Advancing education

Two decades of school trusts improving pupil outcomes

Methodology

Steve Rollett

September 2026

Confederation of School Trusts

The voice of school trusts

Confederation of School Trusts

The Confederation of School Trusts is the national organisation and sector body for academy trusts. Our members educate more than four million children through the dedicated work of their 600,000 staff.

We believe school trusts offer the best model for improving education because they uniquely unite schools in organisations with a strong, single purpose: education. This encourages the sharing of expertise, ideas, and resources to give pupils a better start in life.

We advocate for trusts by helping politicians, policy makers, and the wider public better understand their power to help children. We support trusts through practical advice, strategic guidance, and consultancy. We connect them with each other, and with expert thinking and research.

Published in September 2026 by the Confederation of School Trusts. Registered offices c/o Stone King LLP, Boundary House, 91 Charterhouse Street, London, EC1M 6HR. Registered charity number 1107640.

How to cite this publication: Rollett, S (2026). *Advancing education: Two decades of school trusts improving pupil outcomes - Methodology* London: Confederation of School Trusts

Methodology

1. Purpose and scope

This technical paper explains the methods used in the Trust Impact research. It describes the construction of comparable school-performance measures across successive qualification regimes, the definition of the trust and local authority comparison groups, the approach used to estimate causal effects, the robustness tests applied to those estimates, and the limits of the conclusions that can reasonably be drawn. It should be read alongside the accompanying policy paper.

2. Data and sample

The study combines the Department for Education's published performance data for all state-funded schools in England from 2007 to 2025 with Get Information About Schools records on school and trust membership. Ofsted management information supplies inspection outcomes, while free school meal eligibility is taken from the data published with the performance tables. School-level deprivation is represented by the Income Deprivation Affecting Children Index.

The resulting dataset contains 12,443 schools across 2,303 academy trusts and 9,314 schools maintained by 152 local authorities. There are no national assessment observations for 2020 or 2021 because tests and examinations were cancelled; Key Stage 4 school-level results return in 2022 and Key Stage 2 results in 2023. School histories are linked across changes in Unique Reference Number, including those associated with academy conversion and movement between trusts. Where the same school has results at both key stages in a single year, the two phase observations are averaged rather than selecting one.

3. Measuring performance across a changing assessment system

Direct comparison of raw outcomes across the full period would be invalid. Both Key Stage 2 and Key Stage 4 assessment arrangements changed repeatedly, as did the headline measures reported nationally. A result published in 2010 is therefore not on a common scale with the corresponding headline result in 2024.

We address this by converting each published outcome into a within-year national rank. For every school, the relevant result is expressed as its percentile position among schools in the same phase in that year; the national median is assigned a score of 50. Because the distribution is reconstructed separately for each year, the method does not require different assessment regimes to be placed on a common raw scale. What remains consistent is the question being asked: how did the school stand relative to its national peers at that point in time?

The same ranking procedure produces three indicators. NASattain ranks the headline attainment outcome; NASprog ranks the published progress or value-added outcome; and NAScomp is the mean of attainment and progress where both measures are available.

For school *i* in year *t*, with *N* schools in its phase that year, and for each published outcome *m*: $NAS_m(i,t) = 100 \times rank_m(i,t) / N(t)$

Ties are assigned the average of the ranks the tied schools would otherwise occupy. Applying the formula to attainment and progress produces NASattain and NASprog respectively. Where both are available, NAScomp is their arithmetic mean; where no progress measure is published, the composite takes NASattain alone.

Attainment is treated as the principal outcome and progress as a separate secondary test. NAScomp was constructed as a convenient summary of the two dimensions, but it is less suitable as the main analytical measure because averaging can hide meaningful divergence between attainment and progress. In 30 per cent of school-years, attainment differs from the composite by more than 10 percentile points; for progress the equivalent figure is 38 per cent. For that reason, every analysis was run separately for each outcome.

A central feature of percentile ranking is that, within a single year, it is zero-sum. The mean is 50 by construction: movement upwards by some schools necessarily means that others are overtaken. A fall in NAS should therefore be interpreted as a loss of relative position, not automatically as a deterioration in the school's underlying result.

