



Impact evaluation of the South West Yorkshire teaching assistants scale-up campaign

July 2019

Independent researchers:



Luke Sibieta, Barbara Sianesi



The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus (formerly Impetus Trust) and received a founding £125m grant from the Department for Education. Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:



Camilla Nevill
Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP



0207 802 1679



Camilla.Nevill@eefoundation.org.uk



www.educationendowmentfoundation.org.uk



Executive summary

The project

In 2015, the EEF launched a campaign aimed at helping schools to make better use of their teaching assistants (TAs). It involved direct communication to schools, engagement with various organisations and policymakers (in the U.K. and worldwide), as well as a media and publicity campaign to raise awareness of key issues (Appendix C). Specific areas in the U.K. (Lincolnshire and South and West Yorkshire) were targeted for more specific, practical support; this report focuses on the efficacy of the additional support provided to South and West Yorkshire.

The national campaign had one main component:

- Promoting the EEF-authored **evidence-based guidance report on the use of TAs (Sharples, Webster and Blatchford, 2015; henceforth ‘the EEF guidance’)**. By November, for example, 70% of headteachers nationally were aware of it (see Appendix C).

In addition, the campaign in South and West Yorkshire was extended to include two more elements:

- the commissioning and support of advocacy providers; and
- an opportunity for schools to participate in randomised controlled trials (RCTs) of structured, TA-led interventions.

This report evaluates the impact of the second element—advocacy provision—within this wider context. Advocacy providers (APs) ran programmes for schools from September 2015 to July 2016. These included a launch event and core workshops designed to encourage schools to engage with the EEF guidance on making best use of TAs and to implement its recommendations.

Scope of the project

Originally, there was a target to work with about 525 state-funded primary schools across all nine local authorities in South and West Yorkshire (about half of the 1,050 primary schools in the area), focusing on disadvantaged and under-performing schools. Whilst the campaign managed to work with about 480 schools, advocacy providers could only be recruited in seven of the nine local authorities and the campaign as a whole was less successful in focusing on disadvantaged or under-performing schools.

In this report, we estimate the impact of the overall campaign across all nine local authorities, whether or not they participated in advocacy; this assesses the impact on an ‘intention to treat’ basis and includes the effects of recruitment and participation. Our two primary outcomes are Key Stage 2 (KS2) maths and English scores measured in 2016/2017 (one year after the campaign in 2015/2016). The treatment group contains 43,000 pupils in South and West Yorkshire taking Key Stage tests in 2017, or about 350,000 pupils across all year groups.

Given that the treatment group was deliberately non-random, we could not analyse the effects of the treatment using a randomised controlled trial. We instead make use of ‘synthetic control’ methods. These were pioneered to look at the effects of treatments or policies affecting large geographic areas; for example, they have been used to look at the effects of Basque terrorism, smoking prevention in California, and German reunification. This report is the first to make use of such methods for an EEF trial.

Synthetic control methods seek to construct a control group by weighting other aggregate units (here, local authorities) that best approximate the outcome for the treated group (here, South and West Yorkshire) over the pre-treatment phase (all years up to 2014/2015). The difference between the synthetic control and the treatment group in the post-treatment phase then represents the estimated

impact. Further qualitative and mixed-methods work was undertaken by Sheffield Hallam University in the **Formative evaluation of the South and West Yorkshire teaching assistants scale-up campaign report**.

Key conclusions

1. There is evidence that KS2 English test scores showed an improvement of 0.03 standard deviations in South and West Yorkshire as compared with our synthetic control group.
2. There is no evidence of an impact on maths scores.
3. New KS2 tests and curriculum began from 2016. This could be biasing our results, but the size and direction of any bias is uncertain.
4. The cost per school was relatively low, about £130 per school per year (when spread over three years).
5. Synthetic control methods worked well using education administrative data to approximate pre-treatment trends in South and West Yorkshire. They also appear to perform better than other non-experimental estimators like OLS and matching. We recommend they are used more for appropriate large-scale educational interventions.

Findings

The results suggest that KS2 English test scores—when measured in 2016/2017, one year after the campaign took place—were about 0.03 standard deviations higher in South and West Yorkshire than in the synthetic control group. This is outside the normal year-to-year variation in differences across local authorities. About 7% of local authorities saw a larger increase in KS2 English in 2016/2017 relative to differences observed up to 2014/2015.

There is little evidence of any improvement in maths test scores. There is also little evidence to suggest differential effects for pupils with English as an Additional Language ('EAL pupils'), or for those eligible for free school meals ('FSM pupils').

The estimated impact for English is sizeable given that it covers 43,000 pupils taking KS2 tests across the whole of South and West Yorkshire. The results are consistent with an impact of 0.06 across the 40% schools that took up the offer of advocacy. If about half of these schools then changed their TA practices and this affected 30% of pupils at those schools, these results would equate to an effect size of over 0.3 for the effect of changed TA practices.

It seems likely that these results reflect the impact of advocacy provision. Whilst there were also additional RCTs taking place in South and West Yorkshire, these were largely focused on younger ages and were mostly related to numeracy (rather than English); furthermore, many of the trials implemented have failed to find a positive impact. We also perform a range of robustness checks that confirm our overall results and show that synthetic control methods out-perform more standard non-experimental methods (such as OLS regression and matching) in accounting for unobservable differences between the treatment and control.

However, interpretation of the results is compromised by the fact that national KS2 tests also changed radically in 2016, just after the treatment. Whilst synthetic control methods should create a group of LAs most likely to respond to these new tests in the same way as South and West Yorkshire, such reforms are infrequent and could create a bias in our estimates. Unfortunately, the size and direction of any bias is uncertain.

Alongside this report, the EEF will publish the implementation and process evaluation (IPE) of the South and West Yorkshire campaign (Maxwell et al., 2019). The EEF will also publish an IPE of a different approach to scaling-up the TA guidance in Lincolnshire and a 'lessons learned' report that compares the two approaches (also Maxwell et al., 2019). All of the reports are on the website [here](#).

Security of findings



This was a well-conducted synthetic control method that achieved reasonable levels of balance between the treated unit and its synthetic control, closely following pre-intervention trends in the outcomes of interest. However, the concept of inference in synthetic controls remains complex given that only one unit is treated which requires creating ‘placebo treatment effects’ to compare the magnitude of the results in the treated unit with those obtained where no effect would be expected (because they are placebos). Even if this method provides an indication of the ‘statistical significance’ of results, this method remains in its infancy and should be treated with caution. More detail on the padlocks awarded to this evaluation can be found in Appendix B.

Cost

The estimated cost per school involved in advocacy was about £390 or just over £130 per year when spread over three years.

Impact

Summary of impact on primary outcome

Outcome/ group	Effect size	P value	Security rating
KS2 English points	0.03 ¹	0.07	
KS2 maths points	0.015	0.56	

¹ If converted to additional months progress, this result would be just under one month. The impact of this regional scale-up work is not comparable to funded EEF projects which are typically interventions in schools. It has, therefore, not been presented as additional months progress.

Introduction

Background evidence

Teaching assistants (TAs) make up approximately one quarter of the school workforce yet the evidence—summarised in the EEF Teaching and Learning Toolkit and elsewhere—suggests that the current deployment of TAs in English schools is not improving pupil outcomes and can even be detrimental to learning. Blatchford, Russell and Webster (2012) show that pupils who spend more time with TAs make less progress in English and maths, even after controlling for pupil characteristics. TAs also spend a disproportionately large amount of time with disadvantaged pupils so the attainment of the EEF’s target group is particularly vulnerable to harmful practice.

There is now a wide evidence-base suggesting better ways TAs can be deployed. For example, as summarised in the EEF guidance (Sharples et al., 2015), when TAs are used to provide one-to-one or small group support using structured interventions they can generate average gains of between three and four additional months (for example, Switch-on Reading programme, Gorard et al., 2014, and the Catch Up Numeracy programme, NfER, 2014).

Intervention

In summer 2015, the EEF launched a national campaign to improve the ways in which TAs are used. It produced a guidance document setting out ways that schools can maximise the impact of teaching assistants and launched a multi-stranded campaign to increase uptake of this advice and guidance, including:

- national campaign materials, such as sending hard copies of the EEF guidance report to all schools in England, emailing all schools with links to the guidance, various events, and using social media;
- advocacy support provided in a range of ways (such as training events, coaching, and consultancy) to all primary schools in South and West Yorkshire; this was particularly aimed at underperforming schools and focused on enabling schools to act on the guidance; and
- targeted interventions working with schools with high numbers of disadvantaged pupils within South and West Yorkshire; this involved evaluation grants, scoping grants, and scale-up grants provided by the EEF, each of which will be subject to a separate evaluation.

The **implementation and process evaluation** report by Sheffield Hallam University provides further detail of the programme including reach and an in-depth process evaluation to evidence the theory of change.

Evaluation objectives

The focus of this evaluation is the impact of the second strand of the campaign—the provision of advocacy support—exploring the question:

- What impact does offering advocacy to primary schools in South and West Yorkshire have on pupil attainment?

This is research question (c) from the evaluation protocol published on the EEF website.² Research questions (d) to (k) are answered in the **IPE report**; (a) and (b) will be addressed in a separate report.

Assessing the likely size of any effect is not simple. The advocacy campaign aimed to improve overall use of TAs and any effect on pupil attainment will depend on a range of factors including past TA practices, which pupils have most interaction with TAs, uptake of advocacy, uptake of TA-focused interventions, how many schools change their practices, which practices they change, and the effect of those practices. Furthermore, part of the aim of the intervention was to generate system-wide change and collaboration, the likely effect of which is hard to predict in advance. However, we think it is useful to set out some simple calculations as to the potential effect size under a number of assumptions.

We start by assuming that all the effect of the advocacy occurs via use of structured TA-led interventions, the expected effect of which is about 0.2 to 0.3 standard deviations (Sharples et al., 2012). As confirmed in Blatchford et al. (2012), TAs tend to be focused on disadvantaged pupils, though not exclusively. With just under 20% of our treatment group eligible for free school meals, we assume that about 25% of pupils are affected by changed TA practice. If we further assume that 25% of schools shift over to structured TA-led interventions and the effect of the treatment on the treated is 0.3, then we might expect to find an overall effect size of just 0.02 standard deviations.

Although this calculation is very simplistic, it helpfully shows that even under an optimistic scenario where 25% of schools take up structured TA-led interventions and where this affects 25% of pupils, we would only find a small effect size of 0.02. There are a range of factors that could then shift this up or down (how many pupils are affected, other changes in practice, system wide change). However, it is hard to imagine anything more than 0.1, which would equate to an incredibly optimistic scenario where over half of schools change to the more effective practice and this affects over half of pupils at those schools. Any plausible effect is therefore going to be small.

Unfortunately, we could not implement an RCT and use experimental methods to address our main research question as the whole of South and West Yorkshire represented a single treatment group. South and West Yorkshire was deliberately a non-random choice and may be subject to different time trends in pupil performance as compared with other similar sized regions or England as a whole. One cannot use a matched comparison group of schools since an overall aim of the intervention was also to generate system wide change: assessing the likely size of any system wide change requires comparison with other aggregate units, rather than individual schools. However, using local authorities or regions will also be problematic if South and West Yorkshire is subject to particular time trends.

In principle, one requires a counterfactual group with similar characteristics and subject to similar time trends. In order to do this, we make use of synthetic control methods (precise details described below). Synthetic control methods have not been used in previous EEF evaluations and the authors have not used them in previous work either. They have only been used in a relatively small number of education settings; for example, Hinrichs (2012) uses synthetic control methods to study the effect of affirmative action and Belot and Vandenberghe (2014) evaluate the effects of grade repetition. Much of the analysis was already pre-specified in the Statistical Analysis Plan published on the EEF website,³ which we follow closely. However, this work is necessarily exploratory as it is partly aimed at understanding the potential value for using synthetic control methods in other EEF evaluations.

²

https://educationendowmentfoundation.org.uk/public/files/Publications/Campaigns/Campaigns_TA_Campaign_Protocol_FINAL.pdf

³

https://educationendowmentfoundation.org.uk/public/files/Publications/Evaluation_Reports/Campaign_-_SW_Yorkshire_TA_campaign_SAP.pdf

Ethics and trial registration

Ethical review was sought from Sheffield Hallam University as part of an overarching ethical review for the evaluation as a whole. See **main report** for further details.

The evaluation team decided not to register the trial as the evaluation does not constitute a randomised controlled trial.

Data protection

This evaluation is based on anonymised data at the local authority level derived from the National Pupil Database (NPD) over a 15-year horizon. Given the anonymity of the data, it was therefore impossible to seek informed consent, issue privacy/processing notices, or provide any information to parents. To derive and process the NPD data we rely on Schedule 2 Conditions 5d and 6(1) of the Data Protection Act (1998) for the processing of personal data (FSM eligibility) on the basis that the data is necessary for our research and we believe this work to be in the public interest. For the processing of sensitive, personal items (SEN provision, Ethnic group major/minor, EAL), we rely on Paragraph 9 of the Data Protection (Processing of Sensitive Personal Data) Order (2000). This research is conducted for public benefit, with publicly available findings, for the purpose of understanding what works in raising pupil attainment.

Permission to process the data was granted in late 2017. However, processing of the NPD data at the Institute for Fiscal Studies (IFS) is also compliant with the more recent Data Protection Act (2018). The NPD is listed in the IFS' Register of Data Processing Activities and the specific purpose is listed in a Privacy Notice covering all NPD usage. The legal basis for processing the data is registered as 'Legitimate Interests 6(1)f'. The IFS has carried out a Legitimate Interest Assessment. The legal basis for processing Special Category NPD data is Article 9(j). There is a named data controller (Zoe Oldfield) and the IFS is registered with the ICO.

Project team

This evaluation was led by Luke Sibieta (Research Fellow at the Institute for Fiscal Studies and Education Policy Institute), with expert assistance and guidance from Dr Barbara Sianesi (Research Fellow at the Institute for Fiscal Studies).

Methods

Impact evaluation design

Advocacy and training support was offered to all 1,049 primary schools in South and West Yorkshire with the aim of working with approximately 525 schools. The support was provided by a set of intermediaries or advocacy providers who bid to the EEF for the contracts to undertake the advice and training. Support consisted of a range of activities, such as training events, coaching, and consultancy. It was particularly aimed at underperforming schools and focused on enabling schools to act on the guidance.

This advocacy and training was offered from September 2015. We therefore use outcomes from the years 2002/2003 to 2014/2015 as pre-treatment outcomes, whilst 2015/2016 and 2016/2017 represent post-treatment outcomes.

In this case, the treatment unit essentially represents all nine local authorities in South and West Yorkshire. In order to evaluate this treatment, one needs to construct a counterfactual that best captures how outcomes in South and West Yorkshire would have evolved in the absence of the advocacy campaign. This should allow for the fact that the treatment could have aggregate effects and allow for potentially non-linear time trends up to treatment. As noted, one cannot use experimental methods as the treatment area is deliberately non-random; this also presents a challenge for widely-used non-experimental methods like OLS and matching, partly because there is just one overall treatment unit and partly because achieving a good balance on non-linear time trends can be challenging in practice. OLS and matching are also vulnerable to unobservable determinants of outcomes over time.

Table 1: Summary of study design and outcomes

Trial type and number of arms		Synthetic Control (non-experimental)
Treatment area		South and West Yorkshire (Sheffield, Rotherham, Doncaster, Barnsley, Leeds, Wakefield, Calderdale, Kirklees, Bradford)
Control area		Weighted average across other local authorities in England
Pre-treatment period		2002/2003 to 2014/2015
Post-treatment period		2015/2016 to 2016/2017
Primary outcome	variable	Key Stage 2 fine points score in maths and English in 2017
	measure (instrument, scale)	Both standardised at the national level
Secondary outcome(s)	variable(s)	% of pupils meeting expected levels in Key Stage 2 maths and English in 2017, Key Stage 2 fine points score in maths and English in 2016
	measure(s) (instrument, scale)	% of pupils meeting expected levels measured in raw terms, Key Stage 2 fine points scores standardised at the national level

We therefore make use of synthetic control methods. The synthetic control method was first proposed by Abadie et al. (2003) in the context of estimating the effects of Basque Terrorism, further applied to assess the impact of a tobacco control law in California (Abadie et al., 2010) and of German reunification

(Abadie et al., 2014). They have also been used to analyse the value of financial connections (Acemoglu et al., 2016), pay-for-performance policies in healthcare (Krief et al., 2016) and a small number of educational settings (Hinrichs, 2012; Belot and Vandenberghe, 2014).

From a ‘donor pool’ of aggregate units (local authorities in our setting) not affected by the policy, a synthetic control group is constructed as the weighted average of those units that best resemble the treatment unit in terms of pre-treatment characteristics and outcomes for as long a time period as possible. The hope is that by matching closely on pre-treatment trends, the synthetic control unit would provide a good counterfactual of what would have happened in the absence of the intervention. In particular, by seeking to replicate pre-treatment trends, the method seeks to ensure there is a match between all factors that determine both the level of outcomes and their trend over time.

The context of the advocacy component is well suited to this kind of analysis as we are able to ensure comparability of pre-treatment outcomes stretching as far back as 2002. In particular, we will take local authorities as our potential donor pool and find the set of weights across local authorities that best allows us to match pre-treatment trends in primary outcomes (average Key Stage 2 fine points scores in maths and English) and other characteristics in South and West Yorkshire.⁴

We take local authorities as our main unit of analysis as the treatment was targeted at a group of nine local authorities. A school-level analysis would not allow for comparing any potential system-wide effects and changes. Using regions would have been another option, though this would have led to a very small number of observations as the treatment takes place at a sub-regional level. We effectively treat South and West Yorkshire as a single local authority.⁵

To take a simple example, the weights could take the form of 0.5 for Manchester, 0.25 for Liverpool, and 0.25 for York. These would then be used to re-create a synthetic comparison group for South and West Yorkshire between 2002 and 2014 that would have time trends in the pre-treatment outcomes that are as similar as possible with the actual pre-treatment outcomes of South and West Yorkshire. We would then apply these weights to the post-treatment outcomes across local authorities and compare this with the post-treatment outcomes in South and West Yorkshire. This will give an estimate of the effect of the advocacy campaign.

