CITS for comparability with other such studies in which it is not always possible to convert to pupil-level standard deviations. School-level standard deviations are 1.5-3 times larger than the pupil-level standard deviations (Kraft, 2020), but this is purely a matter of scaling, rather than indicating a meaningfully larger effect.

0.10 pupil-level standard deviations (Kraft, 2020; Evans & Yuan, 2020). Our estimate of bias thus represents 10-14% of the effect size that researchers might expect to find in similar research.

Table 2: Results of the four within-study comparisons

Intervention	**RCT**	**CITS**	**RCT-CITS**
	Effect Size	**Effect Size**	**Effect Size Difference**
Flipped Learning	0.26	0.13	+0.13
Learner Response System*	0.00	-0.03	+0.03
RISE*	0.19	0.16	+0.03
Teacher Observation*	0.15	0.22	-0.07
		Mean:	0.03

Answering RQ2: How does the closeness of the CITS and RCT estimates vary depending on the shape of the trends in the outcome variables in the pre-treatment period?

The pre-treatment trends for each of the four interventions can be found in Figure A1 in the appendix. The two interventions that showed parallel trends in the pre-treatment period (Learner Response System and RISE) showed the smallest absolute differences between the experimental and CITS impact estimates: 0.03 in both cases. The interventions showing non-linear, non-parallel trends (Teacher Observation and Flipped Learning) showed slightly larger differences on average (0.07 and 0.13, respectively). In sum, in line with the theory and intuition behind the CITS design, the more parallel and stable the pre-treatment trends are, the closer that the CITS designs get to reproducing the experimental impact estimates. In the technical report, we provide alternative results using different CITS specifications for each intervention.

Answering RQ3: Does CITS or matching get closer to reproducing the impact estimates from the EEF RCTs?

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. Our CITS impact estimates for the LERS intervention differs from the LERS RCT impact estimate by approximately

0.01 pupil-level standard deviations.[10] The equivalent result in Weidmann & Miratrix is reported graphically (their Figure 1) rather than numerically, but appears very similar in magnitude to our estimate. Our CITS impact estimate for the FLIP intervention differs from the FLIP RCT impact estimate by approximately 0.05 pupil-level standard deviations, which, again, appears very similar to the equivalent result in Weidmann and Miratrix. The results reviewed in this paragraph suggest that CITS and PSM methods, as implemented in these two papers, perform similarly in the context of EEF-funded interventions. It should be noted, however, that this finding is based on only two interventions and should therefore be treated with caution

[10] Table 2 reports this result in school-level standard deviations (0.03). However, we have converted this into pupil-level standard deviations (~0.01) in order to make it comparable to the Weidmann and Miratrix result.

3. Implications for EEF

A clear finding from our project is that the data requirements for conducting CITS designs are, in fact, rarely fulfilled. Recall from Section 2 that our CITS requires a consistently recorded outcome measure across four pre-intervention years and one post-intervention year.[11] If the relevant time series for a given intervention includes any of the kinds of disruption discussed above, then a CITS is unlikely to be a suitable evaluation method. For example, we identified the following events which caused disruptions in the relevant time series:

- in secondary schools, new GCSEs were phased in in 2016 and 2017. This involved changed exam specifications and replacement of the old alphabetic grading system with a non-equivalent numeric grading system;
- in primary schools, in 2012/13, the Key Stage 2 English qualification was substantially reformed, involving changes to the content and format of the exams;
- in primary schools, in 2015, Key Stage 2 exams were again reformed, this time by introducing a new grading system;
- in primary schools, in 2010, there was a widespread boycott of Key Stage 2 examinations;
- in general, the school-level data contain certain school-average performance metrics in some years, but not in others.

For a mix of these reasons, we were only able to conduct CITS evaluations for a minority of the interventions that we initially identified as being potentially re-analysable for this project.

In addition to the above, since we began this project, the COVID-19 pandemic has caused further disruption to the examination system in England, with all exam results in the 2019/20 academic year awarded through teacher assessment.

When should CITS designs (not) be used as the primary evaluation design?