The longitudinal improvement statistic is different because schools do not all share the same observation window. Trust schools are measured from the period immediately before joining, whereas maintained schools are measured from their earliest available results. Mean change across these differently timed windows is therefore not mathematically constrained to zero. The zero-sum property returns when a common window is imposed. Section 7 does this by starting all schools in 2016; across that comparison the mean movement is +0.25 percentile points, appropriately close to zero. The fact that a trust-maintained difference remains under this common-baseline design is an important check that the main result is not being created by the ranking construction itself.

Section 5 explains how these movements in relative rank are translated into equivalent changes in pupil outcomes.

4. Baselines and the measurement of improvement

Each school's change is calculated from an individual baseline. For trust schools, that baseline is the mean NAS across the two academic years immediately preceding entry to the trust. Two years are used because conversion often follows an unusually weak year; treating that single observation as the normal pre-trust position would risk exaggerating subsequent recovery. Maintained schools have no corresponding intervention date, so their baseline is defined as the mean of their first two available results, usually 2007 and 2008.

For a school joining a trust in year *J*, where *J* is the last full year before joining:

baseline*attain*(i) = ½ [NASattain(i, J) + NASattain(i, J−1)]

and for a maintained school, where e_1 and e_2 are its two earliest years with a result:

baseline*attain*(i) = ½ [NASattain(i, e_1) + NASattain(i, e_2)]

Improvement is measured from that baseline to the school's most recent result, in year *T*:

improvement*attain*(i) = NASattain(i, T) − baseline*attain*(i)

If only one of the two intended baseline observations is available, that single year is used. This affects 647 schools, equivalent to 6.9 per cent of schools with a baseline. Identical baseline rules are applied to NASprog and NAScomp. Unless explicitly stated otherwise, reported results use NASattain.

The unequal baseline construction is the principal limitation of the descriptive comparison. Trust schools are typically observed for about nine years from conversion, while maintained schools have a window of around seventeen years from the start of the dataset. The groups also enter the analysis for different reasons: one around a defined governance intervention and the other simply because observations become available. Consequently, the two descriptive estimates are not answers to precisely the same counterfactual question. A direct comparison implicitly assumes that maintained schools would have followed the same pattern had their measurement window begun later, which these data cannot demonstrate. We therefore use two additional designs: the causal analysis compares treated cohorts and controls over the same calendar years, while the common-baseline analysis gives every school an identical starting year.

5. Descriptive analysis

The descriptive sector estimate is obtained by averaging each school's movement from baseline to its latest available observation across all schools for which a valid baseline can be constructed.

sector improvement = (1/n) **Σ** improvement*attain*(i)

Using NASattain, the mean change is +5.57 percentile points for schools entering trusts and −0.27 for schools remaining maintained. Their mean current positions are 51.33 and 49.73 respectively.

The improvement calculation uses 9,412 trust schools and 8,260 maintained schools with valid baselines. Current standing is calculated over the larger populations with a recent result: 11,338 trust schools and 8,280 maintained schools, including recent joiners for which no pre-trust baseline is yet available. These are therefore different samples and the figures should not be subtracted from one another. Within the trust schools included in the improvement analysis, mean current standing is 52.37 rather than 51.33.

To make the percentile movement interpretable in terms of pupil attainment, we map the observed shift back onto the latest national distribution of the underlying published measure. The value of measure M is identified at the sector's mean baseline percentile and again at its mean current percentile; the difference between those values is treated as the outcome equivalent of the ranking movement.

equivalent = M(p_current) – M(p_baseline)

For any percentile p, M(p) is defined as the median published outcome among schools within one percentile point of that position. Trust schools move, on average, from the 46.8th to the 52.4th percentile on attainment. At Key Stage 2, that distance corresponds to roughly two percentage points more pupils meeting the expected standard in reading, writing and mathematics, equivalent to around 5,400 pupils nationally each year. At Key Stage 4, it corresponds to about 1.5 Attainment 8 points per pupil, or roughly one to two GCSE grades in one subject. Using a threshold measure, the same ranking movement equates to 3.1 percentage points more pupils achieving grade 4 or above in English and mathematics, around 15,800 pupils across the annual trust Key Stage 4 cohort. These are translations of the observed rank movement into equivalent raw outcomes; they are not direct pupil counts.