The advocacy and training took place at the same time as a national campaign led by the EEF to improve the way in which TAs are used. This campaign involved a large number of activities (see Appendix C for a list of activities, which was provided by the EEF). By autumn 2015, about 70% of headteachers were aware of the guidance and about 40% had read it. The synthetic control group will thus be one that is potentially more informed about better ways to use TAs than would otherwise have been the case.

A separate experiment was conducted around the national campaign to test different ways of emailing schools about the campaign. This will report at a later date. However, it is worth stating that the experimental variations to the national campaign (different forms of emails and incentives to engage with the information) had relatively low impact on levels of engagement with the TA campaign and are thus unlikely to be any source of bias in our impact analysis of advocacy and training on pupil attainment in South and West Yorkshire. To make this point directly, we also show the average outcomes in primary

⁴ Weights are constant over time, constrained to be non-negative, and add up to one. The argument to constrain weights to non-negative values that add up to one is to limit extrapolation between the observed cases and the synthetic control generated (Abadie et al., 2010).

⁵ This is similar to the approach taken in Krief et al. (2016) who use hospital-level data to recreate a synthetic control group for the North-West English region. Acemoglu et al. (2016) use an alternative approach for cases with multiple treated units. They create synthetic control groups for each treated unit and then aggregate the treatment effect over the multiple units. We sought to replicate this approach as part of our robustness checks, but the prediction errors for individual local authorities within South and West Yorkshire were sufficiently high so as to be considered unreliable. Indeed, five of out of the nine local authorities would be dropped by our inference procedure as being unreliable.

schools outside South and West Yorkshire that did, and did not, receive incentive to engage with the national campaign materials.

Our main focus is the impact of advocacy and training. However, additional interventions were also targeted at schools in South and West Yorkshire. Indeed, a stated aim of the advocacy was to encourage schools to make greater use of structured interventions and become involved in trials.

Our main analysis will therefore combine the effects of advocacy with any effects of additional interventions targeted at South and West Yorkshire over the period. We will not be able to separate out the pure effect of offering advocacy and guidance as opposed to any direct effect of additional interventions.

We will undertake a small element of analysis where we compare the outcomes of schools in South and West Yorkshire that received advocacy or received targeted interventions with those that did not. This is not robust. Any difference in outcomes could result from the effect of additional interventions or advocacy, or it could reflect the willingness of different schools to participate in trials and advocacy which might be correlated with better outcomes.

There were three main changes to the protocol and statistical analysis plan along the course of the trial:

- The date of the primary outcome of interest was moved from one year after the start of the advocacy campaign (summer 2016) to two years after (summer 2017). This decision was made in 2015 before any outcomes were realised. It was felt one year was not long enough for the advocacy campaign to have any effect on pupil attainment. TA deployment decisions are generally made with a relatively long lag. It is certainly possible that the effect could grow after two years as well, but it was felt that two years represented a suitable lag to conduct an evaluation.
- The protocol and statistical analysis plan both proposed creating the synthetic control groups before outcome data for summer 2016 and summer 2017 was available. This was not possible due to delays in accessing National Pupil Database data. However, we still closely follow the procedure set out in the statistical analysis plan for selecting the synthetic control group. Outcomes data was ignored during the process of creating the control groups.
- The statistical analysis plan and previous versions of the protocol make reference to creating different versions of the synthetic control group based on schools that did, and did not, receive incentives to engage with the national campaign. However, given the very low impact of the incentives on information engagement, the impact on attainment is likely to be extremely small. We therefore do not believe that the work required to undertake the separate control groups would be proportionate to the value it would generate. Comparing post-test outcomes across the two sets of schools randomised into incentives groups will already generate an unbiased estimate of the impact of the incentives, and hence allow us to gauge the potential for this to affect our interpretation of the control condition.

Participant selection and sample sizes

The main comparison is between a treatment group consisting of all primary schools in South and West Yorkshire and a comparison group consisting of primary schools outside of South and West Yorkshire. The average outcome for South and West Yorkshire represents the (pupil-weighted) average across the nine local authorities. The synthetic control outcome represents the average across other local authorities in England,⁶ weighted by the synthetic control weights (calculation of which is detailed below). We will exclude the following groups of local authorities for reasons as specified:

- Isles of Scilly and City of London; and

⁶ These averages across donor local authorities are not weighted by pupil numbers as the process does not seek to estimate a representative a national average, but the best match for South and West Yorkshire

- Inner London.

The Isles of Scilly and City of London are excluded as they are both very small and unusual relative to the rest of England. Inner London is excluded due to very different time trends, with inner London Key Stage 2 performance growing by over 0.2 standard deviations relative to the rest of England for those taking the tests between 2002 and 2013. Using the ONS statistical definition of 'inner London',⁷ this leaves us with 128 local authorities out of an initial donor pool of 143 before these exclusions.

Since the treatments were strongly focused on schools within South and West Yorkshire, we judge that spillover effects are likely to be small and include neighbouring local authorities in the donor pool. This is important as neighbouring local authorities are likely to be more similar to South and West Yorkshire than those further away. For robustness checks, we re-estimate the results with neighbouring local authorities removed from the potential donor pool.

A number of alternative samples will be constructed to analyse the likely impact of advocacy and guidance over and above additional treatments occurring as part of the wider TA campaign.

First, a large number of schools outside of South and West Yorkshire received micro-incentives to engage with the information in the TA campaign. To compare this with the effect of the advocacy and guidance, we will compare the average outcome of schools in South and West Yorkshire that did and did not receive such incentives.

Second, some schools in South and West Yorkshire also received additional TA-led interventions. To benchmark the likely impact of these additional interventions on the average outcome for South and West Yorkshire, we will describe the outcomes of schools in South and West Yorkshire that did, and did not, report receiving these additional interventions (list supplied by Sheffield Hallam University).

Outcome measures

All our outcomes are derived from national Key Stage 2 assessments. These are nationally recognised and externally valid measures of literacy and numeracy for primary school pupils. They are also available for practically all pupils in state-funded primary schools. We consider performance in both English and maths as primary outcomes as the intervention aimed to improve overall TA usage rather than pupil performance in a particular subject. Both English and maths outcomes represent the main outcomes for considering overall primary school performance.

Our two primary outcomes are:

- KS2 English fine points score 2016/2017; and
- KS2 maths fine points score 2016/2017.

Both outcomes are standardised amongst all pupils in state-funded primary schools taking KS2 tests in 2016/2017 and then averaged at the local-authority level. The primary outcome for the treatment group is the (pupil-weighted) average across all nine local authorities in South and West Yorkshire and the control outcome is the average across the synthetic control group (local authorities weighted by the final synthetic control weights).

We also analyse secondary outcomes derived from KS2 test scores:

- proportion achieving at least the expected level in KS2 English in 2016/2017;
- proportion achieving at least the expected level in KS2 maths in 2016/2017;
- KS2 English Fine Points Score 2015/2016; and
- KS2 Maths Fine Points Score 2015/2016.

⁷ <https://www.ons.gov.uk/methodology/geography/ukgeographies/administrativegeography/england>

The first two secondary outcomes seek to test whether there is a greater impact lower down the KS2 distribution as it is likely to be struggling pupils who are the biggest focus of TAs. The second set of secondary outcomes test for early impact one year, rather than two years, after the start of the intervention.

Importantly, all primary and secondary outcomes have different synthetic control weights. This is because the ideal set of weights can and does differ across outcomes.

There were substantial reforms to Key Stage 2 assessments taking place from 2016 onwards, the first post-treatment period. These included reforms to the curriculum, assessments, and the way tests are scored. This timing is unfortunate from the perspective of this evaluation. The synthetic control group was chosen to best approximate the treatment group over the whole course of the pre-treatment period. We might therefore hypothesise that it will respond to the new tests in the same way as the treatment. However, it is impossible to predict with any degree of certainty the plausible effects of changes to the tests. The effect could be small or large, and any bias could be positive or negative.

By way of example, the inner London advantage declined by 0.08 standard deviations when KS2 English tests were changed in 2013 (Blanden et al., 2015). This could have been driven by the effect of the test or a genuinely large improvement (inner London was already on an improving trajectory). This does not indicate the likely size or direction of any bias, but does show that effects of changes in test can be large and apply systematically across large geographic areas. Unfortunately, this inevitably means that it will be hard to be certain that any observed effect is the result of the treatment and not the changing test, even though the control group has been chosen in a way that might be likely to respond in the same way as the treatment group.

Figure 1, below, gives a sense of the extent to which the scoring of Key Stage 2 assessments in maths and English changed after the treatment started in 2015. This shows the average (raw) Key Stage 2 fine points score across maths and English over time for South and West Yorkshire (orange line), the rest of England (dashed black line) and the donor pool (the rest of England excluding inner London and the Isles of Scilly⁸). We see that average scores in South and West Yorkshire are generally below that for the rest of England prior to the TA campaign in 2015. We also see that the sample exclusions actually make little difference to the average scores across the rest of England.

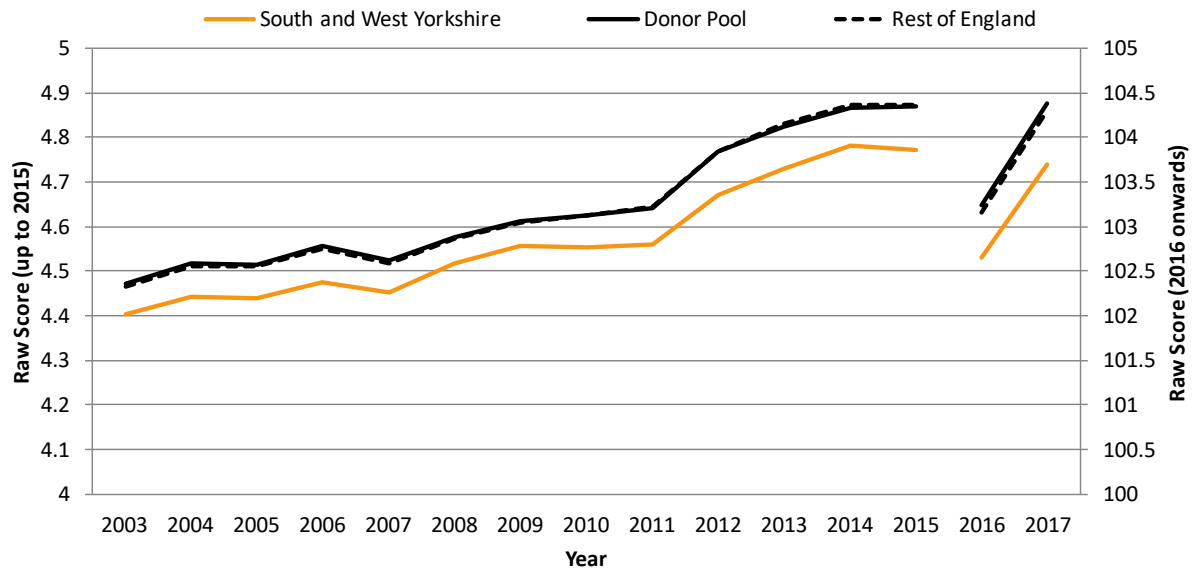
After the reforms to Key Stage 2 assessments for summer 2016, however, we see that the average scores can only be shown by displaying them on a different axis, such is the change in the scoring system. After 2015, the raw gap between South and West Yorkshire and the rest of England closes completely for English, but remains stark for maths. It is not clear if this is driven by the new tests or factors associated with the treatment.

For our analysis, we standardise continuous outcomes at the national level to ensure consistency over time. However, to interpret our estimates as causal, we would still need to maintain the strong assumption that they would react in a largely identical manner.

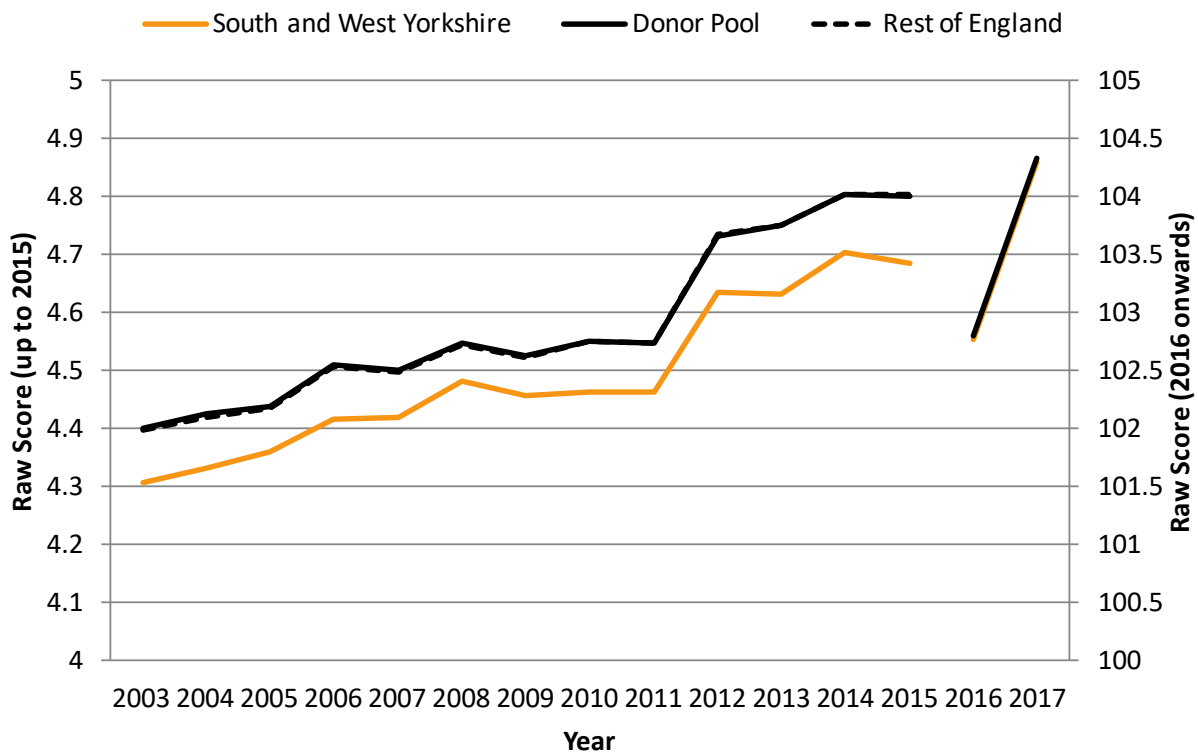
⁸ We drop reference to excluding City of London here as this is included within inner London,

Figure 1: Raw Key Stage 2 maths and English fine points score 2002 to 2017

a) Maths



b) English



Note: 'Rest of England' represents all local authorities outside South and West Yorkshire; 'Donor Pool' excludes inner London and the Isles of Scilly.

Statistical analysis

In order to undertake the synthetic control approach, one needs to construct time-constant weights for each local authority in the donor pool that best approximate how outcomes for the treatment group would have evolved in the absence of the treatment.

To do so, we used pupil-level data from the National Pupil Database (Key Stage 2 results data linked to pupil census data) from 2002/2003 to 2016/2017. The period from 2002/2003 to 2014/2015 is classified as the pre-treatment phase and 2015/2016 to 2016/2017 as the intervention period, with 2016/2017 representing our primary outcome period of interest.

Outcomes for the treatment group represent the average (pupil-weighted) outcome across all nine local authorities in South and West Yorkshire.

The control group outcome is then a weighted outcome from the donor pool of all other local authorities in England (excluding inner London and the Isles of Scilly). The weights are calculated at the local authority level to best approximate the evolution of the lagged outcomes and characteristics of South and West Yorkshire over the pre-intervention phase. In particular, our main measure to assess the quality of the match is the root mean squared prediction error (RMSPE)⁹ between South and West Yorkshire and the synthetic controls in the lagged outcome over the pre-treatment period. The RMSPE is effectively a measure of how close our predictions come to capturing the true changes in the outcomes before the treatment.

The resultant weights can, and do, differ across our two main primary outcomes and our two secondary threshold outcomes but are identical for the 2015/2016 equivalents of our primary outcomes (that is, weights are the same for KS2 English fine points scores in 2015/2016 and 2016/2017). The fact that the weights are different across different outcomes is desirable as it is likely that different underlying factors drive different outcomes and, as a result, different sets of local authorities are likely to serve as suitable controls for different outcomes. One is seeking to find the best synthetic control for each specific outcome.

Although our main assessment of the match quality is based on the RMSPE in lagged outcomes, it is often helpful to seek a balance on wider characteristics that predict levels and trends in outcomes. We will select these from the list of control variables specified below, mainly those that are readily available in our main data source (the NPD). The lagged outcomes are the most important ones to match so we include all years of data here. We then also include a range of characteristics that are likely to affect attainment trends over time, including deprivation, ethnicity, and language. There is a large range of empirical evidence suggesting that deprivation, ethnicity, and language are all strong predictors of pupil attainment (Blanden et al., 2015; Wilson, et al., 2011). TA ratios and academy status are hypothesised to impact on pupil attainment, as well as the potential effect of the treatment, based on the fact that (a) academies have more freedom to make resource choices and (b) response to guidance on making use of TAs is likely to depend on the number of TAs at the school.

Whilst we include the proportion of pupils with a statement of special educational needs, we do not include the other types of special educational needs as this is subject to more discretion and likely to be affected by the way in which TAs are, and can be, used. Data on funding is not included because of missing data for many schools after they became academies between 2010 and 2014.

The full set of predictor variables used were:

- KS2 English/reading fine points score for 2002/2003 to 2014/2015
- KS2 maths fine points score for 2002/2003 to 2014/2015;
- %FSM average over period ;
- %White-British average over period;

⁹ The prediction error is the difference between the actual value and predicted value. This is then squared in order to weight big deviations more highly than small deviations, and to ensure negative and positive errors are treated in the same way. We take the mean average of these squared errors. Then we take the square root of this mean to ensure the final value can be interpreted as the difference between the actual and predicted value.

- %Black average over period;
- %Asian average over period;
- %EAL average over period;
- % of pupils with Statement of Special Educational Needs average over period;
- TA ratio average for 2010/2011 to 2014/2015; and
- % pupils in academies 2010/2011 to 2014/2015.