In the four interventions for which we were able to conduct within-study comparisons, CITS came close to reproducing the results of experimental evaluations. It might therefore be considered as a sensible second-best option when experimental evaluation methods are infeasible (Case 1) or unethical (Case 2). However, in either case, it will be necessary:

a) to be able to identify which schools are exposed to the intervention, in which years. This may be possible using survey methods if the intervention can be described with minimal ambiguity. Alternatively, it may be possible to identify this using administrative/management data from an organisation involved in implementing or delivering a particular intervention;

[11] Four pre-intervention time points allows for the modelling of non-linear pre-trends. For more on this, see St. Clair and Cook (2015).

b) for a sufficient number of schools to already be participating in the intervention.

Furthermore, the pre-treatment trends among treated schools strongly influence how persuasive the CITS results will be. The theory behind the CITS design, the empirical results presented here, and findings from other within-study comparison research (St. Clair, Hallberg, & Cook, 2016), suggest that CITS come closer to reproducing the results of experimental evaluations when the pre-treatment trends are parallel and/or easier to model (see the technical report for details of how we define 'easier to model'). Where this is not the case, CITS will be a less suitable evaluation design.

It should be noted that other non-experimental methods, such as careful propensity score matching designs (Weidmann & Miratrix, 2020), constitute an alternative to the CITS method. In the empirical results presented above, these two non-experimental methods did a comparably good job of reproducing the results of experimental evaluations in the two interventions where we were able to make this comparison. Again, it should be noted that this was only possible in two interventions, so this result should be interpreted with caution. 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 research has illustrated, is the need for a consistently recorded time series, which are often not available due to e.g. exam reforms. The relative benefits of the two non-experimental designs are therefore likely to depend on data availability for the specific intervention.

When should CITS designs (not) be used as an additional or complementary evaluation design?

There may be other cases in which an experimental approach to evaluation is initially adopted but it subsequently becomes worthwhile implementing a CITS design:

- There may be problems with recruitment to the experimental evaluation (Case 3). In this scenario, the considerations around whether to conduct a CITS design are very similar to those outlined in the previous section: is the necessary time series available and can a sufficient number of schools already receiving treatment be identified in the population at large? In addition to these, however, researchers should consider including the schools randomly allocated not to receive the intervention in the control group for the CITS design. This is because even where recruitment has been inadequate, the schools randomly allocated to control are likely to be more similar to the treatment group than those in the population in general.
- Alternatively, there may be problems with control group schools in an experimental evaluation accessing the same or similar intervention through other routes (Case 4). In this scenario, a CITS evaluation that uses schools from outside the trial can be run in parallel with the experimental evaluation. The within-study comparison results presented here suggest that the CITS will come close to the results from the experiment had the experimental control

group not accessed the treatment. The CITS results can therefore be used as an empirical check on whether contamination has significantly biased the experimental impact estimates.

- Finally, there may be situations in which schools' random allocation affects their probability of dropping out of the evaluation entirely (Case 5). Having to remove these schools from the analysis due to missing outcome data undermines the similarity of the experimental treatment and control groups. By contrast, removing such schools from the CITS analysis is possible without compromising the internal validity of the design. A complementary CITS can, therefore, be used as an empirical check on whether non-compliance in the treatment group has significantly biased the experimental impact estimates.

References

Anders, J., Brown, C., Ehren, M., Greany, T., Nelson, R., Heal, J., Groot, A., Sanders, M., & Allen, R. (2017) *Evaluation of complex whole-school interventions: Methodological and practical considerations*. Education Endowment Foundation.

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.

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.

Hallberg, K., Wing, C., Wong, V., & Cook, D. (2013). Experimental design for causal inference: Clinical trials and regression discontinuity. *The Oxford Handbook of Quantitative Methods, Volume 1: Foundations*. Oxford University Press.

IfEE [Institute for Effective Education] (2016). *Teacher Effectiveness Enhancement Programme: Evaluation report and executive summary*. Education Endowment Foundation.

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., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis, 37*(3), 314-332.

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

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.

Sims, S., Ander, J., Zieger, L. (2021). The internal validity of the school-level comparative interrupted time series design: evidence from four within-study comparisons. *Paper under review.*

St. Clair, T., & Cook, T. D. (2015). Difference-in-differences methods in public finance. *National Tax Journal*, *68*(2), 319-338.

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.