6. Causal identification

The descriptive estimates show what happened after schools entered trusts, but cannot by themselves establish causation. Schools that convert are not a random sample of all schools, and those entering after a period of decline might have recovered to some extent without conversion. The causal analysis therefore seeks to estimate the counterfactual outcome for the treated schools had they remained outside trusts.

The principal estimator is the staggered-adoption difference-in-differences method of Callaway and Sant'Anna (2021). It is designed for settings in which treatment begins at different times and avoids the problem in conventional two-way fixed-effects models of using already-treated observations as controls. For each conversion cohort and each post-treatment year, the model compares the cohort's change from its pre-joining baseline with the change among never-treated maintained schools over the same calendar period. The control population is restricted throughout to schools that have never held academy status, preventing former academies from entering the maintained comparison group. Cohort effects are then combined using cohort size as the weight. The analysis includes joining cohorts from 2011 through 2020, with multiplier bootstrap used to estimate standard errors.

Four alternative specifications probe whether the estimate depends materially on the construction of the counterfactual or outcome. One adjusts for free school meal rates. A second removes the first year after joining, recognising that transfer itself may create short-term disruption. A third uses outcome regression, predicting each treated school's counterfactual from a model estimated on comparison schools. The fourth reweights comparison schools so that the control group more closely matches the joining cohort in both pre-treatment trajectory and free school meal profile.

Specification	Estimate
Difference-in-differences (primary)	+4.00 (95% CI 3.48 to 4.52)
Adjusted for free school meal rates	+4.14 (95% CI 3.64 to 4.63)
Excluding the first year of membership	+4.42 (supplementary; no interval computed)
Outcome-regression specification	+4.34 (95% CI 3.84 to 4.85)
Comparison group weighted on pre-joining trajectory and free school meals	+6.08 (95% CI 4.38 to 7.85)
Progress measure (secondary outcome)	+3.63 (95% CI 2.80 to 4.37)

The direction of the result is unchanged across the specifications. Every model with a calculated confidence interval has an interval wholly above zero; the variant excluding the first year is supplementary and no interval was computed. Reweighted models generate larger point estimates than the unweighted estimator, but this is not interpreted as showing that the primary estimate is a lower bound. Reweighting can improve comparability, but it can also increase extrapolation, concentrate influence on a relatively small number of controls, or alter the effective sample. The unweighted difference-in-differences estimate is therefore retained as the headline result. Alternative specifications are reported as sensitivity tests, and none is treated as defining either a minimum or maximum effect.

Falsification and robustness

Three falsification and robustness exercises test whether the estimated effect could be an artefact of the design. First, an in-time placebo moves each joining date three years earlier, when no conversion had occurred; the resulting estimate is only 0.002 of the real effect. Second, a leave-one-cohort-out exercise recalculates the model after removing each joining cohort in turn. The full spread of those estimates is 1.15 percentile points, indicating that the headline result is not dependent on a single cohort. Third, examination-cohort sizes are compared for treated and control schools over the five years after conversion. The two series remain closely aligned, with the difference below one index point, providing no indication that the attainment effect is explained by a differential change in which pupils enter the measured cohort.

Treatment timing presents a further identification challenge. Schools often enter trusts after losing ground relative to their peers, a pattern described in the literature as Ashenfelter's dip. A formal pre-trend test rejects identical pre-treatment trajectories between trust and comparison schools. Difference-in-differences reduces, but cannot eliminate, the resulting concern: the post-treatment comparison is made against maintained schools observed during the same years rather than against the treated schools' own earlier decline, and the trajectory-matched specifications address the issue more directly. Any remaining bias depends on the unobserved path the treated schools would have followed had they not joined a trust.

7. Further analyses

Common baseline

A common-baseline analysis removes the different starting dates used in the descriptive model. Every school is measured from 2016 to its latest available result and is grouped according to current status. Trust schools gain 2.53 attainment percentile points over that window, while maintained schools lose 2.70. However, the groups were already different in 2016: schools subsequently in trusts averaged 49.6 compared with 53.2 among schools that remained maintained. We therefore repeat the comparison by matching each trust school to maintained schools in the same phase, within two percentile points of its 2016 rank and in the same FSM quintile. The matched difference is 1.71 points (95 per cent confidence interval 0.93 to 2.47), with 97.7 per cent of trust schools successfully matched. The finding persists with a narrower matching caliper and when pre-2016 trajectory is added. Around two thirds of the unmatched sector difference disappears after matching, indicating that much of the raw gap reflects different starting compositions.