Performing synthetic control analysis effectively involves choosing two sets of weights: (1) the weights applied to each donor unit (here, local authorities) and (2), the weight applied to each predictor variable (list above). The first set of weights can be found as the solution to an optimisation problem, but the latter is effectively a subjective decision. We follow a similar process to that laid out in Abadie (2003) for choosing the second set of weights. In particular, each set of variable weights (call this V) is associated with a set of weights at the donor unit level (call this W). We will choose the V that achieves the W which is associated with the lowest RMSPE on a subset of the predictor variables.¹⁰ This procedure is more efficient and performs better than simply choosing the weights that achieve the lowest RMSPE on the subset of variables (Abadie, 2003).

The initial subset of predictor variables we use for choosing V is:

- lagged outcomes for 2002/2003, 2007/2008, 2012/2013, 2014/2015;¹¹
- %FSM average over period;
- %White-British average over period;
- %Black average over period;
- %Asian average over period;
- %EAL average over period;
- % of pupils with Statement of Special Educational Needs average over period;
- TA ratio average for 2010/2011 to 2014/2015; and
- % pupils in Academies 2010/2011 to 2014/2015.

Quality of the synthetic control is analysed by comparing the level of the match assessment variables listed above across South and West Yorkshire and the weighted synthetic control group. A key overall metric used to assess overall balance will be the RMSPE. As additional comparators, we also include the (pupil-weighted) average across local authorities with non-zero synthetic control weights and the national average. These will provide an indication of the effect of the synthetic control approach in achieving balance between the treatment and synthetic control groups.

Table 2 shows how we select our preferred donor pool and V specification by calculating the RMSPE across a range of potential different specifications for V and size of the donor pool.

Checking sensitivity of the RMSPE to the size of the donor pool is important as the number of donor units in previous applications of this method has only ever varied between 12 and 38 (as compared with 128 in our case). Along the rows, we therefore vary the size of the donor pool based on the difference between the average pre-treatment outcome in South and West Yorkshire and local authorities in the original donor pool. We start with the full donor pool before then focusing on local authorities where the

¹⁰ In principle, there is an infinite set of V weights. We therefore employ a normalisation such that the Euclidean norm of the weights is equal to one.

¹¹ As stated in Abadie (2003), only including a subset of lagged outcomes as predictors can often be more efficient.

average pre-treatment outcome is within 0.25 standard deviations of the average for South and West Yorkshire. This is then narrowed down in incremental steps down to 0.1 standard deviations.

We vary the specification for V along the columns by progressively dropping groups of variables from the initial specification for V listed above. This is done to show how dropping variables (and thus gaining efficiency in the estimation process) affects the potential bias. Column (1) used the full set of predictors listed above. Column (2) drops academy status as a predictor given the lack of empirical evidence suggesting that academy status has a large impact on pupil attainment, Column (3) drops intervening years for lagged outcomes in the set of variables for V. Column (4) uses the averaged lagged outcome over the entire pre-treatment phases instead of the first and last year. Column (5) then drops the proportion of children with Asian or Black ethnicity, instead relying on proportion with White-British ethnicity to capture the effects of ethnicity

Each cell in Table 2 then shows the RMSPE associated with the synthetic control weights for that specification for V and that donor pool. As can be seen, the RMSPE is very low across the board, generally less than 0.02 standard deviations. In general, reducing the number of variables in V increases the prediction errors, particularly when academy status is removed for maths and when intervening years of lagged outcomes are dropped for English. Reducing the size of the potential donor pool can both increase and decrease the prediction errors. The reductions come from having a more similar group of LAs to begin with, and the increases from having a smaller set of LAs to choose from.

Table 2: Selection of synthetic control group by Root Mean Squared Prediction Error

a) KS2 (standardised) mathematics score

LAs where average pre-treatment outcomes are within given range relative to treatment	Number	Specification				
		(1)	(2)	(3)	(4)	(5)
Whole donor pool	127	0.0072	0.0139	0.0142	0.0162	0.0149
< 0.25 SDs	117	0.0055	0.0133	0.0143	0.0197	0.0144
< 0.2 SDs	112	0.0097	0.0143	0.0143	0.0474	0.0141
< 0.15 SDs	93	0.0095	0.0137	0.0151	0.0170	0.0143
< 0.1 SDs	70	0.0089	0.0101	0.0149	0.0431	0.0138

b) KS2 (standardised) English score

LAs where average pre-treatment outcomes are within given range relative to treatment	Number	Specification				
		(1)	(2)	(3)	(4)	(5)
Whole donor pool	127	0.0068	0.0086	0.0079	0.0193	0.0177
< 0.25 SDs	114	0.0082	0.0077	0.0091	0.0196	0.0244
< 0.2 SDs	103	0.0083	0.0083	0.0070	0.0245	0.0271
< 0.15 SDs	79	0.0054	0.0087	0.0084	0.0336	0.0178
< 0.1 SDs	53	0.0073	0.0066	0.0120	0.0257	0.0221

Specification 1: %FSM (avg); %White-British (avg); %Black (avg); %Asian (avg); %EAL (avg); %SEN-Statement (avg); Pupil:TA Ratio (2010/2011 to 2014/2015); %Academy (2010/2011 to 2014/2015); KS2 (std) maths and English Scores (2002/2003, 2007/2008, 2012/2013, 2014/2015);

Specification 2: Specification 1 minus %Academy;

Specification 3: Specification 2 minus KS2 scores for 2007/2008 and 2012/2013;

Specification 4: Specification 3, all KS2 scores averaged over time;
Specification 5: Specification 3, excluding %Asian and %Black.

For maths, the case that leads to the lowest RMSPE is where we use specification (1) and restrict to local authorities where the pre-treatment outcome is within 0.25 standard deviations of the outcome for South and West Yorkshire. For English, the lowest RMSPE comes from using specification (1) and where we restrict the donor pool to within 0.15 standard deviations of South and West Yorkshire.

The SAP specified that we should use the specification and donor pool that delivers the lowest RMSPE. However, we did not specify how this choice would proceed in the event that the lowest RMSPE was different for the two primary outcomes. Rather than having a different donor pool across the two primary outcomes, we therefore decided to use the ideal specification and donor pool for maths across both outcomes for the primary analysis. In practice, this hardly changes the RMSPE for English. It also allows the optimisation procedure to judge whether to use local authorities where the pre-treatment outcomes are a long way from South and West Yorkshire, as opposed to imposing the restriction to be within 0.15 standard deviations. We also perform robustness analysis where we estimate the impact on primary outcomes using the ideal donor pool for English too. The results are almost completely unchanged.

In Tables D1 in Appendix D, we show the local authorities and weights that form the synthetic controls for each primary outcome. There are about nine or ten local authorities receiving positive weight. The list of local authorities is different across the primary outcomes. For maths, Medway, and Poole are weighted highly with weights of around one third. Coventry, Knowsley, and North East Lincolnshire receive a weight close to 10%, with four other local authorities then receiving quite small weights of under 5%. For English, the weights are more evenly spread, with Rochdale, Blackpool, Nottingham, and North East Lincolnshire all seeing a weight between 12% and 22%. Middlesbrough, Walsall, and Peterborough receive weights between 5% and 10%. Three other local authorities then receive smaller weights below 5%.

In Tables D2, we show the equivalent weights for the secondary threshold outcomes. The weights and the set of local authorities are different to the primary outcomes, which is unsurprising as a different set of local authorities may be required to approximate these different secondary outcomes. There is, however, some similarity with Medway, North East Lincolnshire, Peterborough, and North Yorkshire featuring prominently in local authorities receiving positive weight in both primary and secondary outcome analysis.

ITT analysis

The primary Intention to Treat (ITT) analysis then comprises the raw difference between the standardised primary aggregate outcomes for South and West Yorkshire and the synthetic control group (using the weights delivered by our preferred specification and donor pool decided above). These figures do not account for any covariates (beyond those used to calculate the synthetic control weights). Analysis is conducted using Stata 14, with results presented graphically and numbers quoted in the text. All syntax will be submitted to the EEF archive.

This difference will capture the combined effect of *offering* advocacy and training in South and West Yorkshire. This will include the effects of recruitment and participation in advocacy provision. It will also include any additional effect of targeted interventions. The comparison group is likely to be more informed about effective uses of teaching assistants than would be the case in earlier years as a result of the National Campaign.

Sensitivity analyses

We compare our primary ITT analysis with a number of different approaches to both check the sensitivity of our results to different assumptions and show the effect of the synthetic control approach. We compare against raw outcomes to show the effect of controlling for differences across the treatment

and control groups. We then compare the results against three different specifications for choosing the synthetic controls and check the sensitivity of results to excluding neighbouring local authorities (in case there are spillover effects). Our leave-one-out estimator shows the extent to which our results are dependent on comparing against particular local authorities.

We also compare against a range of widely-used non-experimental estimators to show the effects of the synthetic control method and associated assumptions. The synthetic control method seeks to find a set of weights that achieves the best fit in lagged outcomes over time prior to the treatment. First, we compare this against OLS with covariates. OLS effectively seeks to construct a counterfactual outcome by accounting for the estimated linear effects of other observable covariates. Second, we use propensity score kernel matching. This seeks to construct a set of weights that achieves the best match in a balancing score (the propensity score, which is estimated as a function of a set of observable covariates). In so doing, it allows for non-linear effects of covariates. Third, we use Mahalanobis distance matching, which seeks to find a set of weights that achieves the best balanced on a set of observable covariates. It is similar to propensity score matching in that it allows for non-linear effects of covariates, but directly targets these rather than indirectly through the propensity score (and it can therefore be more challenging to achieve balance in practice).

The main difference between the synthetic control method and these other non-experimental methods is how they account for unobservable determinants of the outcomes over time. The synthetic control method implicitly accounts for these by seeking to match actual trends in lagged outcomes. OLS and matching methods will be biased when there are large unobservable differences between the treatment and control group. One of the main reasons we chose the synthetic control approach as our preferred method is that we believe there are likely to be notable unobservable determinants of outcomes over time. All the approaches, however, will be biased if the effect of unobservables is changing over time differentially across treated and control units, for example, differential responses to the new Key Stage 2 tests.

Taking these together, we compare the results against the following alternative ways of calculating the counterfactual outcome:

- national average, that is, raw baseline comparison;
- pupil-weighted average across local authorities with positive synthetic control weights;
- three alternative synthetic control groups with the next lowest levels of RMSPE;
- leave-one-out SC estimate (the idea is to iteratively exclude one LA which had received positive weight; this loses some goodness of fit but can help assess whether the results hinge on a specific LA being included);
- OLS with all covariates listed above and a dummy variable for South and West Yorkshire (additional to Statistical Analysis Plan);
- Kernel Propensity Score Matching; and
- Mahalanobis distance matching.

OLS and the two matching estimators are undertaken at both local authority and school level with the same predictor variables as in our preferred specification.¹² The local authority level analysis is more in keeping with the synthetic control approach in seeking to find a counter-factual at the local authority level and allowing for aggregate effects, but reduces the number of observations in the control group. A school-level approach does not allow for aggregate effects but increases the effective sample size of the control group. Despite using the same sets of covariates, these approaches can give different

¹² In both estimators, we employ an Epanechnikov kernel with bandwidth 0.03 and maintain common support. We experimented with various bandwidth values. The value of 0.03 gave the best overall match and the estimated effect was largely unchanged in any case.

results given the different levels of aggregation.¹³ As already noted, these are all likely to be less reliable than the synthetic control approach in this setting but will be useful to show the extra value of the synthetic control approach in practice.

As part of our robustness checks, we also imply a further placebo test, where we instead assume the treatment began one year earlier and look at outcomes at the end of 2015 (not mentioned in the Statistical Analysis Plan). Under the assumption that each method for constructing the counterfactual outcome is unbiased, this should produce an estimated treatment effect of zero. They are likely to produce different estimates when there are significant unobservable determinants of the outcomes over time.

In addition to advocacy and guidance, the TA campaign also involved a national campaign (with some schools receiving micro-incentives to engage with information) and additional TA-led interventions. We will therefore also describe the outcomes in schools that did, and did not, receive experimental incentives designed to increase information engagement.

We also show the outcomes for schools in South and West Yorkshire offered advocacy and training split by whether or not they received additional interventions, and whether they actually took up the offer of advocacy. We undertake these comparisons just before the treatment in 2015 and in 2017 when our primary outcomes are recorded. We also control for covariates that might drive differences in outcomes even in the absence of treatment effects¹⁴ (we use the same set of predictor variables used in our preferred primary specification). Even so, this is not a robust approach as there could be unobserved factors driving participation in trials and advocacy, and outcomes. However, it is indicative as to the extent to which additional interventions could be driving some of the estimated difference in outcomes between South and West Yorkshire and the synthetic control group.

Secondary ITT analyses

The secondary ITT analysis comprises the raw difference between the standardised secondary aggregate outcomes for South and West Yorkshire and the synthetic control group. These figures are calculated as described in the outcomes section above and do not account for any covariates (beyond using them to construct the weights). The weights used are shown in Tables D2.

Subgroup analyses

We also perform subgroup analysis for pupils eligible for free school meals and those with English as an Additional Language (both key targets groups of TAs). This proceeds in an identical fashion to the overall analysis except that the outcomes for each local authority are defined by those of the subgroup in question and the covariates defining the subgroup are dropped each time (that is, we do not include EAL as a covariate when we are examining the outcomes of pupils with EAL).

Effect size calculation and inference

Continuous outcomes are standardised prior to analysis and raw differences will thus be interpreted in effect size terms. The secondary threshold outcomes (the proportion achieving the expected levels in maths and English) are not standardised as these outcomes are already readily understandable by schools. They are instead presented in their raw form.

¹³ This differs from the SAP in two ways. We have chosen to use the same set of predictor variables across specifications, rather than use a different set for the matching specifications. This increases the chance that any observed difference in impact estimates is driven by different methodologies rather than using different covariates. Second, we impose a balanced panel assumption so that the analysis only draws on schools we can observe and link across all years. This is necessary given that the specification includes a range of lagged outcome variables.

¹⁴ We use simple OLS analysis for this comparison, with separate regressions for 2015 and 2017 and dummies for whether schools participated in advocacy or trials.

Although there is no uncertainty in terms of the values of the aggregate units (in this case, average attainment by local authority), there is uncertainty as to whether the synthetic control group best approximates the counterfactual for the treatment group. Inference cannot therefore be conducted in standard ways. Instead, Abadie et al. (2010) propose a placebo-test method whereby each aggregate unit within the 'donor pool' is in turn assigned as the treatment group in order to create a distribution of treatment effects. We can then compare the estimated treatment effect of the actual intervention with the distribution of placebo treatment effects to ascertain whether the former is large and calculate a p-value.

As part of the approach proposed by Abadie et al. (2010), we undertake the inference procedure in the following steps:

- show the synthetic control estimates for all other donor LAs (with South and West Yorkshire back in donor pool);
- exclude placebo LAs where pre-RMSPE is five times larger than RMSPE in South and West Yorkshire to deal with LAs that are unlikely to have suitable synthetic control; any threshold choice here is arbitrary and we use five as our threshold to match Abadie et al. (2010); and
- compare both the levels and ratio of the post-RMSPE to pre-RMSPE in South and West Yorkshire with that in other LAs.

Costs

The advocacy and training campaign was targeted at the whole of South and West Yorkshire rather than individual schools. The overall campaign was run by the EEF, with provision of advocacy and training run by different providers in each local authority. Repeating anything similar would also require a coordinated approach. We therefore provide costs from the perspective of a policymaker deciding whether or not to conduct a campaign, rather than an individual school deciding whether or not to adopt an intervention.

We do, however, approach costs in a very similar way to EEF guidance on providing costs. We focus on direct costs, excluding staff time, and present costs per school and pupil involved.

Advocacy providers were sent a very short survey with a proforma spreadsheet in May 2018 (available on request). This asked for a breakdown of costs for workshops/events, additional support offered to schools, other training, management/administration, and 'other' costs. Providers were asked to report actual costs rather than funding provided by the EEF. Unfortunately, only four providers supplied such data, despite various reminders. Data was also not provided consistently across the headings.

We therefore report costs for the three providers that did supply data and focus on total costs. Costs are reported per school signed up with each advocacy provider and per pupil at these schools.

Rather than use actual names of providers, advocacy providers are labelled in an identical fashion to the **IPE report**, for example, 'AP1', 'AP2'.

Timeline

Table 3: Timeline

Date	Activity
September 2015–July 2016	Advocacy and guidance campaign in South and West Yorkshire
Summer 2016	Key Stage 2 assessments (first year after treatment)
Summer 2017	Key Stage 2 assessments (second year after treatment)
September 2017	NPD application
May 2018	Receipt of NPD data
May–June 2018	Cost survey
June–August 2018	Analysis
End August 2018	Submission of draft report
Early 2019	Publication of final report

Impact evaluation

Participants

Here we describe the process for selecting local authorities into different groups and how rules for selecting pupils for analysis affects the sample sizes each year.

The unit of analysis in this impact evaluation is at the local authority level, of which there are 152 over this period. The treatment group was predetermined as the nine local authorities in South and West Yorkshire (Sheffield, Rotherham, Doncaster, Barnsley, Leeds, Wakefield, Calderdale, Kirklees, and Bradford). We then exclude the City of London and Isles of Scilly given their relative small size. We also exclude inner London given the very different level and trajectory for pupil attainment. This gives 128 local authorities that form the potential donor pool.

As specified in the SAP, we further restrict the sample to the donor pool that delivers the best prediction for the lagged outcomes of the synthetic control group in the pre-treatment period. As described in the 'Statistical Analysis' sub-section, in our case this means restricting to local authorities where the average KS2 (standardised) fine points score in maths in the pre-treatment period is within 0.25 standard deviations of that for South and West Yorkshire. This reduces the donor pool to 118 local authorities in the primary and secondary analysis.¹⁵

The analysis all proceeds at the local authority level (excluding some robustness checks at school level). The local authority data is derived from pupil-level data and the SAP described using all pupils to create this data. However, we have applied a number of necessary and desirable sample restrictions to derive the local authority data. The effect of sample restrictions on the number of pupils by year used in the analysis is shown in Table 4.