Weidmann, B., & Miratrix, L. (2020). Lurking inferential monsters: Quantifying selection bias in non-experimental evaluations of school programs. *Journal of Policy Analysis and Management,* https://doi.org/10.1002/pam.22236

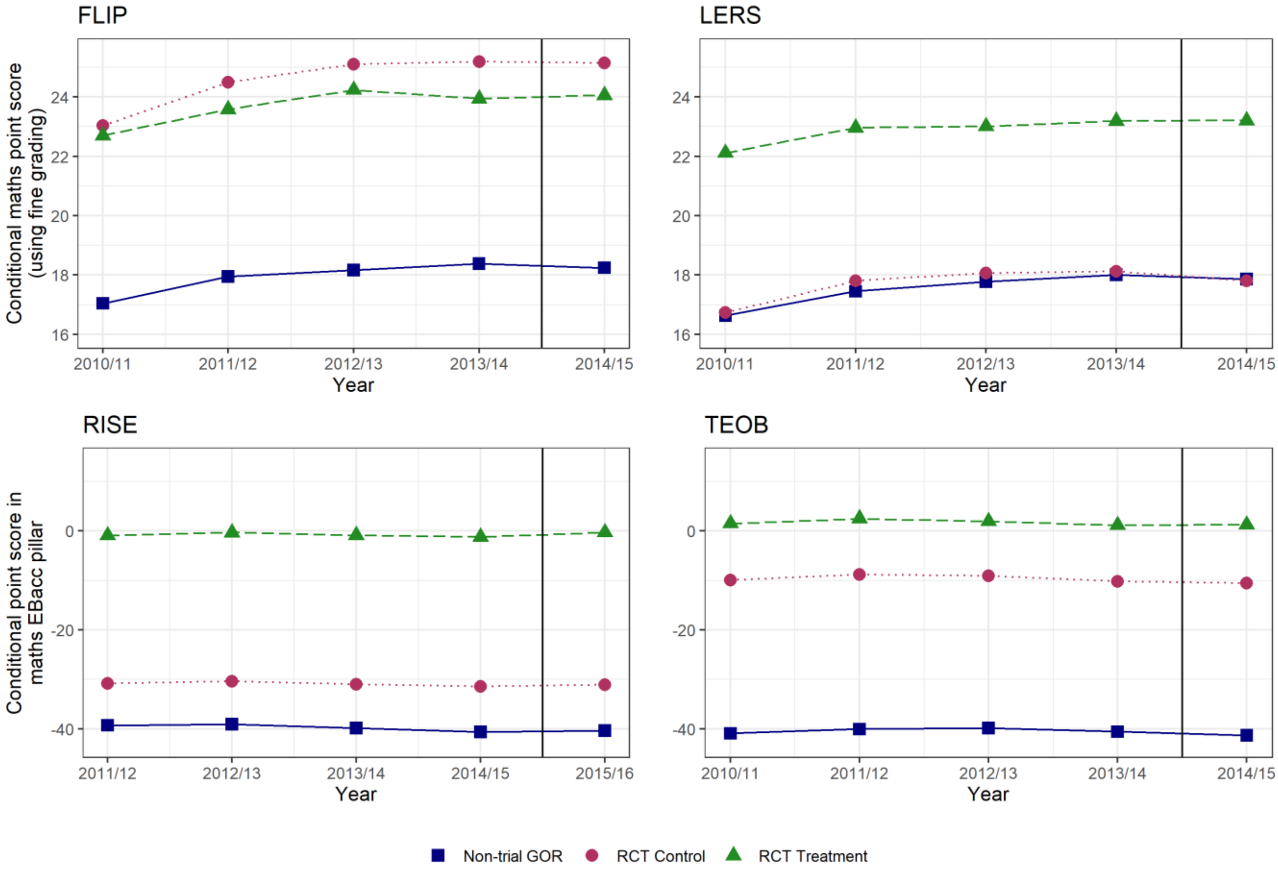
Wiggins, M., Jerrim, J., Tripney, J., Khatwa, M., & Gough, D. (2016). *The RISE Project: Evidence-informed school improvement.* Education Endowment Foundation.

Wiggins, M., Sawtell, M., & Jerrim, J. (2017). *Learner response system: Evaluation report and executive summary*. Education Endowment Foundation.

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 pre-treatment 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.