Rate of improvement

A second supplementary design examines the slope of each trust school's trajectory before and after conversion. Ordinary least squares is fitted separately to the school's own pre- and post-joining history. The average annual rate is 0.26 percentile points before joining and 0.49 afterwards. Trust schools were therefore already improving somewhat faster than maintained schools before conversion; the relevant finding is the subsequent acceleration at the point of joining, not the pre-treatment difference itself.

Entry route

The independent review recommended separating the two principal routes into the trust sector because their starting conditions differ substantially. Converter academies chose to convert. For this group, the primary estimate is 1.29 percentile points (95 per cent confidence interval 0.76 to 1.80); the trajectory- and FSM-weighted specification produces 3.59. Consistent with Section 6, the plain estimate remains the primary result and the two specifications are not interpreted as lower and upper bounds.

Sponsor-led academies cannot be evaluated against the same maintained-school counterfactual. Since 2016, maintained schools judged inadequate have been legally required to enter a trust, so an equivalent population of inadequate schools cannot remain untreated. Maintained schools with similar attainment are also not equivalent on the regulatory condition that generated sponsorship. For the 2011–2015 sponsored cohorts, the comparison group is therefore formed from schools that would later be sponsored but had not yet been. Relative to that group, sponsored academies improve by 15.82 percentile points (95 per cent confidence interval 14.79 to 16.94). The estimate reflects both recovery among sponsored schools and continuing decline among the not-yet-sponsored group, and both components are reported. Their pre-treatment trends are statistically indistinguishable, while the placebo estimate is 0.03 of the real effect. The relevant counterfactual is consequently sponsorship at a later date, not a hypothetical state of never being sponsored.

Trust structure

For the analyses reported in this section, uncertainty is estimated using a both-sides school-level cluster bootstrap: each school receives one bootstrap weight that is retained across all of its observations. The great majority of academies are members of multi-academy trusts, with only a small and declining standalone sector. Restricting treated schools to converters, so that the estimate is not influenced by the turnaround of inadequate schools, we compare schools joining groups with otherwise comparable schools outside groups - maintained schools and standalone academies combined. The plain estimate is 1.41 percentile points (95 per cent confidence interval 0.92 to 1.90), rising to 2.70 under the weighted specification. Against maintained schools alone, the corresponding estimates are 1.52 and 3.09. The similarity of these pairs suggests that academy status without group membership accounts for only a small part of the overall estimated effect.

Two additional analyses place limits on what can be inferred about the group mechanism. First, the prior trajectory of the schools already within a trust does not predict the subsequent improvement of a school joining that trust. Second, the analysis of schools that were already academies before entering a multi-academy trust is inconclusive: the two comparator constructions produce effects with opposite signs. We therefore cannot establish a distinct incremental effect of group membership for an existing academy. The group estimate above should instead be understood as combining membership of the multi-school organisation with the governance change that occurs at conversion.

Deprivation

The distribution of effects by deprivation shows where the largest sector gains occur. Estimating the model separately within each deprivation quintile, with trust and maintained schools compared inside the same quintile and calendar years, produces positive estimates throughout. The effect decreases monotonically with affluence: 4.57 percentile points in the most deprived quintile and 2.17 in the least deprived, with non-overlapping intervals at the two extremes.

A pre-specified converter-only analysis clarifies the source of that gradient. Once sponsor-led academies are removed, the estimates from most to least deprived are +1.23, +0.37, +0.88, +1.47 and +1.63. The gradient therefore disappears and slightly reverses; only the two least deprived quintiles remain individually significant. This indicates that the sector-level deprivation pattern is principally compositional. Sponsored academies, whose estimated effects are much larger, are disproportionately located in disadvantaged areas: 36 per cent are in the most deprived quintile compared with 17 per cent of converters. The appropriate interpretation is therefore that trusts have delivered their largest gains in the most disadvantaged communities largely through the sponsorship route used to address entrenched underperformance. It is not evidence that deprivation itself necessarily magnifies the treatment effect for an otherwise equivalent school. Progress results show a related but less orderly pattern: +6.48 in the most deprived quintile and +3.65 in the least, with intermediate estimates of +4.69, +3.04 and +1.63.