First, we exclude pupils with missing Key Stage 2 or pupil covariate data. This equates to about 20–30,000 or about 5–6% of pupils in most years. The one exception is 2010 where about 30% of pupils have missing data owing to the boycott of Key Stage 2 in that year. This has minimal effect on our analysis as 2010 is not used as one of the key predictor variables. Excluding pupils with missing data is unlikely to be a source of bias in this context as these variables are only used to predict lagged outcomes, which the method seeks to minimise in any case.¹⁶

Second, to ensure a fair comparison, we also focus on state-funded primary schools, which are the focus of the trial in South and West Yorkshire. This largely means excluding special and independent schools. This amounts to about 5% of the original sample in 2003, falling to 2% by 2017. This fall over time is likely to be down to fewer pupils in special and independent schools sitting Key Stage 2 tests over time. This illustrates that many of these pupils do not sit Key Stage tests (our primary outcomes), which is often a choice for the school and a further reason why these pupils are excluded.

Third, we show the effect of excluding inner London, City of London, and the Isles of Scilly, which together represent about 5% of pupils each year. Note that the sum of the number of pupils excluded may not sum to the difference between the original and final sample as some pupils may be excluded

¹⁵ Regression to the mean is unlikely to be a problem in this context. The synthetic controls are chosen to best approximate the pre-treatment trends in outcomes, which could be non-linear, and these exclusion criteria seek to ensure that synthetic controls are chosen from a pool of broadly similar LAs to start with. There is no reason to expect that outcomes in South and West Yorkshire would diverge positively or negatively relative to the synthetic controls after the intervention.

¹⁶ It would only be a source of bias if using full or imputed data could achieve an even better match on pre-treatment trends. In excluding missing data, we are implicitly assuming that pupils with missing data have the mean values for each local authority for cases with non-missing data. An alternative would be to use multiple imputation methods. However, most of the cases have missing data across multiple variables.

for multiple reasons (for example, many pupils in independent schools have missing pupil characteristics too).

The final sample is generally about 90% of all pupils in the data each year. This includes pupils in the potential control group (donor pool) and in the treatment group (South and West Yorkshire). The treatment group in the final sample in 2017 includes 1,031 primary schools with non-missing Key Stage 2 outcomes and 42,973 pupils. This represents the total number of pupils in the analysis sample. However, looking across all year groups, there were about 350,000 pupils in state-funded primary schools in South and West Yorkshire in January 2017.¹⁷

Table 4: Effect of sample restrictions on pupil sample sizes

Year of KS2 tests	Original sample	Missing pupil data	% of original sample	School restrictions	% of original sample	LAs exclusions	% of original sample	Final sample	% of original sample
	637,675	37,224	5.8%	29,528	4.6%	30,266	4.7%	572,826	89.8%
2004	613,726	33,277	5.4%	27,107	4.4%	29,470	4.8%	553,171	90.1%
2005	610,918	40,719	6.7%	25,630	4.2%	29,380	4.8%	543,477	89.0%
2006	595,306	34,032	5.7%	22,377	3.8%	29,391	4.9%	534,310	89.8%
2007	588,122	34,360	5.8%	22,239	3.8%	29,530	5.0%	526,554	89.5%
2008	597,332	34,450	5.8%	21,233	3.6%	29,931	5.0%	535,446	89.6%
2009	580,033	30,218	5.2%	18,596	3.2%	28,974	5.0%	523,158	90.2%
2010	571,341	171,913	30.1%	14,903	2.6%	28,839	5.0%	380,028	66.5%
2011	554,888	22,135	4.0%	13,122	2.4%	28,808	5.2%	505,532	91.1%
2012	544,222	18,555	3.4%	12,047	2.2%	28,519	5.2%	498,450	91.6%
2013	540,197	19,610	3.6%	11,963	2.2%	28,959	5.4%	493,174	91.3%
2014	561,543	20,387	3.6%	12,712	2.3%	30,368	5.4%	512,533	91.3%
2015	579,263	20,300	3.5%	12,888	2.2%	31,079	5.4%	529,454	91.4%
2016	592,272	35,538	6.0%	13,061	2.2%	31,984	5.4%	527,083	89.0%
2017	604,575	33,477	5.5%	12,101	2.0%	32,401	5.4%	541,089	89.5%

Note: Missing pupil data relates to missing Key Stage 2 test data, FSM eligibility, EAL status, SEN status or ethnicity; School restrictions focus on state-funded primary schools only (mostly excluding special and independent schools); LA exclusions relate to inner London, City of London and the Isles of Scilly.

The main aim of our impact evaluation is to estimate the impact of offering advocacy across the whole of South and West Yorkshire. This will include the effects of participation and the actual effects of advocacy provision itself.

To help interpret our impact estimates, it is important to be aware that advocacy providers could only be recruited in seven of the nine local authorities in South and West Yorkshire. This will immediately limit the potential size of any impact estimate.

The IPE report shows that about 42% of schools across South and West Yorkshire signed up for advocacy provision, with about 38% attending at least one workshop. Amongst local authorities where there was a provider, recruitment varied from 29% to 100% of schools.

Despite an initial aim to focus on disadvantaged and under-performing schools, the IPE report shows that the campaign was not successful in targeting underperforming schools or schools with the most disadvantaged pupils in South and West Yorkshire. This could be because providers were not aware of this aim. Indeed, the IPE report argues that there was very limited targeting of these schools by advocacy providers.

¹⁷ <https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2017>

Pupil and school characteristics

In Table 5, we compare some key pupil and school characteristics across South and West Yorkshire, our Synthetic Control group (which differs across the two primary outcomes), and our donor pool of local authorities (that is, all local authorities outside of South and West Yorkshire excluding inner London and the Isles of Scilly). All pupil characteristics are averaged over the whole pre-treatment phase, whilst the pupil to teacher assistant ratio (pupil:TA ratio) is averaged over 2010/2011 to 2014/2015 and the proportion of primary schools that are academies is averaged over 2010/2011 to 2014/2015. With the exception of the pupil:TA ratio, all variables are derived from the National Pupil Database. The pupil:TA ratio is derived from School Workforce Census statistics across various years.

A comparison between columns (1) and (3) shows how South and West Yorkshire differs from the rest of England (these are the same for both outcomes). It is more deprived, on average, with 19.5% of pupils eligible for free school meals (FSM) as compared with 15.3% in the rest of England. The mix of ethnicities is also slightly different from the rest of England, with a slightly smaller share of White-British ethnicity, a smaller share of Black ethnicity, and larger share of Asian ethnicity. There is also a larger share of pupils with English as an Additional Language (EAL): 15.6% as compared with 11.9% for the rest of England. The proportion of pupils with a statement of special educational needs (SEN) is similar across groups. However, we see a slightly smaller pupil:TA ratio in South and West Yorkshire, suggesting a greater reliance on teaching assistants as compared with the rest of England. However, in both cases there is approximately one TA per class of just under 30. A slightly smaller share of primary schools, on average, were academies in South and West Yorkshire (7.0%) than in the rest of England (8.1%).

The second column shows the characteristics of the synthetic control group for both outcomes. For maths, all the pupil and school characteristics are within one percentage point of each other (often less) and pupil:TA ratio is also very similar across the two groups. The standardised differences shown in column (4) are also all small in value.

For English, some slightly larger differences remain with the difference in the percentage of pupils eligible for FSM around four percentage points and the difference in the share of schools that are academies gets worse relative to the comparison with the rest of England. Both of these differences are large in standardised terms too, for example, the difference in the proportion of children eligible for FSM is equivalent to about one half of a standard deviation. Such differences are not necessarily problematic, however, the aim of the synthetic control method is to minimise differences in pre-treatment outcomes rather than predictor variables.

Table 5: Comparison of time-constant control variables across treatment and control (South and West Yorkshire, synthetic control, selected donor pool)
a) KS2 (standardised) mathematics score

	South and West Yorkshire	Synthetic control	Rest of England (donor pool)	Standardised difference (SW Yorkshire – synthetic control)
% Eligible for Free School Meals	19.4	19.8	15.3	-0.05
% White British	77.5	77.3	79.1	0.01
% Black	2.0	2.0	3.5	0.00
% Asian	13.2	12.8	6.7	0.05
% English as an Additional Language	15.6	15.5	11.9	0.01
% Statement SEN	2.0	2.0	1.9	0.00
Pupil:TA ratio (2010/2011–2014/2015)	28.06	29.42	33.30	-0.09
% Academy (2011/2012–2014/2015)	7.0	7.1	8.1	-0.01
Root Mean Squared Prediction Error		0.006	0.096	

b) KS2 (standardised) English Score

	South and West Yorkshire	Synthetic control	Rest of England (donor pool)	Standardised difference (SW Yorkshire – synthetic control)
% Eligible for Free School Meals	19.4	23.4	15.3	-0.55
% White British	77.5	76.6	79.1	0.04
% Black	2.0	2.0	3.5	0.00
% Asian	13.2	12.2	6.7	0.11
% English as an Additional Language	15.6	15.5	11.9	0.01
% Statement SEN	2.0	2.1	1.9	-0.05
Pupil:TA ratio (2010/2011–2014/2015)	28.06	29.12	33.30	-0.07
% Academy (2011/2012–2014/2015)	7.0	10.2	8.1	-0.27
Root Mean Squared Prediction Error		0.007	0.126	

Notes and Sources: Authors' calculations using the National Pupil Database and School Workforce Census statistics (<https://www.gov.uk/government/collections/statistics-school-workforce>). Unless otherwise stated, all characteristics are averaged over 2003 to 2015. Differences are standardised by the standard deviation in each characteristic for the donor pool.

At the bottom of the tables, we then show the root mean squared prediction error (RMSPE) in the pre-treatment phase as compared with South and West Yorkshire. For maths, the RMSPE falls from 0.096 national standard deviations in the rest of England to 0.006 national standard deviations amongst the synthetic controls. For English, even though the balance on pupil and school characteristics is slightly worse, the RMSPE falls even more, from 0.126 national standard deviations in the rest of England to 0.007 amongst the synthetic controls. This suggests it is harder to create a synthetic control group for English, which may explain why some predictor variables appear imbalanced.

In general, however, these figures suggest that our synthetic control group is doing a relatively good job in matching the pupil and school characteristics in South and West Yorkshire, as well as the average pre-treatment primary outcomes.

Figure 2 then shows the evolution over time of the primary outcomes for South and West Yorkshire, the synthetic control group and the donor pool. This is shown for both the pre-treatment phase, where we are seeking to explicitly match trends between South and West Yorkshire and the synthetic control group, and the post-treatment phases.

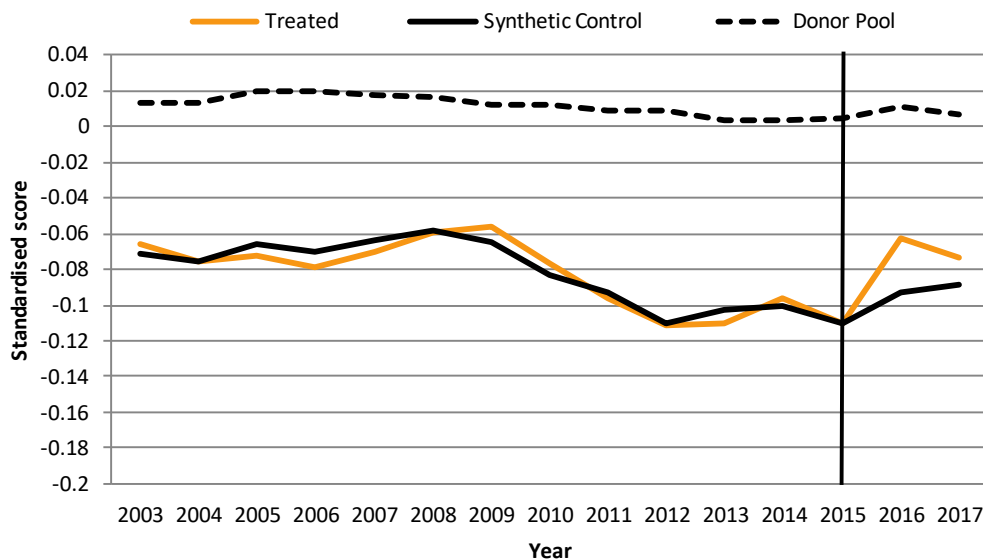
As can be seen, the primary outcomes are lower on average in South and West Yorkshire as compared with the rest of England in the donor pool. This gap also increases over time for both maths and English. For maths, the gap goes from about 0.08 to 0.11 standard deviations between 2003 and 2015. For English, the increase in the gap is even faster, going from 0.12 to 0.17 standard deviations between 2003 and 2015. These different time trends represent an important reason why simply matching on pre-treatment outcomes just before the point of treatment could be misleading, as outcomes for South and West Yorkshire might be on a different trajectory.

The solid black lines then show the estimated outcomes over time amongst out synthetic controls. This shows that our synthetic control group is able to replicate the pre-treatment outcomes across maths and English very closely. This includes the largely flat trend up to 2009 and decline thereafter up to 2015.

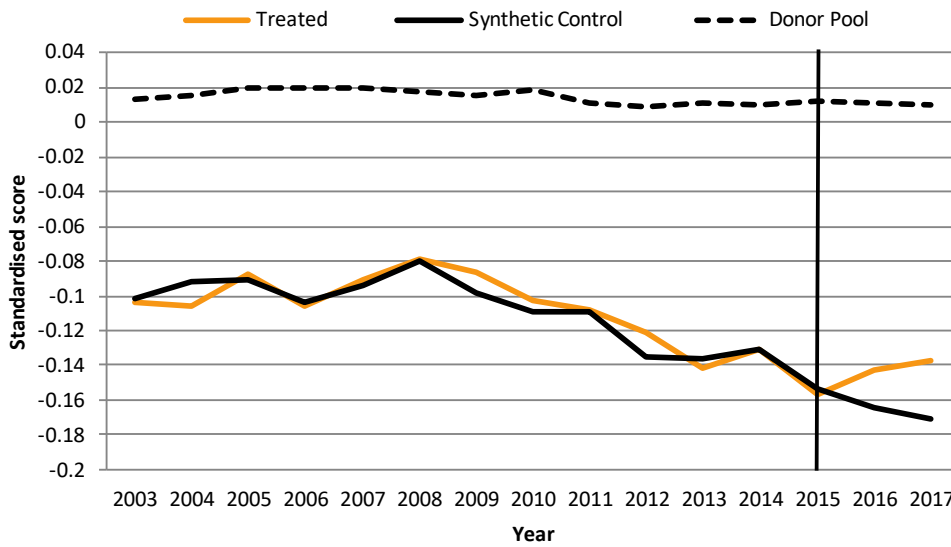
In the post-treatment phase in 2016 and 2017, we see an uptick in both maths and English scores in South and West Yorkshire relative to the synthetic control group. This suggests that the TA campaign may have improved outcomes in South and West Yorkshire. In the next section, we therefore turn to estimating this impact more directly and whether it is within the normal bounds for variation over time in the performance of different areas.

Figure 2: Key Stage 2 maths and English standardised fine points score 2002 to 2017

a) Maths



b) English



Note: 'Donor Pool' represents all other local authorities in England, excluding inner London, City of London and the Isles of Scilly. All outcomes are standardised at the national level by year.

Outcomes and analysis

Primary ITT analysis

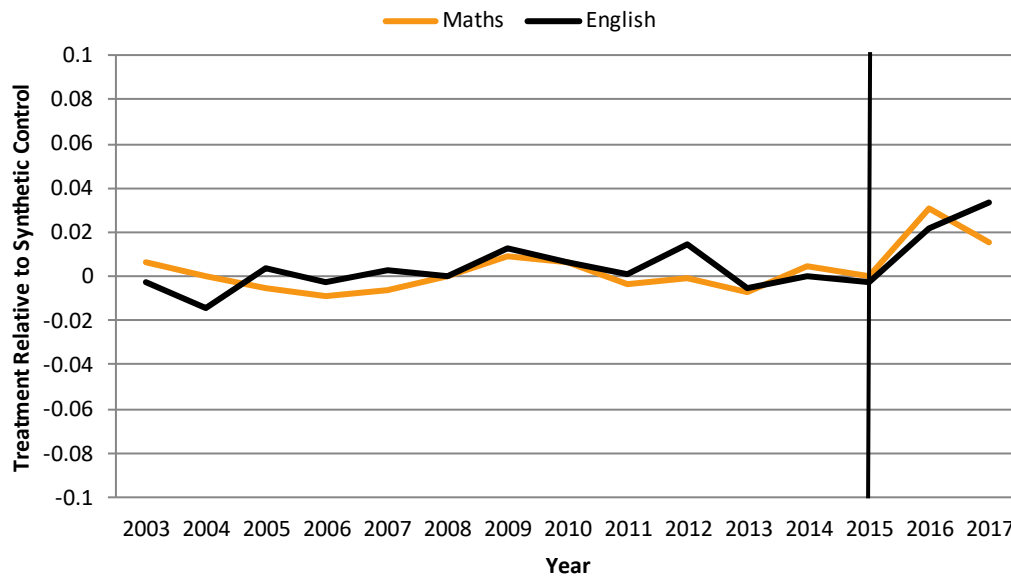
In Figure 3, we report the differences between South and West Yorkshire and our synthetic control groups for our primary outcomes.

For maths, we see that the post-treatment gap is around 0.03 standard deviations in 2016 and 0.015 standard deviations in 2017. These differences are both higher than any seen over the pre-treatment phase, but only just in the case of the primary outcome (the largest difference in the pre-treatment phase is 0.009 standard deviations).

For English, we see more evidence of a growing impact of the treatment. In 2016, the implied impact estimate is 0.02 standard deviations, growing to over 0.03 standard deviations in 2017. The latter is about twice the size of any other absolute difference seen over the pre-treatment phase.

The larger gap in outcomes in 2016 for maths could reflect a positive response to the new KS2 tests in South and West Yorkshire, particularly as we have already hypothesised that it would be slightly early to see an impact of the treatment. However, this is not replicated for English, where the impact appears to grow with time.