8. Limitations

The estimates should be interpreted subject to the following limitations.

- Baseline asymmetry: The descriptive estimates necessarily use different observation windows because only trust schools have a conversion event. This is the most important constraint on the raw sector comparison. The causal and common-baseline designs test the finding without relying on that same asymmetry.

- Non-linearity: Improvement need not proceed at a constant rate. Schools entering trusts from a weak position may make rapid gains soon after conversion and then level off, so a conversion-anchored window can capture a faster phase than the longer maintained-school series. Annualising the figures would not solve this: it would assume linear change while leaving different starting positions untouched.
- Selection: Schools are not assigned randomly to trusts. The causal models compare schools over common periods, include specifications that account for prior trajectory, and survive the falsification tests, but unobserved differences between treated and comparison schools cannot be excluded.
- Measurement: Percentile rank turns the underlying result into an ordinal position and can therefore be affected by floor and ceiling effects. A robustness analysis replacing ranks with normal scores yields materially identical standardised effects. Further analysis using raw outcomes and z-scores alongside ranks remains a priority. The present estimates should always be read as changes in relative national standing, not absolute attainment.
- Intake: FSM eligibility is used to test whether pupil composition explains the estimated effect, but it is only a partial proxy. It captures one important dimension of disadvantage rather than the full range of differences between school intakes.
- Composite measure: Published progress measures end after 2023 for Key Stage 2 and 2024 for Key Stage 4. Recent composite observations therefore contain attainment alone whereas earlier baselines can combine attainment and progress. This has negligible impact on the sector estimate, but composite results are not reported for subgroups.
- Data coverage: No national assessment results are available for 2020 or 2021. In 2022, only secondary school outcomes were published and the remaining maintained secondary sector was already small. Comparisons confined to that year consequently depend on an unrepresentative maintained sample.
- Variation: These are average effects across thousands of schools, not predictions for individual institutions. Some trust schools do not improve, and this study does not identify which trusts or which organisational and improvement practices generate the observed average effect.

9. Verification and independent review

The analytical pipeline writes a validation record at each stage, testing outputs against expected ranges, sample sizes and internal cross-checks. Every result appearing in the policy paper can be traced to one of these records. The supplementary analyses in Section 7 were specified in writing before results were calculated, and their outcomes are reported irrespective of whether they support the main hypothesis, including null findings.

An independent academic expert reviewed both the methodology and the underlying analysis. That process led to several substantive changes: attainment and progress were treated as separate outcomes rather than relying on the composite; entry routes were analysed separately; the common-baseline exercise was extended to school-level matching; and greater care was taken in interpreting differences between alternative specifications. The trust-structure and deprivation analyses, together with the two measurement limitations described in Sections 5 and 9, were developed after that review.

As a further verification step, the analytical pipeline was rebuilt from the verified data foundation and executed independently of the original run. This reproduced every reported figure in the attainment and progress pipelines exactly. The additional Section 7 analyses were conducted later and are supported by their own written pre-specifications and validation records.

10. Use of AI assistance

In line with Resnik & Hosseini, we declare for transparency the use of AI in this work.[1] The analysis pipeline for this paper, including the Python scripts used to compute all figures from source data, was developed with the assistance of Claude, an AI assistant produced by Anthropic. The code used to generate the underlying school and trust level data, and the

1 Resnik, D B, & Hosseini, M (2026) 'Disclosing artificial intelligence use in scientific research and publication'. *Accountability in Research*, 33(2).

ranking calculations, were validated by an external independent expert validator specialized in AI coding. Subsequent calculations were independently verified by an academic expert using manually reviewable files (Excel) prior to publication, including the replication of calculations from scratch. Methodological decisions, interpretation of findings, and editorial judgements were made by the author. As such, while this significant body of work has been supported by AI, it has been subjected to verification and has been human led at each stage.

Confederation
of School Trusts

cstuk.org.uk