Figure 3: Difference in primary outcomes between South and West Yorkshire and synthetic controls, 2002 to 2017



Notes and sources: Authors' calculations using the National Pupil Database. Both KS2 English and maths outcomes are standardised at the national level by year.

Following on from this analysis, the obvious question is whether the estimated impacts are statistically significant or not. As synthetic control analysis makes use of aggregate units (local authorities in our case), there is no uncertainty about the outcomes across the different units. There is uncertainty as to the extent to which the synthetic controls represent a valid counterfactual. The nature of synthetic control analysis, however, means that inference cannot be conducted in the usual fashion through calculation of standard errors and confidence intervals.

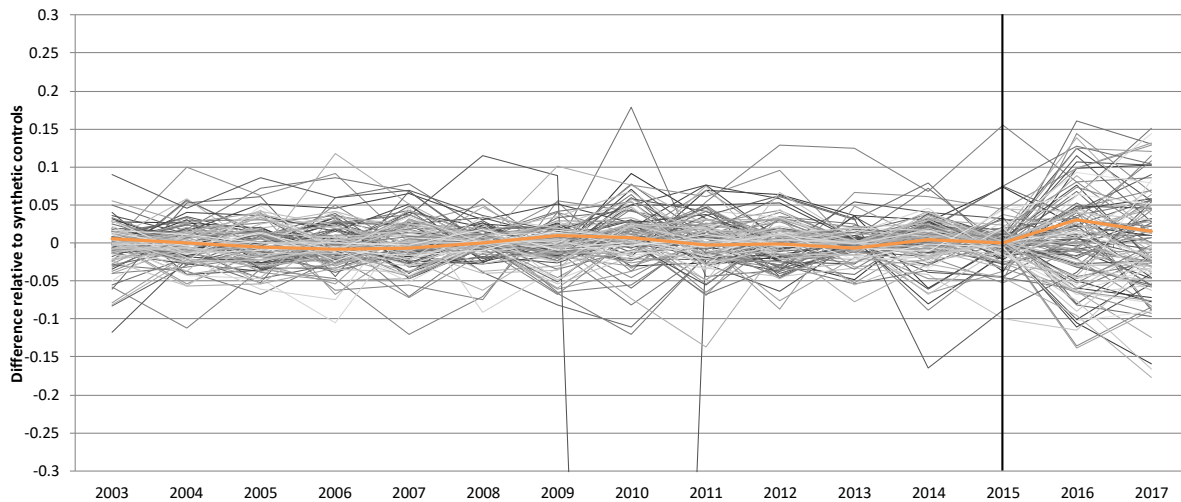
Following Abadie et al. (2010), we instead use permutation or placebo tests. We return South and West Yorkshire to the donor pool and then iteratively treat each local authority as the treatment unit. We find synthetic controls for that local authority in order to match pre-treatment outcomes and characteristics, and calculate the difference in the post-treatment phase. Our interest lies in whether the estimated difference for South and West Yorkshire is in some sense unusual.

Figure 4 shows the results of this iterative procedure for maths and English. Each light grey line represents the estimated difference between each local authority and its respective synthetic controls. In each case, the orange line shows the difference between the actual treatment unit (South and West Yorkshire) and its synthetic controls.

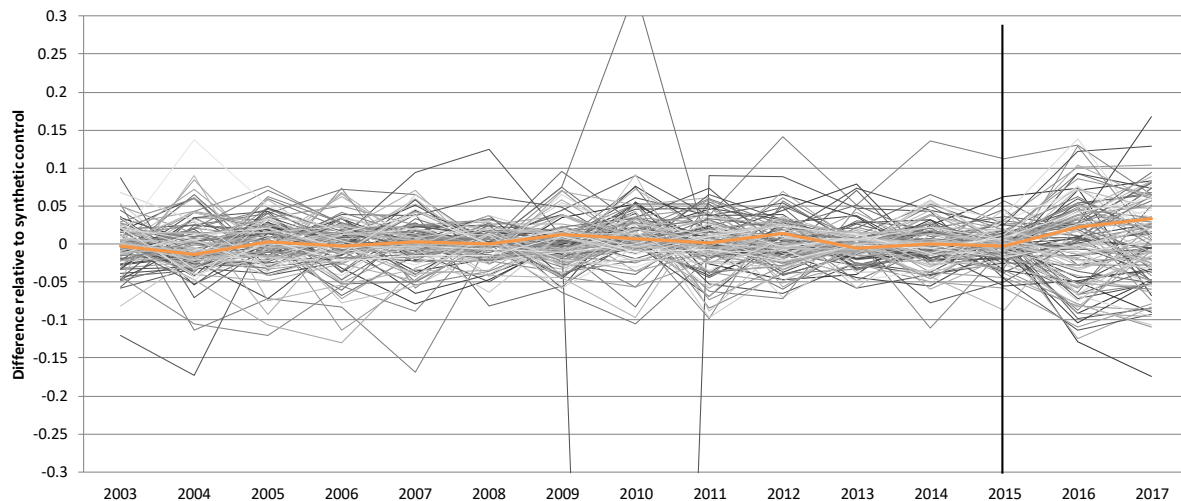
For both primary outcomes, we see that most local authority differences relative to their synthetic controls lie within 0.05 to -0.05 standard deviations. However, there are also a number of local authorities where the differences are much larger than that. There is then more variation after 2015. This is to be fully expected given that the synthetic control group will seek to match differences up to 2015. This is unlikely to cause any bias. These post-treatment differences explicitly capture the normal variation over time in the performance of different areas relative to each other, but also the different responses of local authorities to the new Key Stage 2 curriculum and tests from 2016 onwards. The question we are seeking to ask in these permutation tests is whether the observed change for South and West Yorkshire is larger than these 'normal' changes over 2016 and 2017.

Figure 4: Permutation tests for KS2 maths and English (std) fine points scores

a) Maths



b) English



Notes: 'Rest of England' represents all local authorities outside South and West Yorkshire, 'Donor Pool' excludes inner London and the Isles of Scilly. Each grey line represents the difference between a local authority and its synthetic control groups. The orange line show South and West Yorkshire relative to its synthetic control.

It is also very clear that our main synthetic control analysis for South and West Yorkshire (orange line) does a better job of matching pre-treatment trends in outcomes than the vast majority of our placebo tests across local authorities. Our main synthetic control analysis manages to reduce pre-treatment differences to less than 0.01 standard deviations across all but one pre-treatment year. The differences are much larger for other local authorities in the donor pool. This is almost certainly driven by the fact that each local authority is smaller than our treatment unit of South and West Yorkshire (which is nine local authorities combined together). Matching the pre-treatment trends in a single local authority is naturally more challenging than for a larger group combined together. Unfortunately, however, there is no natural set of geographic units to compare with South and West Yorkshire, which is larger than individual local authorities but smaller than the set of regions defined by the Office for National Statistics. We discuss how this could be addressed in future research in the conclusion, such as simulating regions of a similar size to South and West Yorkshire (as per Acemoglu et al. (2016) and Krief et al. (2016)).

Notwithstanding this important point, the permutation analysis at the local authority level is still quite revealing. For both maths and English, we see that the post-treatment differences for South and West Yorkshire lie well within the differences we see for other local authorities. However, these figures include

a number of local authorities where pre-treatment differences between each local authority and its respective synthetic control are very large. As specified in the Statistical Analysis Plan, we therefore exclude a number of local authorities based on the pre-treatment difference in the RMSPE as compared with our actual treatment unit.

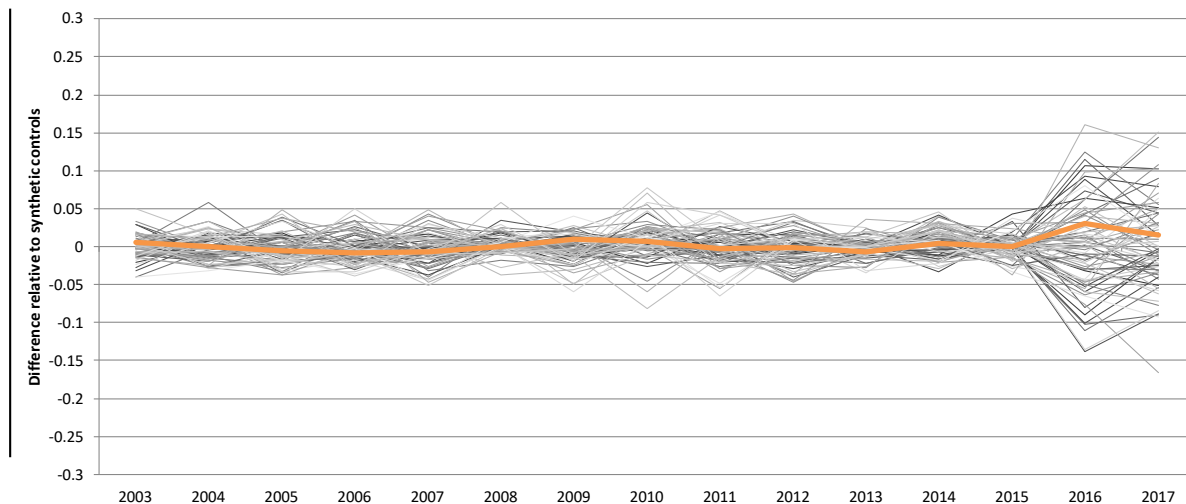
We did not specify the difference we would use in the Statistical Analysis Plan so we have chosen to implement a cut-off of five times the RMSPE for our main analysis. This matches that used in Abadie et al. (2010). This is essentially an arbitrary threshold designed to exclude outliers.

For English, the RMSPE is 0.007, so this means excluding any local authorities where the RMSPE in pre-treatment outcomes is larger than 0.036 standard deviations, leaving 90 local authorities outside of South and West Yorkshire. For maths, the RMSPE is 0.006, so we exclude any local authorities where the RMSPE is above 0.027 standard deviations, leaving 71 local authorities outside of South and West Yorkshire. Tables D5 and D6 in Appendix D provides further details about each local authority in terms of the RMSPE for each local authority, the implied impact estimate for 2017, and the ratio between the two. These tables also show which local authorities are excluded based on the ‘five times RMSPE’ rule.

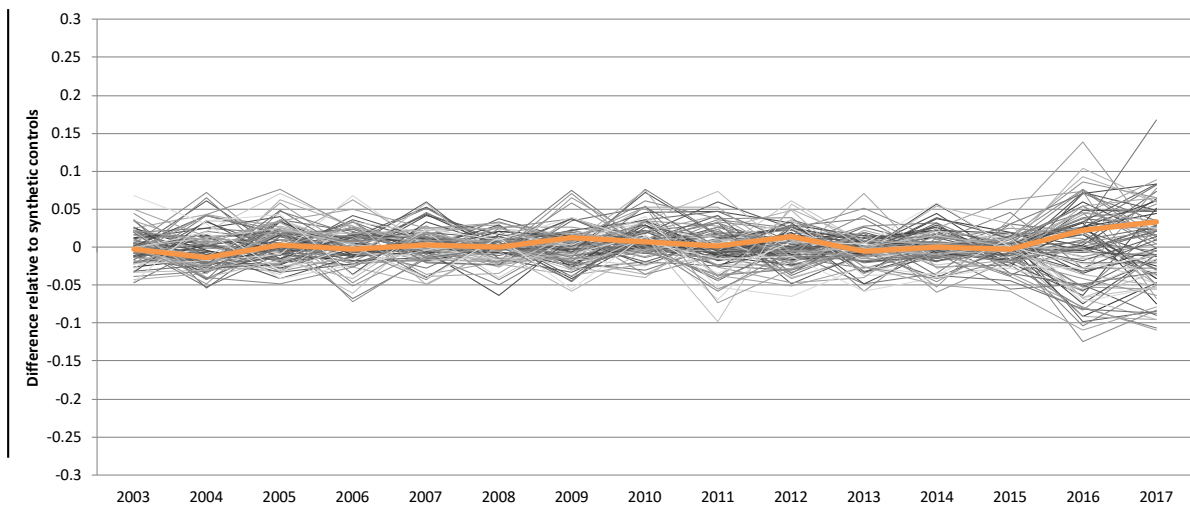
The resultant set of figures is shown in Figure 5. This suggests that the differences for South and West Yorkshire are well within the normal range. For English, out of the 90 remaining local authorities, 47 had larger post-treatment differences in outcomes and 43 showed smaller differences. For maths, 55 have larger post-treatment outcomes as compared with South and West Yorkshire, out of a total of 71 remaining local authorities. On their own, these figures would suggest the estimated impacts for South and West Yorkshire are well within the normal range over time.

Figure 5: Permutation tests for KS2 maths and English (std) fine points scores, excluding LAs where pre-treatment RMSPE is greater than five times that for treatment

a) Maths



b) English



In Figure 6, we then show a histogram of the ratio between post-treatment difference in 2017 and the RMSPE in the pre-treatment phase. This allows us to judge whether the uptick for South and West Yorkshire is large relative to the RMSPE in the pre-treatment phase as compared with the pattern seen for other local authorities.

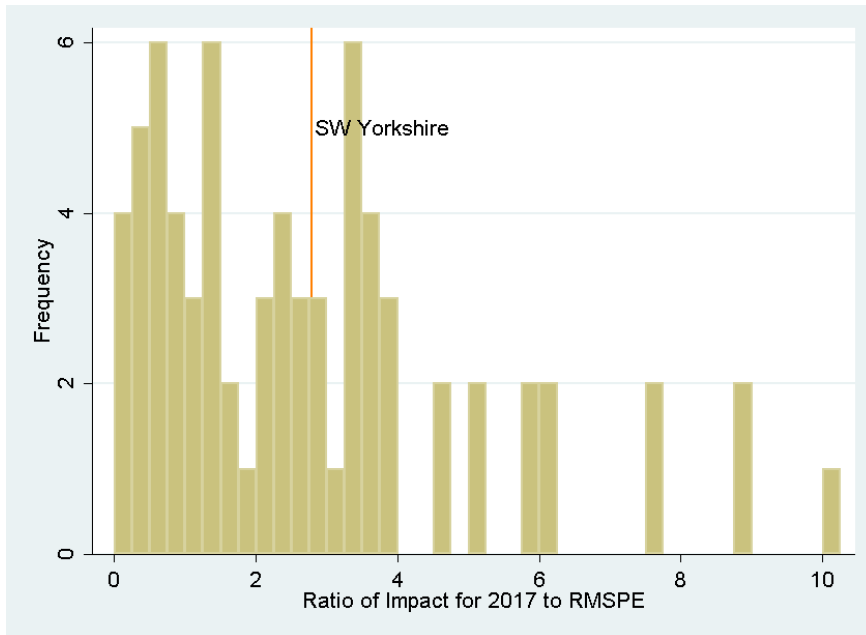
For maths, the ratio between the post-treatment difference and the RMSPE for South and West Yorkshire is 2.8. This is larger than that seen for 42 of the 72 remaining local authorities, so is well within the normal range of possibilities.

For English, we see that the ratio for South and West Yorkshire is 4.6, which is the seventh largest out of a total of 91 local authorities remaining in the donor pool (those where the RMSPE is less than five times the size of that for South and West Yorkshire). This suggests that the post-treatment difference for English in South and West Yorkshire is unusual relative to other local authorities where we can perform a good match of pre-treatment outcomes. We can interpret the range of local authority effects relative to their synthetic controls as a simulation of the range of local authority effects one might expect in normal circumstances. About 7% of local authorities see a larger increase relative to pre-intervention differences as compared with South and West Yorkshire. Abadie et al. (2010) treat this like a p-value, and we can thus see the difference for English as being on the boundary of statistical significance. This is not a standard way to conduct statistical inference and should thus be interpreted with some caution.

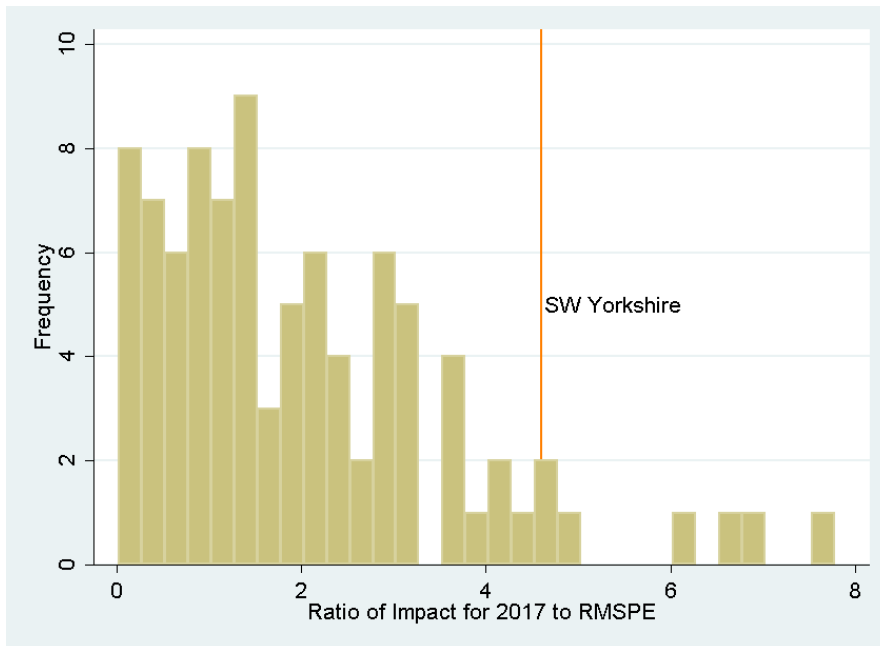
It should also be noted that we implemented an essentially arbitrary threshold for excluding local authorities from this analysis. However, if we had instead included all local authorities, then the implied p-value would have been smaller as all excluded local authorities have a smaller ratio of the treatment effect relative to RMSPE as compared with South and West Yorkshire. For example, if we had set a threshold of ten times the RMSPE, we would have obtained a p-value of 5%.

Figure 6: Histogram of ratio of impact estimates for 2017 relative to pre-intervention RMSPE

a) Maths



b) English



Robustness checks

In Tables D1, D2, D3 and D4 we report the results of our robustness checks.

Table D1 reports the estimated impact on our primary outcomes by iteratively leaving out each local authority that receives positive weight and is included in the synthetic controls. In both cases, the estimated impacts vary slightly, but there is little suggestion that the estimated impact is driven by a big difference for just one local authority.

Second, Table D3 reports a variety of raw and synthetic control methods for calculating the impact of the treatment. First, we show the raw differences as compared with the rest of England, which show large negative differences as one might expect given the ways in which South and West Yorkshire differs from the rest of England. In column (2), these raw differences are reduced to near zero once we only include local authorities receiving positive weight in the synthetic control analysis. The third column (3) then shows our main results applying the synthetic control weights.

The next set of three columns show the results based on alternative synthetic control specifications (as specified in the Statistical Analysis Plan, these are the specification with next lowest RMSPE). Column (4) uses the whole donor pool, which slightly increases our impact estimates for English. Column (5) uses the preferred sample for English (within 0.15 standard deviation of the pre-treatment outcome for South and West Yorkshire) and the first specification. Results are almost identical to the main estimates. Column (6) uses the preferred sample for English and the second specification. Again, the results are very similar to the main estimates.

The final column shows our synthetic control estimates when we exclude neighbouring local authorities from the donor pool. The results for English are largely unchanged. The result for maths moves closer to zero though does not change our qualitative finding of no evidence of an effect.

Table D4 then compares our estimates against a wide-range of different non-experimental methods at the local authority and school level. Starting with the local authority level results, OLS largely matches our preferred synthetic control estimates for both English and maths. OLS also seems to find a zero estimate for 2015 in the case of maths, though finds a negative estimate in the case of English. Given that we expect this difference to be zero, this provides some initial indication that there may be unobservable factors biasing our OLS results.

Propensity score matching suggests a near-zero effect in 2017, whilst Mahalanobis matching produces clear negative estimates. Unfortunately, however, diagnostic checks on both matching approaches suggest they are achieving a very poor balance in covariates, with Rubin's B falling well above the threshold of 25 recommended by Rubin (2001). Furthermore, when we conduct our validation test by (falsely) assuming the treatment began a year earlier, we see that the kernel matching estimators become slightly more positive and the Mahalanobis estimators remain negative. The diagnostic checks also suggest a poor balance.

The rest of Table D4 then repeats this analysis at the school level. In this case, OLS produces near zero results for our primary outcomes in 2017, though negative ones for 2015. The balance achieved in matching is better than at the local authority level. These generally produce negative estimates for 2017, which become more negative in 2015. This, again, suggests that unobservable determinants may be biasing these non-experimental estimators.

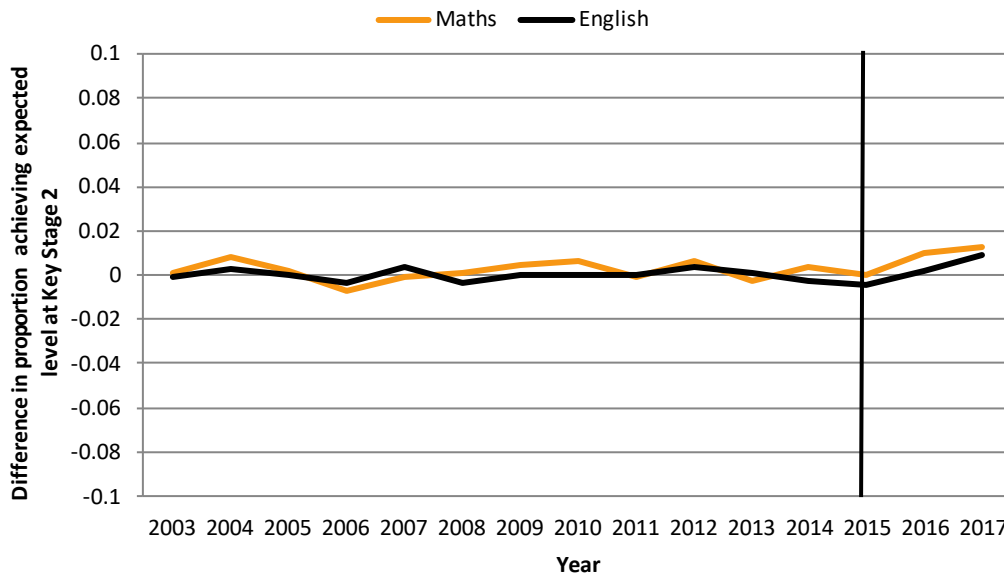
In summary, robustness checks on our synthetic control methods suggest that our results are robust to alternative specifications and excluding neighbouring local authorities. Non-experimental estimators produce quite different estimates of the results, generally zero or negative. However, the balance achieved in matching is poor at the local authority level. These non-experimental estimators also generally produce non-zero estimates for the effect in 2015, which should be close to zero if there are no major unobservable determinants of outcomes over time. In contrast, synthetic control methods produce near-zero results for 2015, suggesting a greater ability to control for unobservable determinants of outcomes. As a result, our synthetic control results remain our preferred method.

Secondary ITT analysis

Figure 7 then repeats the analysis for our secondary outcomes for achieving the expected levels in Key Stage 2 maths and English (our secondary outcomes for fine points scores in 2016 are already included within Figure 3). This shows a similar pattern to our primary outcomes. Differences between the

treatment and synthetic controls are small in the pre-treatment phase up to 2015, with an uptick in the post-treatment phase of around one percentage point in both cases. Both upticks are also larger than any difference seen over the pre-treatment phase. However, we repeated the permutation analysis for these threshold outcomes, which would imply p-values of 25% for maths and 22% for English,

Figure 7: Difference in secondary outcomes between South and West Yorkshire and synthetic controls, 2002 to 2017



Notes and sources: Authors' calculations using the National Pupil Database.

Subgroup analysis

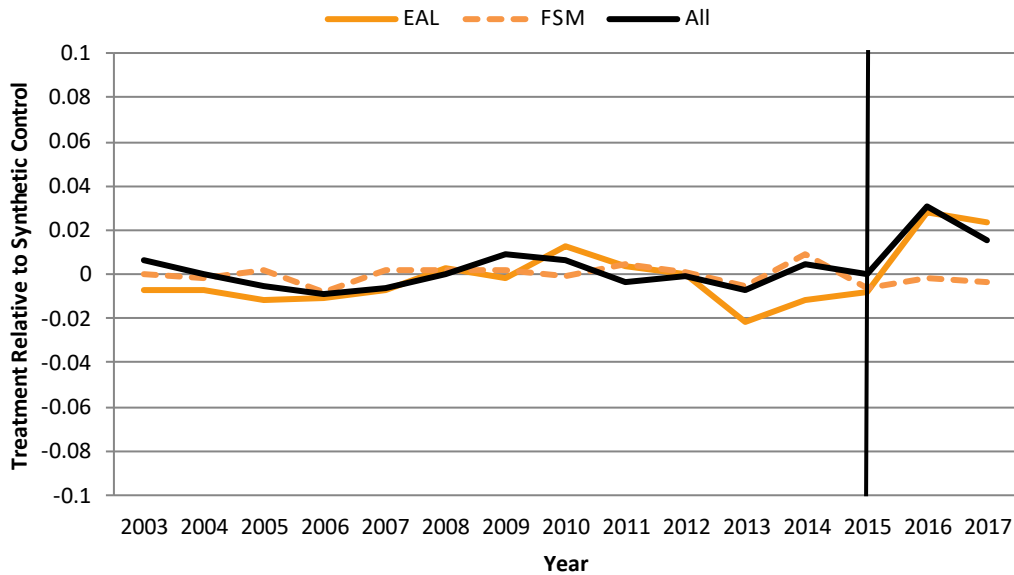
In Figure 8, we report the results of subgroup analysis and repeat the primary analysis for all pupils for reference. We follow the same process as for our main analysis except that the outcomes are in each case only defined for pupils eligible for Free School Meals (FSM) and those with English as an Additional Language (EAL). We seek to match pre-treatment trends in outcomes for these two groups and look at the post-treatment differences between the outcomes for these two groups.

In both cases, we see almost no difference in the post-treatment phase in 2016 and 2017 for FSM pupils relative to their synthetic controls. We see a small uptick for EAL pupils. This is true across both maths and English, with the uptick about 0.02 standard deviations in both outcomes for pupils with EAL in 2017. However, the estimated differences for pupils with EAL in 2017 are both small (around 0.02 standard deviations). This is very slightly larger than the primary results for maths and smaller than the primary results for English. For EAL pupils, the estimated impact for maths is about 2.4 times the RMSPE in the pre-treatment phase and about 3.3 times for English. Both of these ratios were well within the normal range seen for other local authorities in the main analysis. Furthermore, all the same caveats applied to the interpretation of the main results apply here too.

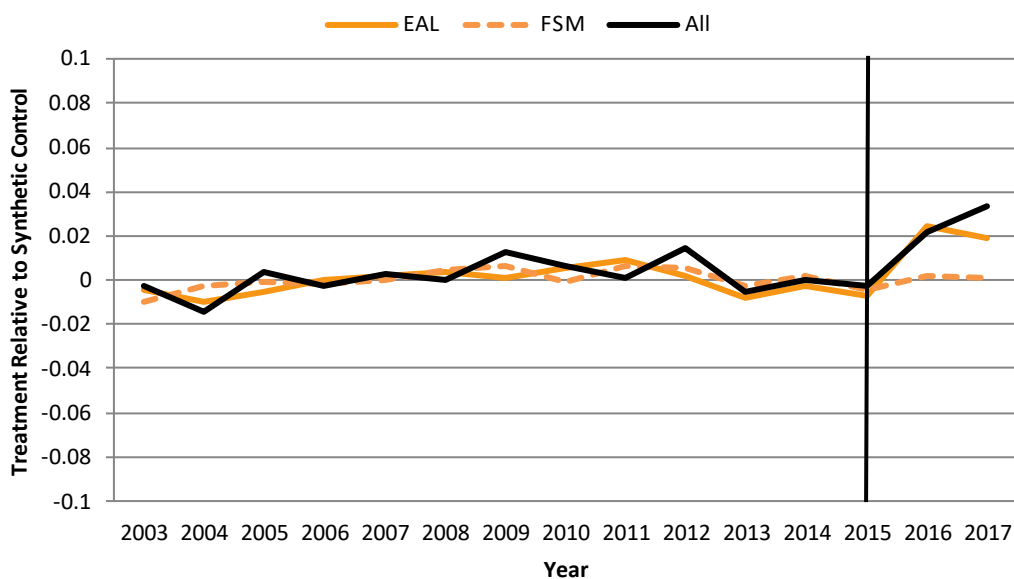
Therefore, there is little evidence of a differential impact of the treatment on pupils eligible for FSM or pupils with EAL.

Figure 8: Difference in primary outcomes between South and West Yorkshire and synthetic controls for FSM and EAL subgroups, 2002 to 2017

a) Maths



b) English



Additional Analysis

We also conducted a small amount of additional analysis to aid interpretation of the impact analysis. First, we undertook some on-treatment analysis amongst schools in South and West Yorkshire. Second, we compared our results against two further national benchmarks based on our experimental email campaigns outside of South and West Yorkshire.

Two concrete ways in which schools in South and West Yorkshire could show engagement with the campaign was to sign up with one of the advocacy providers or participate in a trial of a TA-led intervention, both of which were heavily targeted at the area. According to data collected by Sheffield Hallam University in its **IPE report**, about 480 or 42% of all primary schools in South and West Yorkshire

signed up with one of the advocacy providers and around 20% of primary schools received additional TA interventions as part of the TA campaign. The report, however, expresses some caution in terms of the accuracy of data on trial participation: these two groups of schools are not mutually exclusive and schools could sign up with an advocacy provider and/or participate in a trial. Indeed, schools that signed up for advocacy provision were often targeted by advocacy providers for participation in trials.

Table 6 seeks to estimate the difference in primary outcomes between schools that signed up to an advocacy provider or were involved with a trial as compared with other schools in South and West Yorkshire. We estimate these effects using OLS regression where we include dummy variables for signing up to an advocacy provider and participating in a trial, and control for additional covariates (the same as those used in the primary analysis). We do this for primary outcomes in 2017, and for the period just before the treatment began in 2015. This analysis is not fully robust or causal as there could still be unobservable factors driving both outcomes and participation in trials.

These results provide no clear empirical evidence that schools that participated in advocacy or additional TA-led interventions had higher Key Stage 2 scores than schools that did not. This was the case in 2015 before advocacy provision began and two years later in 2017.

These results are not definitive evidence that advocacy or the TA-led interventions had no effect as these are not causal estimates. The estimates are also consistent with the small positive effects we estimate in our primary analysis. However, large effects of either advocacy or additional trials would only be consistent with the results in Table 6 if the latter are subject to large negative biases (for example, schools that signed up to advocacy or trials were likely to have much worse outcomes, even after controlling for the covariates included in Table 5).

Table 6: Estimated differences in outcomes between schools that signed up for advocacy provision or signed up to a trial and other schools in South and West Yorkshire, 2015 and 2017

	2015	2017
KS2 English		
Signed up to advocacy	0.020	0.012
Standard Error	[0.022]	[0.022]
Participated in a trial	-0.019	-0.030
Standard Error	[0.026]	[0.026]
KS2 Maths		
Signed up to advocacy	0.025	0.018
Standard Error	[0.022]	[0.024]
Participated in a trial	-0.036	-0.034
Standard Error	[0.026]	[0.030]

Note: Outcomes represents the average standardised KS2 points score for each subjects. Effects are estimated using OLS regression with dummy variables for each group of schools and the same set of covariates used as predictors in our primary analysis. Only includes schools in South and West Yorkshire.

In a separate experiment, we tested different ways to incentivise schools outside of South and West Yorkshire to engage with the guidance and national campaign. Some schools were randomly selected to receive a control standard email, whilst others were randomly determined to receive emails with incentives to engage with the information.¹⁸

Table 7 shows the average outcomes for schools within the email experiment that did, and did not, receive incentives to engage with the guidance materials. The average outcomes are near identical to

¹⁸ The level of these incentives was set relatively low in order to act as a small nudge (For example, a £5 amazon voucher). This was largely an arbitrary decision.

one another. This strongly suggests that our email incentives experiment had little impact on engagement with the guidance in the national control group for South and West Yorkshire.

Table 7: Average levels of Key Stage 2 scores in 2017 for schools that did, and did not, receive email incentives to engage with guidance materials

	KS2 (std) English point scores	KS2 (std) maths point scores
No Incentives	0.026	-0.020
Standard Error	[0.005]	[0.005]
Number of Schools	7043	7043
Incentives	0.025	-0.018
Standard Error	[0.005]	[0.006]
Number of Schools	5038	5038
Difference	-0.001	0.002
Standard Error	[0.007]	[0.008]

Note: Outcomes represents the average standardised KS2 points score for schools that were randomly determined to receive incentives or no incentives. The sample of schools is slightly reduced compared with all schools observed in 2017 as not all schools were randomised in the original experiment (due to missing school data or missing email addresses). We also exclude schools that cannot be matched across 2015 and 2017 due to changing school URNs (most were matched based on the database of open academies, but a small number could not be).

Summary and interpretations

To summarise, our synthetic control approach is able to achieve good balance in pre-treatment outcomes in both maths and English, as well as pupil and school characteristics. We then find a post-treatment difference in English scores of about 0.03 standard deviations. This is large relative to pre-treatment differences in outcomes between the treatment and synthetic control and relative to changes over time seen in other local authorities. For maths, we find little evidence of any impact.

These figures represent our estimates of seeking to implement an advocacy campaign right across South and West Yorkshire. There are then a number of ways to interpret these results and the mechanisms driving them. Here we speculate on the plausibility of these different mechanisms.

First, it could reflect a genuine causal impact of the campaign. If this is the case, a 0.03 impact need not necessarily be seen as small, given the large number of schools and pupils involved. The analysis covers about 43,000 taking KS2 tests in South and West Yorkshire, or about 350,000 pupils across all year groups. Furthermore, advocacy providers could not be found for two of the nine local authorities and only about 40% of schools signed up for advocacy. Based on these figures for participation, the results would suggest an impact of over 0.06 amongst schools that took up the offer (assuming no effect on those that did not take up the offer). This in itself would also be substantial. It would be consistent with over half of these schools implementing a change in practice that improved pupil attainment by 0.3 standard deviations (such as greater use of structured interventions) and which then affected more than one third of pupils at these schools. This is in excess of the optimistic scenario we set out in the introduction. If there is a causal effect of 0.03, then we can certainly interpret this as large.

Furthermore, the control group would have been positively affected by the national campaign and guidance materials. Over 40% of headteachers reported reading the guidance document within six months of publication in 2015. This could have improved overall usage of teaching assistants in the control group over time. Whilst this would have also affected South and West Yorkshire, it does make the search for a differential impact in South and West Yorkshire slightly more demanding, and would need to reflect more than just becoming aware of the guidance. This said, our on-treatment analysis suggests that schools that took up the offer of advocacy had similar outcomes to those of other schools in South and West Yorkshire. Although we cannot give this a causal interpretation, it would caution against interpreting our main results as potentially reflecting a large causal effect of advocacy.

Second, it may be that the treatment effect is driven by the additional RCT interventions in South and West Yorkshire. The evidence for this interpretation is relatively weak. In the **IPE report**, colleagues show that around 20% of primary schools received additional TA interventions as part of the TA campaign. Although we cannot establish a reliable benchmark for what was happening in the rest of England, it seems reasonable to assume the figure was a lot lower outside of South and West Yorkshire. However, three of the four interventions promoted by advocacy providers related to numeracy, where we find little impact. Only one focuses on literacy (Switch-on Reading and Writing) and the published EEF evaluation finds little evidence that the trialled version of the treatment had a positive impact on literacy outcomes (NatCen, 2017). Furthermore, we found little evidence that schools in South and West Yorkshire participating in these trials showed higher test scores than schools that did not.

Third, it could be that South and West Yorkshire was unusual in its response to the new Key Stage 2 tests and curriculum in being an area that was best able to improve its outcomes. By definition, this is hard to test. Any bias could be either positive or negative, large or small.

Fourth, it may reflect the fact that we are better able to minimise pre-treatment differences for South and West Yorkshire. Indeed, it has the lowest RMSPE amongst all the permutations used in the inference procedure. Indeed, given that methods for calculating the statistical significance of synthetic control estimates are still in their infancy, particularly for multiple treated units, some caution is therefore needed in interpreting these results.

Costs

The estimated costs of providing advocacy and training are shown in Table 8 below. These are presented from the perspective of a policymaker deciding whether or not to organise such a campaign across a large geographic area. Cost data was only supplied for four of the seven providers involved. We therefore focus on the average figures.

The average cost per school involved was relatively cheap at around £392 per school, equating to £131 per year when spread over a three-year period.

These are relatively low costs per school and per pupil. However, it is important to remember that these low unit costs were only achievable because they covered a large number of schools and pupils. Even just the four providers shown here covered 306 schools and 106,000 pupils. To gauge the value-for-money of the campaign, one must compare the total costs of the campaign against the overall improvements. For example, our synthetic control implies an improvement of 0.03 standard deviations in English test scores at age 11 across 306 schools and the overall programme across these four groups cost about £120,000.

Table 8: Cost of advocacy provision and guidance campaign

	AP1	AP4	AP 5	AP6	All
Number of schools	65	51	134	56	306
Number of pupils	24,388	15,519	50,933	15,053	105,894
Total reported costs	£26,293	£27,600	£37,875	£28,100	£119,868
Cost per school	£404.50	£541.18	£282.65	£501.79	£391.72
Cost per pupil	£1.08	£1.78	£0.74	£1.87	£1.13
Cost per school over 3 years	£134.83	£180.39	£94.22	£167.26	£130.57
Cost per pupil over 3 years	£0.36	£0.59	£0.25	£0.62	£0.38

Note: Cost data supplied by advocacy providers. Advocacy providers are anonymised and labelled in an identical fashion to the IPE report. Numbers for schools, pupils, and total costs represent totals across all three providers.

Conclusion

Key conclusions

1. There is evidence that KS2 English test scores showed an improvement of 0.03 standard deviations in South and West Yorkshire as compared with our synthetic control group.
2. There is no evidence of an impact on maths scores.
3. New KS2 tests and curriculum began from 2016. This could be biasing our results, but the size and direction of any bias is uncertain.
4. The cost per school was relatively low, about £130 per school per year (when spread over three years).
5. Synthetic control methods worked well using education administrative data to approximate pre-treatment trends in South and West Yorkshire. They also appear to perform better than other non-experimental estimators like OLS and matching. We recommend they are used more for appropriate large-scale educational interventions.

Interpretation

This evaluation estimated the impact of offering advocacy and guidance aimed at improving the use of teaching assistants across the whole of South and West Yorkshire. Given the lack of a standard control group, we instead applied a synthetic control approach. This seeks to re-weight other local authorities to best mimic the outcomes in South and West Yorkshire during the pre-treatment period. This is the first time such methods have been used in an EEF evaluation and a further goal was therefore to understand how such methods could be applied in other evaluations.

The synthetic control approach was largely successful in finding a set of suitable weights across local authorities, which allowed us to closely mirror the outcomes in South and West Yorkshire over the pre-treatment phase. This is despite the notably lower and worsening outcomes for South and West Yorkshire relative to the rest of England as a whole. Our comparison with other non-experimental methods further suggests that the synthetic control approach is better able to mirror pre-treatment outcomes in a situation where unobserved factors are likely to be an important determinant of outcomes. The synthetic control method estimates no difference in outcomes before the treatment begins. In contrast, other non-experimental estimators (here, OLS and matching) produce non-zero estimates, suggesting unobserved factors remain important.

The results suggest that Key Stage 2 English scores were about 0.03 standard deviations higher in South and West Yorkshire two years after the beginning of the treatment. Although inference cannot be undertaken in standard ways, our tests results suggest the difference for English in South and West Yorkshire is larger than we might expect compared with what we see in other local authorities over time. The procedure would suggest a p-value of about 0.07. This would make the result statistically significant at the less conventional 10% level. The overall result is robust to changing the size of the donor pool, using different covariates, and is not sensitive to including particular local authorities in the synthetic control group.

For maths, there is little evidence of any difference in outcomes between South and West Yorkshire and the synthetic controls.

In order to ensure the results can be interpreted in a causal manner, one needs to be sure that there are no factors likely to be influencing outcomes in South and West Yorkshire differentially to the synthetic control group over time. Fortunately for interpretation, most changes to school funding and school labour markets were happening in a constant manner across England. School funding per pupil was largely frozen in cash terms per pupil, public sector pay was subject to a 1% national cap, and we are not aware of any additional initiatives that were focused on South and West Yorkshire in particular.

There are, however, two main factors that complicate interpretation. First, the treatment happened at exactly the same time as a major change in the Key Stage 2 tests. Whilst the synthetic control group is constructed to replicate trends over time in the treatment group and responses to national policy changes, the changes to the KS2 tests were substantial and represent an 'infrequent event'. It is therefore uncertain whether the results represent a genuine causal impact or a different response to the tests in South and West Yorkshire. Second, additional interventions were focused on South and West Yorkshire at the same time as the advocacy campaign. About 20% of schools were participating in a trial of TA-led structured interventions in South and West Yorkshire. However, three of the four trials promoted by advocacy providers were focused on numeracy, where we find little impact. Furthermore, there is no evidence of any difference in outcomes between schools that did, and did not, participate in such trials in South and West Yorkshire.

If there was a causal impact of 0.03 standard deviations on English test scores, as suggested by our evaluation, then there are good reasons to interpret such an effect as sizeable. It would represent an improvement amongst 43,000 pupils across South and West Yorkshire, and is estimated relative to a national control group that was relatively well informed by the EEF national campaign and guidance documents. Any causal effect would also likely be concentrated amongst the seven local authorities where the EEF was able to appoint advocacy providers and the 40% of schools that took up the offer. Advocacy provision was also relatively inexpensive, with an average cost per school of about £390 per school, or £130 per school spread over three years. Such small unit costs are only achievable due to the largescale nature of the project, however.

Limitations

The fact that the treatment happened at the same time as the introduction of new Key Stage 2 test scores has already been identified as a limitation for interpreting the results, and also applies to their generalizability.

Beyond this, the main limitation of the analysis is the calculation of statistical significance. Because standard errors cannot be calculated in the normal way, we instead rely on permutation or placebo tests. We effectively treat all local authorities as if they were South and West Yorkshire, and then examine whether the improvement in results after the treatment for South and West Yorkshire is unusual relative to other local authorities. For English, this appears to be the case; for maths, it is not.

However, individual local authorities are smaller than South and West Yorkshire, which comprises nine different local authorities. As a result, we find it harder to re-create pre-treatment trends in outcomes for individual local authorities than we do for South and West Yorkshire. Our placebo tests are therefore very strong tests since what might be unusual variation for a group of nine local authorities might be quite normal for a single local authority.

In future evaluations using synthetic control analysis, we therefore recommend specifying treatment and donor geographic units of similar size to avoid this problem. This might involve applying treatments at a region level, local authority level, or predefining something in between. One could also create groups of local authorities within regions, for example, grouping together large urban conurbations within regions such as Greater Manchester, Merseyside. or the West Midlands conurbation.

Alternatively, one could follow the approach taken in Krief et al. (2016) and Acemoglu et al. (2016). They also seek to use synthetic control methods in the context of multiple treatment units. From the donor pool of control units, they randomly simulate regions or groups of a similar size to the treatment group and then apply synthetic control methods to these simulated groups. They repeat this approach up to 5,000 times in order to obtain standard errors and confidence intervals. This is computationally intensive and we will seek to explore using such methods in future work, including the appropriateness of such methods to cases where aggregate effects are of interest.

Future research and publications

In a further evaluation for the EEF, we will seek to publish more detail on the results of the experimental variations in emails sent to schools and their effect on engagement with information.

We will also seek to publish the results of this synthetic control analysis in academic journals. In particular, we will seek to further explore the implications of the differences between the synthetic control approach and other non-experimental estimators to show the effect of different assumptions as well as different approaches for conducting inference.

Alongside this report, the EEF will publish the implementation and process evaluation (IPE) of the South and West Yorkshire campaign (Maxwell et al., 2019). The EEF will also publish a process and implementation evaluation of a different approach to scaling-up the TA guidance in Lincolnshire and a 'lessons learned' report that compares the two approaches (also Maxwell et al., 2019).

References

- Abadie, A. and Gardeazabal, J. (2003) 'The Economic Costs of Conflict: A Case Study of the Basque Country', *American Economic Review*, American Economic Association, 93 (1, March), pp. 113–132.
- Abadie, A., Diamond, A., Hainmueller, J. (2010) 'Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program', *Journal of the American Statistical Association*, 105 (490), pp. 493–505.
- Acemoglu, D., Johnson, S., Kermani A., Kwak, J. and Mitton T. (2016) 'The Value of Connections in Turbulent Times: Evidence from the United States', *Journal of Financial Economics*, 121, pp. 368–391.
- Belot, M. and Vandenberghe, V. (2014) 'Evaluating the "threat" effects of grade repetition: exploiting the 2001 reform by the French-Speaking Community of Belgium', *Education Economics*, 22 (1).
- Blanden, J., Greaves, E., Gregg, P., Macmillan, L. and Sibieta, L. (2015) 'Understanding the improved performance of disadvantaged pupils in London', CASE—Social Policy in a Cold Climate Working Paper 21, Centre for Analysis of Social Exclusion, LSE.
- Blatchford, P., Russell, A. and Webster, R. (2012) *Reassessing the impact of teaching assistants: How research challenges practice and policy*, Abingdon, Oxon: Routledge.
- Gorard S., See, B. H. and Siddiqui, N. (2014) 'Switch On Reading', EEF Evaluation Report: <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/switch-on-reading#research-results>
- Hinrichs, P. (2012) 'Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities', *Review of Economics and Statistics*, 94 (3).
- Krief, N., Grieve, R., Hangartner, D., Turner, A., Nikolova, S. and Sutton, M. (2016) 'Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units', *Health Economics*, 25 (12).
- National Foundation for Educational Research (2014) 'Catch up Numeracy', EEF Evaluation Report: <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/catch-up-numeracy/>
- NatCen Social Research (2017) 'Switch-On Reading Effectiveness Trial', EEF Evaluation Report: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_Switchon_Effectiveness.pdf
- Rubin, D. B. (2001) 'Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation', *Health Services and Outcomes Research Methodology*, 2, pp. 169–188.
- Sharples, J., Webster, R. and Blatchford, P. (2015) 'Making Best Use of Teaching Assistants: Guidance Report', Education Endowment Foundation: https://v1.educationendowmentfoundation.org.uk/uploads/pdf/TA_Guidance_Report_Interactive.pdf
- Wilson, D., Burgess, S. and Briggs, A. (2011) 'The dynamics of school attainment of England's ethnic minorities', *Journal of Population Economics*, Springer: European Society for Population Economics, 24 (2, April), pp. 681–700.

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF’s approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

Cost rating	Description
£ £ £ £ £	<i>Very low:</i> less than £80 per pupil per year.
£ £ £ £ £	<i>Low:</i> up to about £200 per pupil per year.
£ £ £ £ £	<i>Moderate:</i> up to about £700 per pupil per year.
£ £ £ £ £	<i>High:</i> up to £1,200 per pupil per year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

OUTCOME: KS2 Maths and English

Rating	Criteria for rating	Initial score	Adjust	Final score
	Design	MDES	Attrition	
5	Randomised design	<= 0.2	0-10%	
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%	
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%	
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%	2
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%	
0	No comparator	>=0.6	>50%	

Adjustment for threats to internal validity [0]

Threats to validity	Threat to internal validity?	Comments
Threat 1: Confounding	Moderate	The synthetic control creates a plausible counterfactual for South and West Yorkshire achieving very low Root Mean Squared Prediction Errors over the whole period (0.006 for Maths and 0.007 for English) and for each year. Pre-intervention trends in the treated region are closely replicated by the synthetic control. However, given the complexities inherent to this method and the emerging methodological literature on its implementation and interpretation, results from this method should be interpreted with caution.
Threat 2: Concurrent Interventions	Low	Other interventions were focused on South and West Yorkshire at the same time as the advocacy campaign. About 20% of schools were participating in a trial of TA-led structured interventions in South and West Yorkshire. However, three of the four trials promoted by advocacy providers were focused on numeracy, where little impact was found. Furthermore, there is no evidence of any difference in outcomes between

		schools that did, and did not, participate in such trials in South and West Yorkshire
Threat 3: Experimental effects	No information available	No information available as part of this report
Threat 4: Implementation fidelity	No information available	No information available as part of this report
Threat 5: Missing Data	Low	As other data-driven methods, the synthetic control method used for this evaluation assumes complete data. Even if there might be some school-level data missing, by aggregating school-level data at the region-level there is effectively no missing data even if that aggregation might be done with slightly different school-level data in each period.
Threat 6: Measurement of Outcomes	Moderate	This evaluation used the KS2 fine point scores in Maths and English which are valid, reliable and educationally relevant measures of attainment. These tests, however, were changed in 2016 just after the treatment was implemented. This could pose a threat to the validity of the results if these changes could be expected to differentially affect the area subject to the TA campaign (South and West Yorkshire). However, there is no clear reason to believe this is the case.
Threat 7: Selective reporting	Moderate	Data-driven approaches like the one used in this evaluation generally require an iterative approach for data analysis. However, the main analyses and approach in this study were pre-specified with changes to the protocol and SAP appropriately recorded and explained. When decisions were not pre-specified they generally followed some usual approaches used in the (still incipient) literature for this method.

- **Initial padlock score:** [2] Padlocks – This was a well-conducted synthetic control method that achieved reasonable levels of balance between the treated unit and its synthetic control, closely following pre-intervention trends in the outcomes of interest. However, the concept of inference in Synthetic Controls remains complex given that only one unit is treated which requires creating ‘placebo treatment effects’ to compare the magnitude of the results in the treated unit with those obtained where no effect would be expected (because they are placebos). Even if this method provides an indication of the ‘statistical significance’ of results, this method remains in its infancy and should be treated with caution.
- **Reason for adjustment for threats to validity:** [0] Padlocks – No severe threats to the validity of the results were identified. However, given the limitations of the method used results should be interpreted with caution.
- **Final padlock score:** initial score adjusted for threats = [2] Padlocks

Appendix C – EEF National Campaign Activities

Communications schools and stakeholders

- Hard copies of the *Making best use of Teaching Assistants* guidance report to all schools nationally (using Raise Online database) June 15
- Two emails from EEF Chief Executive to schools nationally - a follow up email to the national posting in July 2015 and a further email, in January 2017, emphasizing a new set of implementation resources and the online course.
- Letter from EEF Chief Executive to all LAs – June 2015
- Letter from EEF Chief Executive to key contacts and national stakeholders – June 2015
- Jonathan Sharples Blog on TA-led interventions – March 16
- A range of organisations sent the guidance to their schools and associates, including: Challenge Partners, Achievement for All, The Key, Unison, National Education Trust, Cornwall LA, Berkshire LA (hard copies to all SENCOs) – Summer/Autumn 2015
- Presentations of work to policy makers – OECD, EU, US government, What Works Global Summit, Australia federal and state.
- 95,000 unique page views of the EEF’s materials up to November 2017.
- NFER Omnibus Survey data suggests ~70% heads were aware of the guidance (~40% read it, ~15% acted on it) – Nov 2015
- Sheffield Hallam University’s independent evaluation of the *Making best use of Teaching Assistants* campaign in Lincolnshire (see below) suggests ~90% of heads in the two sample regions (Lincolnshire, Kent and Medway) have read the guidance.

Training/events

- Campaign to disseminate and implement the *Making best use of Teaching Assistants* guidance across 480 schools in eight LAs within S&W Yorkshire – 2015-16. Sheffield Hallam University’s independent evaluation of this campaign finds demonstrable changes in school practices relating to TA deployment, in line with the research.
- Campaign to disseminate and implement the *Making best use of Teaching Assistants* guidance across ~300 schools in Lincolnshire (£190,000 of funding secured through Lincolnshire Learning Partnership) – 2016-17. Sheffield Hallam University’s independent evaluation of this campaign finds demonstrable changes in school practices relating to TA deployment in line with the research, plus the development of county-wide infrastructure to support evidence-informed school improvement.
- Six EEF Research Schools are currently delivering courses on the guidance to schools in their region (Stockport, Suffolk, Norfolk, Devon, York, Oldham). Approx. 200 schools – Sept ember 2017
- 75-100 presentations by Jonathan, Rob and other EEF colleagues at events for schools (e.g. EEF events, The Key, Optimus, Osiris, Derbyshire LA, Suffolk LA, MITA events) – ongoing
- *Making best use of Teaching Assistants* online course accessed >50,000 times – June 2017

National press/media

- Press release for the publication of the guidance report – Mar 2015
- Press release for the new implementation resources – Jan 2017
- Press releases for publication of evaluation reports for TA-led interventions (e.g. ABRA, Catch Up) – ongoing through 2015/16.
- Articles on the TA guidance in: TES, The Economist, Radio 4 Today Programme, Education Business, National Governors’ Association magazine

Policy

- Full page update on the TA campaign in the OfSTED School Inspection Update (p12/13) – Nov 2016
- Reference in Teaching School Council review ‘Effective Primary Teaching’ (aka Keeble review), section on TAs – Autumn 2016

- Reference in National Audit Office review on Pupil Premium spending – June 2015
- Reference in DfE guidance on work force planning – ‘School workforce planning - Guidance for schools’ – Jan 2017
- Reference in Unison materials – Summer 2016
- New Zealand government are using the guidance in their support for schools - <http://inclusive.tki.org.nz/guides/teacher-aide-practice/>
- Western Australian government are using the guidance in their support for schools around SEN provision
- Around 500 schools have accessed MITA since 2014.
- About 3,000 TAs have been trained via the Maximising the Practice of TAs CPD programme, plus about 1,200 teachers.
- 40 licensed MPTA trainers nationally and 25 Reviewers.
- Reviewers can be commissioned to deliver the **TA Deployment Review**.. The Review supports peer-to-peer support and has a self-evaluation framework for school to conduct internal audits. Over 2,000 copies have been downloaded since it was launched in Sept 2017.
- 2 online auditing tools: **Staff survey** and **TA Audit**
- **maximisingtas.co.uk** is a comprehensive resource bank for TA-related materials
- Media and press <http://maximisingtas.co.uk/research/research-blogs-articles.php>

Appendix D – Additional Analysis

Table D1 – Weights used and Effects of Leaving out each Local Authority on Primary Outcome Impacts

a) Maths

Local Authority	Weight	Impact Estimate in 2017 Leaving out this Local Authority
Medway	0.311	0.025
Poole	0.037	0.018
Knowsley	0.032	0.036
Coventry	0.020	0.021
North East Lincolnshire	0.070	0.017
Peterborough	0.337	0.020
Oldham	0.094	0.010
North Yorkshire	0.092	0.017
Barking and Dagenham	0.005	0.020
All	Sum = 1	0.015

b) English

Local Authority	Weight	Impact Estimate in 2017 Leaving out this Local Authority
Rochdale	0.212	0.037
Blackpool	0.180	0.024
Nottingham	0.163	0.044
North East Lincolnshire	0.128	0.038
Middlesbrough	0.092	0.040
Walsall	0.078	0.044
Peterborough	0.065	0.022
Tameside	0.048	0.040
Poole	0.020	0.044
Knowsley	0.012	0.043
All	Sum = 1	0.033

Table D2 – Weights used and Effects of Leaving out each Local Authority on Secondary Threshold Outcome Impacts

a) Maths

Local Authority	Weight	Impact Estimate in 2017 Leaving out this Local Authority
Peterborough	0.206	0.010
Oldham	0.143	0.005
Medway	0.137	0.015
Blackburn with Darwen	0.132	0.005
Kingston Upon Hull, City of	0.097	0.015
Walsall	0.074	0.015
North Yorkshire	0.061	0.013
Nottingham	0.059	0.012
Tameside	0.053	0.015
North East Lincolnshire	0.037	0.014
All	Sum = 1	0.014

b) English

Local Authority	Weight	Impact Estimate in 2017 Leaving out this Local Authority
Oldham	0.207	0.007
North East Lincolnshire	0.202	0.010
North Yorkshire	0.153	0.005
Blackburn with Darwen	0.144	0.006
Walsall	0.135	0.006
Peterborough	0.083	0.010
Medway	0.049	0.012
Nottingham	0.026	0.009
Tameside	0.001	0.008
All	Sum = 1	0.009

Table D3 – Synthetic Control Robustness Checks

Local Authority	Raw	LAs receiving positive weight	Synthetic Control (main spec)	Synthetic Control (alt spec 1)	Synthetic Control (alt spec 2)	Synthetic Control (alt spec 3)	Synthetic Control (no neighbours)
Maths							
Impact on South and West Yorkshire	-0.092	-0.026	0.015	0.023	0.019	0.024	-0.009
English							
Impact on South and West Yorkshire	-0.128	-0.002	0.033	0.051	0.035	0.038	0.030

Alternative Specification 1 uses the full donor pool and specification (1) from Table 2. Alternative Specification 2 uses the preferred sample for English (within 0.15 standard deviations of South and West Yorkshire) and specification (1) from Table 2. Alternative Specification 3 uses preferred sample for English and specification (2) from Table 2. No Neighbours excludes all local authorities directly adjacent to South and West Yorkshire.

Table D4 – Non-experimental Robustness Checks

Local Authority	Synthetic Control (main spec)	OLS	Propensity Score Kernel Matching	Mahalanobis Matching	OLS	Propensity Score Kernel Matching	Mahalanobis Matching
Maths 2017							
Impact on South and West Yorkshire	0.015	0.006	-0.006	-0.011	0.021	0.001	-0.048
Standard Error		<i>0.011</i>	<i>0.015</i>	<i>0.019</i>	<i>0.022</i>	<i>0.067</i>	<i>0.024</i>
Rubin's R			1.19	1.15		1.24	1.18
Rubin's B			6.97	22.20		7.31	17.34
English 2017							
Impact on South and West Yorkshire	0.033	-0.017	-0.048	-0.051	0.027	-0.003	-0.084
Standard Error		<i>0.010</i>	<i>0.015</i>	<i>0.019</i>	<i>0.026</i>	<i>0.072</i>	<i>0.048</i>
Rubin's R			1.19	1.15		1.33	1.44
Rubin's B			6.97	22.20		>100	>100
Maths 2015							
Impact on South and West Yorkshire	0.011	-0.024	-0.028	-0.034	0.000	0.031	-0.064
Standard Error		<i>0.010</i>	<i>0.013</i>	<i>0.018</i>	<i>0.015</i>	<i>0.046</i>	<i>0.032</i>
Rubin's R			1.24	1.18		n/a	2.65
Rubin's B			7.31	17.34		n/a	>100
English 2015							
Impact on South and West Yorkshire	0.004	-0.058	-0.078	-0.087	-0.028	0.016	-0.107
Standard Error		<i>0.010</i>	<i>0.014</i>	<i>0.019</i>	<i>0.013</i>	<i>0.047</i>	<i>0.039</i>
Rubin's R			1.24	1.18		n/a	2.65
Rubin's B			7.31	17.34		n/a	>100

School-level OLS and matching estimators are for a balanced panel of schools and use the same specification as our preferred one from Table 2 (excluding data at Teaching Assistants, which is often suppressed at school-level for data confidentiality reasons). When 2015 outcomes are used, 2015 covariates are naturally excluded from the specifications. For matching estimators, Rubins' B measures the absolute standardised difference of the means of the linear index of the propensity score in the treated and control group and Rubin's R measures the ratio of treated to control group variances of the propensity score index). Rubin (2001) recommends that B be less than 25 and that R be between 0.5 and 2 for the samples to be considered sufficiently balanced. "n/a" indicates a convergence failure.

Table D5 – Additional details of permutation analysis by local authority for KS2 Maths

Local Authority	Root Mean Square Prediction Error (RMSPE)	Impact Estimate in 2017	Ratio (Impact /RMSPE)
SW Yorkshire	0.006	0.015	2.794
Dorset	0.009	-0.053	6.262
Wirral	0.009	-0.088	10.051
East Riding of Yorkshire	0.009	-0.052	5.837
Lancashire	0.009	-0.032	3.451
Cambridgeshire	0.009	-0.031	3.328
Oxfordshire	0.009	-0.005	0.530
Cornwall	0.011	-0.043	3.883
Essex	0.012	0.044	3.761
South Tyneside	0.012	0.103	8.804
Norfolk	0.012	-0.090	7.533
Devon	0.012	-0.023	1.903
Oldham	0.012	-0.040	3.268
West Sussex	0.013	-0.035	2.671
County Durham	0.013	0.028	2.095
Worcestershire	0.014	-0.016	1.124
Derbyshire	0.014	-0.038	2.702
Gloucestershire	0.014	-0.033	2.316
Staffordshire	0.014	0.018	1.278
Swindon	0.015	-0.006	0.440
Croydon	0.015	0.091	6.037
Surrey	0.015	0.051	3.419
Manchester	0.015	0.010	0.636
East Sussex	0.015	-0.055	3.614
Nottinghamshire	0.016	-0.025	1.601
Leicestershire	0.016	-0.012	0.763
Warrington	0.016	0.143	8.803
Blackburn with Darwen	0.017	0.078	4.741
Dudley	0.017	-0.077	4.621
Warwickshire	0.017	-0.062	3.622
Medway	0.017	-0.013	0.741
Wiltshire	0.018	0.023	1.268
Coventry	0.018	-0.008	0.461
Kent	0.019	0.015	0.796
Somerset	0.019	-0.004	0.224
Birmingham	0.019	-0.026	1.353
Liverpool	0.019	-0.015	0.765
Hampshire	0.019	0.031	1.620
Hillingdon	0.019	0.056	2.898
Cumbria	0.019	-0.003	0.156
Ealing	0.020	0.043	2.198
Cheshire	0.020	-0.059	2.929
South Gloucestershire	0.021	-0.030	1.441
Suffolk	0.021	0.030	1.457
Kingston Upon Hull, City of	0.021	0.108	5.199
North Lincolnshire	0.021	0.044	2.118

Bournemouth	0.021	0.071	3.392
Torbay	0.021	0.064	3.049
Salford	0.021	-0.011	0.501
Northamptonshire	0.022	-0.023	1.053
Lincolnshire	0.022	-0.166	7.618
Wolverhampton	0.022	0.032	1.463
Southend-on-Sea	0.022	0.050	2.284
Blackpool	0.022	0.050	2.285
Brighton and Hove	0.022	0.016	0.725
West Berkshire	0.022	0.007	0.320
Thurrock	0.022	0.083	3.714
Waltham Forest	0.023	0.058	2.477
North Tyneside	0.024	0.028	1.176
Bracknell Forest	0.024	-0.092	3.865
Barking and Dagenham	0.025	0.130	5.261
Bedfordshire	0.025	-0.085	3.365
Havering	0.025	0.151	5.950
Gateshead	0.026	0.102	3.983
Northumberland	0.026	-0.018	0.678
Leicester	0.026	0.013	0.497
Sefton	0.027	0.000	0.013
Shropshire	0.027	0.002	0.077
Wigan	0.027	0.027	0.990
Enfield	0.027	-0.072	2.671
Milton Keynes	0.027	-0.018	0.666
Stockport	0.028	-0.046	1.660
5 x RMSPE for SW Yorkshire = 0.027 (LAs below here excluded from inference analysis)			
Hertfordshire	0.028	-0.047	1.669
North East Lincolnshire	0.029	-0.057	1.968
Reading	0.029	-0.021	0.733
Nottingham	0.029	0.012	0.419
North Somerset	0.029	-0.023	0.800
Sunderland	0.030	0.121	4.017
Halton	0.030	-0.125	4.143
Slough	0.030	0.070	2.302
Bexley	0.031	-0.024	0.787
Bath and North East Somerset	0.031	-0.097	3.109
Merton	0.031	0.057	1.807
York	0.032	0.046	1.452
Middlesbrough	0.032	0.132	4.091
Hounslow	0.033	0.010	0.317
Brent	0.033	-0.026	0.801
Newcastle upon Tyne	0.033	0.055	1.640
Tameside	0.034	-0.082	2.457
Bolton	0.034	-0.023	0.689
Peterborough	0.034	-0.076	2.245
Redbridge	0.034	0.128	3.745
Telford and Wrekin	0.035	-0.040	1.157
Windsor and Maidenhead	0.035	-0.025	0.728

North Yorkshire	0.036	-0.018	0.492
Stockton-on-Tees	0.036	0.007	0.183
Bristol, City of	0.037	-0.047	1.297
St. Helens	0.037	-0.159	4.342
Walsall	0.037	-0.087	2.383
Stoke-on-Trent	0.037	-0.079	2.122
Plymouth	0.038	0.013	0.352
Herefordshire, County of	0.040	-0.030	0.757
Luton	0.041	-0.024	0.575
Southampton	0.041	-0.056	1.352
Portsmouth	0.045	-0.055	1.239
Derby	0.045	0.022	0.486
Redcar and Cleveland	0.046	0.104	2.253
Poole	0.048	0.066	1.377
Bury	0.048	0.056	1.166
Knowsley	0.049	0.115	2.355
Sandwell	0.050	-0.089	1.778
Darlington	0.051	0.038	0.740
Isle of Wight	0.060	-0.177	2.931
Rochdale	0.067	-0.019	0.287
Rutland	0.073	0.054	0.741
Greenwich	0.090	0.087	0.962
Hartlepool	0.688	0.032	0.046

Notes: Each row represents a permutation where the stated local authority is assumed to be the treated unit. The columns then show the error in predicted lagged outcomes, the estimated impact estimate in 2017 and the ratio between the two. The critical value of the RMSPE for inclusion in the inference analysis is highlighted (5 x RMSPE for SW Yorkshire), with the true exclusion based on unrounded numbers

Table D6 – Additional details of permutation analysis by local authority for KS2 English

Local Authority	Root Mean Square Prediction Error (RMSPE)	Impact Estimate in 2017	Ratio (Impact /RMSPE)
SW Yorkshire	0.007	0.033	4.596
Hampshire	0.009	0.029	3.166
Lancashire	0.009	-0.011	1.196
Essex	0.010	0.028	2.904
East Sussex	0.011	-0.010	0.964
Medway	0.012	0.049	4.253
County Durham	0.012	0.014	1.167
Ealing	0.012	0.081	6.612
Cornwall	0.013	-0.036	2.903
Dudley	0.013	-0.046	3.636
Liverpool	0.014	-0.084	6.152
Devon	0.014	0.065	4.648
East Riding of Yorkshire	0.014	-0.001	0.067
Lincolnshire	0.014	-0.107	7.554
Nottinghamshire	0.015	0.021	1.364
Warrington	0.015	0.045	2.909
Croydon	0.017	0.073	4.367
Staffordshire	0.017	0.036	2.082

Cumbria	0.017	0.063	3.619
Redbridge	0.017	0.023	1.319
Derbyshire	0.018	-0.020	1.163
West Berkshire	0.018	-0.026	1.426
Oxfordshire	0.018	0.021	1.180
Sefton	0.018	-0.049	2.655
Warwickshire	0.019	-0.041	2.173
Norfolk	0.019	0.015	0.787
Gloucestershire	0.019	-0.028	1.491
Tameside	0.019	-0.075	3.860
Suffolk	0.019	0.029	1.482
North Lincolnshire	0.019	-0.096	4.929
Gateshead	0.020	0.083	4.239
West Sussex	0.020	0.004	0.229
Wolverhampton	0.020	-0.003	0.171
Hertfordshire	0.020	-0.013	0.642
Wiltshire	0.020	-0.031	1.519
Kent	0.021	0.060	2.866
Somerset	0.021	0.048	2.297
Thurrock	0.021	0.077	3.662
Manchester	0.021	-0.010	0.479
Middlesbrough	0.021	0.061	2.857
Hillingdon	0.022	-0.050	2.296
Coventry	0.022	-0.006	0.253
Leicestershire	0.022	-0.017	0.754
Northamptonshire	0.023	-0.056	2.460
York	0.023	0.026	1.150
Cambridgeshire	0.023	0.047	2.053
Oldham	0.023	-0.049	2.127
Wigan	0.023	-0.033	1.432
Cheshire	0.024	0.022	0.929
Southend-on-Sea	0.024	0.089	3.740
Waltham Forest	0.024	0.167	6.939
Worcestershire	0.024	-0.010	0.396
Telford and Wrekin	0.025	0.008	0.311
Birmingham	0.025	-0.019	0.762
South Gloucestershire	0.025	0.014	0.569
Torbay	0.026	0.024	0.919
Blackburn with Darwen	0.026	0.035	1.309
North Yorkshire	0.027	0.052	1.917
Wirral	0.027	-0.083	3.059
North Somerset	0.027	-0.049	1.795
North Tyneside	0.027	0.082	3.012
Slough	0.027	0.034	1.251
Dorset	0.028	-0.005	0.193
Bolton	0.028	-0.041	1.485
Windsor and Maidenhead	0.028	0.057	2.028
Luton	0.028	-0.015	0.538
Brent	0.029	-0.090	3.154
Sandwell	0.029	-0.051	1.781

Redcar and Cleveland	0.029	-0.028	0.975
Swindon	0.029	0.009	0.318
Salford	0.029	-0.063	2.173
Isle of Wight	0.030	0.027	0.880
Barking and Dagenham	0.030	-0.001	0.031
Stoke-on-Trent	0.030	-0.095	3.127
Herefordshire, County of	0.031	0.004	0.116
Walsall	0.032	-0.087	2.729
Hounslow	0.033	-0.068	2.094
Bexley	0.033	-0.052	1.606
Stockport	0.033	0.017	0.525
South Tyneside	0.033	-0.038	1.157
Reading	0.033	0.011	0.325
Stockton-on-Tees	0.033	-0.033	1.003
Merton	0.033	0.066	1.971
North East Lincolnshire	0.033	-0.011	0.318
Bury	0.034	0.022	0.651
Bedfordshire	0.034	-0.053	1.565
Peterborough	0.034	-0.079	2.317
Blackpool	0.034	-0.009	0.274
Newcastle upon Tyne	0.034	0.060	1.752
Northumberland	0.035	-0.031	0.890
Bracknell Forest	0.036	-0.109	3.072
5 x RMSPE for SW Yorkshire = 0.036 (LAs below here excluded from inference analysis)			
Havering	0.036	0.068	1.866
Shropshire	0.037	-0.001	0.020
Portsmouth	0.037	-0.054	1.470
Knowsley	0.037	0.006	0.166
Brighton and Hove	0.037	0.128	3.455
Enfield	0.037	-0.035	0.948
Southampton	0.038	0.081	2.124
Kingston Upon Hull, City of	0.039	0.073	1.888
Bournemouth	0.039	-0.007	0.184
Bath and North East Somerset	0.039	0.094	2.405
St. Helens	0.040	-0.174	4.393
Poole	0.040	0.005	0.137
Bristol, City of	0.040	0.074	1.844
Milton Keynes	0.040	-0.005	0.121
Surrey	0.043	0.061	1.418
Sunderland	0.043	0.104	2.409
Derby	0.043	0.011	0.259
Darlington	0.049	-0.026	0.527
Plymouth	0.052	-0.012	0.224
Rutland	0.052	0.014	0.264
Halton	0.056	-0.049	0.869
Nottingham	0.059	-0.029	0.489
Leicester	0.063	-0.060	0.951
Greenwich	0.094	0.065	0.690
Rochdale	0.103	-0.034	0.333

Hartlepool	0.718	-0.093	0.130
-------------------	--------------	---------------	--------------

Notes: Each row represents a permutation where the stated local authority is assumed to be the treated unit. The columns then show the error in predicted lagged outcomes, the estimated impact estimate in 2017 and the ratio between the two. The critical value of the RMSPE for inclusion in the inference analysis is highlighted (5 x RMSPE for SW Yorkshire), with the true exclusion based on unrounded numbers

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at www.educationendowmentfoundation.org.uk



Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk