National Institute of Economic and Social Research

ACTION TUTORING'S SMALL-GROUP TUITION PROGRAMME

An impact evaluation using statistical comparison groups

Paolo Lucchino

Date: March 2016

About the National Institute of Economic and Social Research

The National Institute of Economic and Social Research is Britain's longest established independent research institute, founded in 1938. The vision of our founders was to carry out research to improve understanding of the economic and social forces that affect people's lives, and the ways in which policy can bring about change. Seventy-five years later, this remains central to NIESR's ethos. We continue to apply our expertise in both quantitative and qualitative methods and our understanding of economic and social issues to current debates and to influence policy. The Institute is independent of all party political interests.

National Institute of Economic and Social Research
2 Dean Trench St
London SW1P 3HE
T: +44 (0)20 7222 7665
E: enquiries@niesr.ac.uk
niesr.ac.uk
Registered charity no. 306083

This paper was first published in March 2016

ACTION TUTORING'S SMALL-GROUP TUITION PROGRAMME

Paolo Lucchino

Acknowledgements

The author would like to thank the Action Tutoring team for the support provided throughout the evaluation; the Centre for Social Action Innovation Fund for funding this evaluation; and the National Pupil Database team for helping with the data request and granting access to the data. The author is grateful to Richard Dorsett (NIESR) for comments received. All errors and omissions are of the author.

Contact details

Paolo Lucchino (p.lucchino@niesr.ac.uk), National Institute of Economic and Social Research, 2 Dean Trench Street, London SW1P 3HE

Table of Contents

Summary of findings

This evaluation aims to estimate the impact of Action Tutoring's small-group tuition programme on GCSE grades. During the 2014/15 academic year, Action Tutoring provided 6-8 weeks of tutoring in Maths or English to pupils on the C/D borderline. The intervention was directed at schools with more than double the national average of pupils eligible for Free School Meals.

The ambition of this evaluation is to estimate impacts that could be causally attributed to the programme. We attempt to do so using quasi-experimental methods that reply on the assumption of *selection on observables*. These consist of applying statistical methods to construct a comparison group of non-tutored students with characteristics that are as similar as possible to those found among tutored students. To the extent that characteristics that drive programme participation and/or influence outcomes are successfully balanced across treatment and comparison groups, any difference in outcomes would identify the impact of Action Tutoring's programme.

In this section, we highlight the main findings of the evaluation, discuss possible interpretations and avenues for further research.

- We find evidence of positive and statistically significant impacts of the programme on GCSE grades. When considering students tutored in any subject, we estimate the programme positive impact on GCSE point scores of about 2 points. This equates to one third of a letter grade.[1] The impact when considering students tutored specifically in Maths and English is around 2-2.5 points and 1.5 points respectively.
- We fail to find evidence that the programme affects the probability of achieving a grade C or above in the tutored subjects. While we typically find small positive point estimates in the order of 0 to 3 percentage points, these are statistically insignificant, suggesting no impact. Estimates remain statistically insignificant when considering each subject separately and point estimates for Maths are again higher than for English.
- Two possible explanations can help reconcile these apparently inconsistent results. It could be that the positive impacts on GCSE points primarily materialise below or above the C threshold. This is likely to occur if schools did not solely target students on the C/D borderline. Additionally, the evaluation is likely to have been insufficiently powered in relation to the grade C outcome to attribute statistical significance where point estimates are positive. Revisiting this result in future evaluations may be useful.
- These headline results are robust across all 3 of the chosen statistical procedures, and 3 different model specifications.
- We find a strong positive association between the number of tutoring sessions attended and estimated impacts. For example, compared to the whole sample estimate of 2 points, students attending at least 7 sessions are found to have GCSE points that are 3 points higher than the comparison group. However, our evaluation approach rests on the selection on observables assumption and it is likely that the number of sessions attended will be correlated with unobserved personal characteristics such as motivation and commitment. One should therefore be cautious of giving this result a causal interpretation. Further

[1] GCSE Points are a numerical representation of letter grades, where the difference between any two letter grades is 6 points.

evaluations designed specifically to estimate the impact of differing degrees of tutoring would be a promising area of further research.

- There appears to be a strong variation in the estimated effects by broad geographical region. Being able to explain such geographical heterogeneity would appear to be a promising area for further enquiry both internally by Action Tutoring and through future external evaluations. Possible hypotheses to explore could be the quality of tutors, the quality of tutor-tutee match, and the prevalence of alternative tutoring initiatives in each area.
- We find that the effect of the programme is broadly similar across gender and eligibility for Free School Meals.

Finally, the quasi-experimental design adopted here allows for causal inference on the assumption of selection on observables. As discussed in the report, there are a number of features that make Action Tutoring's programme well suited for such an evaluation approach. However, as is always the case, assumptions can be queried. Further evaluations using methodologies that require less stringent assumptions, notably through a randomised controlled trial, may be worth considering to assess the robustness and replicability of these results.

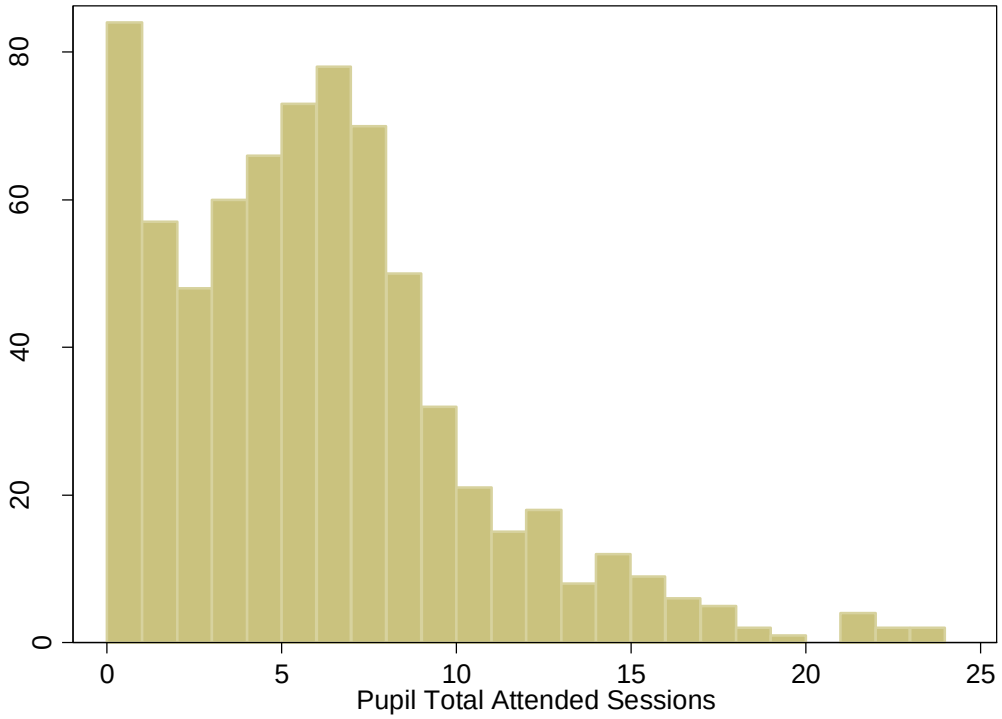
Description of the Action Tutoring programme

The Action Tutoring programme provides tutoring in Maths or English for GSCE students. It targets schools with more than double the national average of pupils eligible for Free School Meals.[2] During the 2014-15 academic year, a tutoring programme consisted of 6-8 weekly 1 hour sessions over the course of a term.

Subject to Action Tutoring's criteria that students must be on or near the C/D borderline, schools select pupils they deem are most in need of additional support. In some cases, schools run multiple programmes across consecutive terms, and often retain the same pupils across terms.

Attendance to tutoring sessions is presented as compulsory by schools. Ultimately, however, participation by the student is voluntary, and compliance is partial. Figure 1 - Distribution of number of sessions attended by treated students in our sample. Just under 20% of students attend do not attend any session, or attend only one. However, we see a clear concentration of students around the target 6-8 sessions. Just under half of students attend at least 6 sessions, thereby completing, or near-completing, at least one full programme.

Figure 1 - Distribution of number of sessions attended



Action Tutoring recruits tutors to deliver the sessions. Tutors are required to be educated to or working towards a degree or have other relevant qualifications/experience. A large proportion of the tutors are students, although Action Tutoring does not prioritise any particular group.

Identification strategy

[2] The evaluation focuses on the 2014-15 academic year. From 2015-16, Action Tutoring has been working with a wider range of schools, and prioritising Pupil Premium students within these.

The chosen evaluation strategy consists of quasi-experimental approaches relying on the *selection on observables* assumption. These 'matching'-type approaches were identified as the solution that best balanced rigour and compatibility with Action Tutoring's current operations.[3] The general objective of our approach is to use data on individuals that have not been effected by the programme to construct a statistical comparison group that is as similar as possible to programme participants in relation to characteristics thought to affect outcomes. A comparison of outcomes between these two groups then identifies the effect of the programme.

The fundamental assumption underlying a matching estimator is that the treatment must be *strongly ignorable* (Rosenbaum and Rubin, 1983). This assumption is satisfied when two main conditions are met. Firstly, because the objective is to ensure that characteristics associated with programme participation and outcomes are balanced across treatment and comparison group, these characteristics must be observable to the researcher. This condition, referred to as *selection on observables*, is not a trivial requirement. It may be the case that a large number of factors determine individuals' participation in a programme. If so, a matching approach would be successful in identifying a casual impact estimate only if the researcher can confidently argue that all such determining characteristics, to the extent that they may affect outcomes, have been balanced across treatment and comparison group. Alternatively, it may be the case that what determines participation and/or outcomes are personal attributes that are fundamentally unobservable. By definition, it would therefore not be possible to statistically balance these attributes across treatment and comparison group, and a matching approach would be at risk of delivering biased estimates.

This first condition makes matching approaches relatively more suited for programmes that have simple and explicit eligibility criteria. When this is case, the eligibility criteria clearly and observably characterise programme participants, as well as help identify eligible individuals who had no involvement with the programme. In particular, the latter group would make suitable comparison group if the reason for their non-participation is not under their control or not determined by their choice.

The above leads naturally to the second condition necessary for strong ignorability. This requires that, for each type of individual, it must be possible to observe some that do and some do not participate in the programme. In other words, the data must satisfy the *overlap condition* (also known as common support condition). Quite simply, if all individuals with a given set of characteristics participate in the programme, it will not be possible to find a match for them among those that did not participate.[4]

Finally, while not technically a requirement, it is particularly important for the matching approach to be able to match individual on baseline outcomes pre-programme. In many cases, baseline levels will very likely influence whether an individual selects (or is selected) into a programme aiming to improve these levels. If this is the case, by virtue of the selection on observables condition discussed above, these must be included among the characteristics one matches on. However, even if this is

[3] More precisely, as discussed later in the report, we use a matching and reweighting approach. In the text, we refer loosely to both as *matching*. This intended to refer to their common aim of making treatment and comparison groups similar through the use of a statistical procedure.

[4] Absence of the overlap condition for a given type of individuals in the treatment group does not affect the ability to construct a comparison group for other types of individuals in among the treated.

not the case, matching on pre-treatment levels is likely to substantially reduce the heterogeneity in outcomes among individuals in the comparison group, thereby increasing statistical power.

Considering the above, we argue that a matching-type approach offers the best balance between maintaining a high level of rigour and minimising operational burden in the context of Action Tutoring's programme. Indeed, the programme has a number of characteristics that lend themselves well to such approaches.

Firstly, the program targets pupils on the C/D borderline, in schools where the share of students eligible for Free School Meals is more than twice the national average. These provide helpful eligibility criteria to use when identifying suitable candidates (in suitable schools) for the comparison group. Additionally, as the eligibility criteria is expressed in terms of pre-programme attainment levels, the matching approach will by necessarily match and balance baseline levels across treatment and comparison groups.

Secondly, because of operational constraints, Action Tutoring does not have a presence in every eligible school in a given area. It follows that data will be available for students on the C/D borderline attending local schools that do not participate in the programme. This helps satisfy the overlap condition. Finally, to the extent that we can assume that Action Tutoring's presence in one eligible school over another is largely random, or at least unrelated to pupils' attainment levels, we can argue that the non-participation of comparison students will be more likely driven by lack of access rather than self-selection.

While keeping these advantages in mind, our identification strategy must nevertheless address the fact that participation into the programmes is likely to be also driven by unobservable characteristics. The effect of these may occur at two main points in the programme. Firstly, as it is teachers who select pupils to be offered tutoring, one may be concerned that this selection is based on information known to the teacher but not the evaluator. For example, teachers might select students whom they believe might benefit most from tutoring from among a group of students otherwise identical on paper. As the evaluator does not share this insight, it is not possible to match participants with comparison students holding the same promise, thereby determining an upwards bias in our estimates. We argue that this is unlikely to be a major concern in the context of Action Tutoring's programmes, as we understand that schools typically tend to offer tutoring to all eligible students in their pool.

Secondly, students offered tutoring ultimately differ in the number of sessions they attend (and some do not attend at all). It may be the case that those who attend more sessions could be more motivated or organised than the general eligible student population used to construct a comparison group. Again, it would not be possible to account for these unobservable differences with a matching process, thereby generating a risk of an upwards bias. We address this concern by conducting an Intention-To-Treat (ITT) analysis. In other words, we consider as 'treated' any student who has been *offered* tutoring, regardless of the extent to which they attend tutoring sessions. By doing so, we bypass the concern of what, if any, unobservable characteristics are driving attendance. By focusing our comparison between the eligible pupils who have been offered tutoring and eligible pupils who did not have access to the programme, we identify the average effect of tutoring on those who were *offered* the opportunity. As we know that a significant minority of students offered tutoring addend few or no sessions, our estimate is likely to be a lower bound of the effect of the programme on those who actually attend.

Finally, we cannot rule out that C/D borderline students in comparison schools may have access to other tutoring programmes similar to the ones offered by Action Tutoring. To the extent this is the case, again, our estimate will be a lower bound estimate of the true impact of Action Tutoring's programme.

Our evaluation approach allows for causal inference on the assumption of selection on observables. This was chosen both in light of a number of features that make Action Tutoring's programme well suited for it, as well as to minimise the operational burden of conducting the evaluation. As discussed above, there are reasons why the selection on observable assumption may not apply. If this is the case, it could determine a (positive or negative) bias on the impact estimates. While aware of these, we nevertheless argue that the approach chosen can at least aim for causal inference and is fit for purpose given the Action Tutoring's current evaluation requirements and priorities. Evaluations using methodologies that require less stringent assumptions, notably through a randomised controlled trial, may be worth considering in future evaluation efforts to assess the robustness and replicability of these results.

Data Sources

The evaluation's core dataset is the National Pupil Dataset (NPD). The NPD is the Department for Education's administrative data source containing information on characteristics and test results of pupils across England. As such, its strengths for the purposes of our evaluation are that: it offers standardised outcome measures across treatment and comparison groups; provides high levels of data quality and sample size; and minimise the need for primary data collection.

Specifically, our data consists of the Key Stage 4 outcomes for candidates sitting exams at the end of the 2014-15, as well as previous attainment histories going back to Key Stage 1. The dataset also includes several pupil characteristics of interest, including gender, age, eligibility to Free School Meals (FSM) and English as an Additional Language (EAL) status. The sample is limited to schools in districts or local authorities where there is at least one school participating in the Action Tutoring programme. This minimises the differences in neighbourhood contexts across treatment and control groups. Additionally, we also make use of the 2015 school census, detailing school characteristics of interest, for example the percentage of pupils eligible for FSM.

A very limited set of data on the tutored students was matched onto the NPD. This consisted primarily of the subject being tutored in and the number of sessions attended. Matching any external data to the NPD required us to obtain opt-out consent. We chose not to seek opt-in consent as we felt natural inertia may have led to an excessive reduction in sample size. As such, we were unable to request data on pupil characteristics deemed sensitive or personal (e.g. ethnicity) for the treatment group. By extension, we did not request these for the comparison group.

The final dataset consists of 724 tutored GSCE students, 59,484 GSCE students in other schools in the same district/local authorities, as well as 4,310 non-tutored GSCE students in schools participating in the Action Tutoring programme. This latter group is excluded from the matching process as we cannot rule out that they may be affected by spillover effects from tutoring of their classmates.

The first two columns in Table 1 present some summary statistics describing the average characteristics of pupils in the tutored group and those in the rest of the NPD sample on which we carry out the analysis. We see that tutored students have statistically significant lower previous attainment levels, are more likely to be female, on FSM and have EAL, and to attend a school that is

smaller and located in more deprived areas. Mean GSCE point scores are similar across treated students and the wider student population but the share achieving at least a grade C is significantly lower among treated students.

Implementation of the matching process

In this section, we detail our approach to constructing a comparison group, and report on the relative success of alternative statistical methods. We construct the comparison group in two steps. Firstly, we manually exclude a number of observations in the potential comparison group that are systematically different from those in the treatment group. Secondly, we test three different statistical procedures to ensure that the characteristics of the comparison group are made to be as similar as possible to those among the treated.

We start by removing from the sample of non-tutored students those whom exhibit characteristics that are systematically different from those among the totality of the tutored group. For example, as none of the tutored students are boarders or part-time students, we exclude any such pupils from the potential comparison group. These total only 34 observations.

We take a similar approach with school characteristics. Specifically, as all schools participating in the programme were either Comprehensive, Modern or Maintained Special, we exclude schools of any other type from the potential comparison group. As all participating schools are in urban areas, we also exclude a small number of schools in rural areas from the potential comparison group.

Comparing the second and third columns of Table 1 shows the effect of these manual exclusions on the average characteristics of pupils in the comparison group. Overall, these manual restrictions do not have substantial effect on the pre-matching difference between treated individuals and candidates for the comparison group. This confirms the need to apply statistical techniques to further balance characteristics treated and comparison groups. Notably, however, the difference in mean GSCE point scores between treated and comparison groups is now negative and statistically significant.

Table 1 - Descriptive statistics

Characteristic	Treatment group	Whole potential comparison group		Trimmed potential comparison group	
KS4 Points	36.4	36.74		37.34	**
% achieving >=C at KS4	0.52	0.6	***	0.64	***
KS3 Level	5.18	5.57	***	5.57	***
KS2 Level	2.44	2.59	***	2.55	**
EAL	0.48	0.28	***	0.32	***
FSM	0.38	0.22	***	0.26	***
Gender	0.59	0.49	***	0.5	***
KS4 cohort size	317.57	374.13	***	374.2	***
School % FSM	28.58	21.57	***	21.56	***
IMD Rank	9198.85	9956.2	***	9954.21	***

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

We then proceed to applying statistical techniques to redefine the comparison group in a way that its characteristics match those of the treatment group as closely as possible. This involves choosing the characteristics that must be comparable across the two groups and then type of estimator to use.

Considering the eligibility criteria of the Action Tutoring programme and best practices for matching approaches, we place a lot of effort on matching on previous attainment as closely as possible. This approach also makes best use of the NPD's strengths. Indeed, while matching approaches often emphasise the need to match on several different factors, this is limited by the relatively small set of individual characteristics in the NPD. On the other hand, the NPD's main advantage is its coverage of the full student population. As such, we exploit this by attempting to compare each tutored student with students with exactly identical attainment histories in the subject tutored in. In most cases, we are able to achieve this all the way back to Key Stage 1, and in all cases this is ensured going back to at least Key Stage 2. Specifically, we create attainment groups for each possible combination of Key Stage 1, 2 and 3 outcomes. Where a tutored student exhibits an attainment history that finds no direct comparison among un-treated students, they are assigned to the largest group that ensures an exact match on Key Stages 2 and 3. As discussed below, we then use estimators that either ensure that comparisons are made only within such groups, or that place a significant more weight on comparisons within groups.

We run three model specifications. All three seek to balance previous attainment between treatment and comparison groups. The first model additionally seeks to balance the school percentage of students eligible for FSM, thereby covering the full set of Action Tutoring eligibility criteria. Model two extends the first specification by comparing across additional individual and school characteristics: gender, individual FSM eligibility, English as an Additional Language (EAL) and the size of the Key Stage 4 cohort. While the comparison group consists only of students in districts/local authorities where Action Tutoring has a presence, comparison of students can still occur across these. The third model therefore adds neighbourhood characteristics, specifically the local Index of Multiple Depravation Rank.

Finally, we consider three different statistical approaches to achieving balance over the variables of interest across treatment and comparison groups. A number of different such statistical procedures exist. Despite the popularity of propensity score matching, recent simulation study testing the performance and reliability of a large number of such estimators indicates that alternative methods provide more precise and reliable results (Frölich et al., 2015). We therefore selected the approaches used in the evaluation by considering these rankings, as well as the feasibility and ease of implementation these in our statistical software, STATA.

The first approach taken consists of a single nearest neighbour matching algorithm. A measure of distance is constructed for each pair-wise combination of treated and potential candidate from the comparison group, and the closest match for each treated observation is selected for inclusion in the comparison group. The distance metric is the Mahalanobis distance, which has the advantage of being scale-invariant and implicitly taking into account of correlations across variables. In our specific implementation, an exact match on previous attainment group is imposed. This means that, for each treated student, the nearest neighbour is chosen only from comparisons students with the same attainment history.

Our second implementation is Inverse Probability Weighting (IPW). Intuitively, IPW keeps all of the observations in the comparison group, but reweights observations so that the average of characteristics in the re-weighted comparison group match those of the treatment group. Observations that are more similar to those in the treated group are given a higher weight, and vice versa. The revised weights depend on an estimate of the probability of being treated. It is easy to implement, and does is not demanding on computational resources. While it can be sensitive to the misspecification of the propensity score (in particular for observations with a score close to zero), simulations in Frölich et al. (2015) indicate it performs well compared to other alternatives.

Table 2 evaluates the relative performance of these two methods by exploring the extent to which target characteristics are balanced across treatment and comparison groups after their application. For each model specification, it displays the average values of the variable of interest across treatment and comparison groups after the procedure has been applied, and reports stars identifying the statistical significance of the difference in these two means. The tables also report the standardised percentage bias for each variable (Rosenbaum and Rubin, 1985). This measures the difference of the sample means in the treated and comparison groups as a percentage of the square root of the average of the sample variances in the groups. The average of the standardised percentage bias across all variables considered is reported in the final row of each panel in the table. This metric gives an immediate evaluation of the degree of balancing achieved by the estimator.

Table 2 - Comparison of statistical approaches

	Nearest neighbour matching				Inverse probability weighting			
Characteristic	Comparison group mean	Treatment group mean		% bias	Comparison group mean	Treatment group mean		% bias
Model 1								
KS3 Level	5.19	5.18		-0.8	5.19	5.18		-0.7
KS2 Level	2.45	2.44		-0.3	2.44	2.44		0.6
EAL	0.35	0.48	**	26.7	0.39	0.48	**	18.2
FSM	0.37	0.38		2.5	0.35	0.38		5.3
Gender	0.52	0.59	**	15.5	0.48	0.59	**	23.2
KS4 cohort size	325.48	317.57		-8.6	352.29	317.57	**	-33.2
School % FSM	28.54	28.58		0.3	28.67	28.58		-0.7
IMD Rank	7240.92	9198.85	**	29.2	7810.12	9198.85	**	19.9
Average across all				10.6				12.5
Model 2								
KS3 Level	5.29	5.18		-14.7	5.18	5.18		0.4
KS2 Level	2.77	2.44	*	-39.7	2.43	2.44		1.1
EAL	0.32	0.48	*	31.7	0.48	0.48		0.3
FSM	0.40	0.38		-4.4	0.38	0.38		-0.4
Gender	0.58	0.59		2	0.59	0.59		0.4
KS4 cohort size	318.82	317.57		-1.9	315.06	317.57		2.5
School % FSM	32.47	28.58		-23.9	28.74	28.58		-1.1
IMD Rank	6593.94	9198.85	*	44.3	7519.13	9198.85	**	24.5
Average across all				19.3				3.6
Model 3								
KS3 Level	5.29	5.18		-15.1	5.18	5.18		0.8
KS2 Level	2.65	2.44		-23.4	2.43	2.44		1.3
EAL	0.37	0.48		21.9	0.48	0.48		0.3
FSM	0.53	0.38		-29.2	0.38	0.38		-0.6
Gender	0.69	0.59		-19.3	0.59	0.59		0.5
KS4 cohort size	319.05	317.57		-2.3	314.46	317.57		3.1
School % FSM	38.18	28.58	**	-54.7	28.84	28.58		-1.8
IMD Rank	7366.57	9198.85		29.2	9264.99	9198.85		-0.9
Average across all				25.9				1.1

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

Table 2 clearly shows the Inverse Probability Weighting approach performs best. Difference in means across groups become statistically insignificant when included in the model specification, and the average percentage bias declines and reaches very low levels as the model specification covers more and more variables. The nearest neighbour matching approach also tends to achieve a statistically insignificant difference in means in the target variables following matching. However, a clear trade-off emerges as matching is sought on more and more variables. In that case, seeking to

balance one variable by selecting a given neighbour can cause the balance on other variables to deteriorate, pushing the estimator to seek an imperfect balance across the difference variables. Nevertheless, we still apply both methods when estimating impacts to check the robustness of the results across different approaches.

Finally, as a further robustness check of our results, we also use a standard regression with attainment group fixed effects and controlling for additional covariates for comparison. By including attainment group fixed effects, it essentially reports the difference in mean outcomes across tutored and non-tutored students when exclusively comparing individuals with the same attainment histories. Additional covariates are then added to the regression specification.

Results

The evaluation focused on two main outcomes of interest. Firstly, we estimate the effect of tutoring on the average GCSE point score in the subject tutored in. We use the highest point score obtained in GCSE Maths and GCSE English as recorded on the NPD. For the case of English, this refers to the highest grade across English GCSEs attempted (i.e. across English Language, English Literature & English Language & Literature). Secondly, we assess the impact of the tutoring programme on the probability that the student obtains a grade C or higher.

Results are presented in Table 3. Estimates using each of the three chosen statistical approaches are displayed across the columns. Going down the rows, the table is split into two panels, presenting results for GCSE point scores and the probability of obtaining a grade C or higher respectively. Within these, we present results across the three model specifications used. Standard errors in brackets can be found below each point estimate. Stars highlight the statistical significance of each estimate. By presenting the full set of results, we are able to assess the robustness of the impact estimates to variations in specifications and approaches, thereby strengthening our confidence in the findings.

Table 3 - Main results

Estimates across all subjects tutored			
	Regression	**NN matching**	**Inverse probability weighting**
GCSE Points			
Model1	2.202 **	2.027 **	2.129 **
	[.538]	[.458]	[.309]
Model2	2.058 **	1.42 **	1.98 **
	[.519]	[.501]	[.311]
Model3	1.98 **	1.988 **	2.075 **
	[.544]	[.505]	[.317]
Prob (>= Grade C)			
Model1	0.012	0.004	0.008
	[.019]	[.023]	[.017]
Model2	0.006	0.031	0.001
	[.019]	[.025]	[.017]
Model3	0.003	0.02	0.001
	[.019]	[.025]	[.017]

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

A couple of results emerge clearly and consistently across the various specifications. Specifically, GCSE point scores of tutored students are estimated to be a statistically significant 2 points higher than those among the comparison groups. On the other hand, the probability of achieving a grade C or higher is estimated to be no different among the treated and comparison groups. While point estimates are consistently positive in the range between 0 and +3 percentage points, they are statistically insignificant across all specifications.

Table 4 - Main results by subject tutored

Estimates for students tutored in Maths

	Regression	NN matching	Inverse probability weighting
GCSE Points			
Model1	2.549 **	2.433 **	2.418 **
	[.696]	[.616]	[.408]
Model2	2.301 **	1.352 *	2.169 **
	[.683]	[.617]	[.415]
Model3	2.264 **	2.522 **	2.211 **
	[.702]	[.647]	[.418]
Prob (>= Grade C)			
Model1	0.018	0.013	0.012
	[.026]	[.028]	[.022]
Model2	0.01	0.032	0.003
	[.026]	[.031]	[.022]
Model3	0.007	0.032	0.002
	[.026]	[.032]	[.022]

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

Estimates for students tutored in English

	Regression	NN matching	Inverse probability weighting
GCSE Points			
Model1	1.59 *	1.237 *	1.56 **
	[.85]	[.616]	[.441]
Model2	1.55 *	1.527 *	1.561 **
	[.797]	[.856]	[.438]
Model3	1.381	1.438 *	1.804 **
	[.876]	[.822]	[.465]
Prob (>= Grade C)			
Model1	0.003	-0.012	0
	[.024]	[.037]	[.028]
Model2	-0.001	0.019	-0.005
	[.024]	[.04]	[.028]
Model3	-0.006	0.015	-0.001
	[.024]	[.041]	[.028]

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

Table 4 replicates Table 3 by subject tutored. Results are very similar to the main results, but suggest a somewhat larger effect on students tutored in Maths compared to those tutored in English. Point scores among those tutored in Maths are typically estimated to be 2-2.5 points higher than in the comparison group. The corresponding figure for students tutored in English is around 1.5 points. As with the main results, we do not find evidence of an impact on the probability of achieving grade C or above. All point estimates are not statistically different from zero. Point estimates for Maths range between 0 and +3 percentage points, while those for English are all around the zero mark.

A positive impact on GCSE points may appear at odds with no evidence of effects on the probability of achieving a grade C or higher. Two main factors may help reconcile these. Firstly, our statistical test may not be sufficiently powered. Indeed, power calculations conducted in the early stages of the project indicated that, with our sample size, we could aim to detect an effect on the probability of hitting the C threshold if this was at least around 10 percentage points. As it turned out, the absolute value of our point estimates is always significantly smaller. In other words, we are not able to distinguish whether our point estimates of, for example, 2 percentage points represent a true effect that goes undetected or a genuine lack of effect.

Secondly, it could be possible that the increases in point scores materialise only at a certain distance below or above the C grade threshold. This could be the case if schools did not fully comply with the requirement to focus tutoring on C/D borderline students, and perhaps included a number well below that working at level. To explore the validity of this hypothesis, we estimate whether programme effects vary by the Key Stage 3 attainment of the tutored students. These are reported in Table 5. Indeed, there is evidence that the impact on GCSE point scores is substantially larger (6 points) for students who achieved no more than Level 4 at Key Stage 3; in line with whole-sample results (2 points) for those who had achieved Level 5 at Key Stage 3; and negative (-2 points) for those who had achieved Level 6 or above. This evidence therefore suggests that the largest improvements are occurring where they do not affect the share of student achieving C or above.

Table 5 - Impact estimates by previous attainment

Level at KS3	**Level <=4**	**Level 5**	**Level >=6**
IPW Model 3	6.894 *** [1.208]	1.28 *** [.377]	-1.426 *** [.318]

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

We also explore the extent to which effects vary according to the gender of the pupil, by eligibility to FSM, by broad region, and by number of sessions attended.[5] We only estimate these with the model and estimator found to perform best at balancing characteristics across treated and comparison groups: Model 3 using Inverse Probability Weighting. The impact estimates for each subgroup are obtained by re-running the estimator on the subsample of treated and comparison students in the given subgroups. For example, only students on FSM are included among potential candidates for the comparison group when considering the effect on tutored students on FSM. When splitting the tutored students by number of sessions attended, the full comparison group is considered as potential matches.

[5] In a number of cases, the estimator failed to converge when attempting to replicate Model 3 Inverse Probability Weighting used in the main results. This occurred when subgroups where too small. To address this, we merge some subgroups together. For example, we had to merge Government Office Regions to create 3 broad regions and grouped the number of sessions attended into three interval ranges.

Results are summarised in Table 6 and Table 7, showing the estimated impact on GCSE points and the probability of achieving grade C or higher respectively. Impact estimates on both outcomes do not differ by gender and eligibility to FSM, and are very similar in magnitude to those estimated for the whole sample. Estimated impacts on GSCE points in London and the South East and in the North West and Yorkshire and the humber are higher than when estimating over the whole sample, with statistically significant point estimates in the order of 3 GCSE points. Similarly, the point estimates on achieving a grade C or above is higher than when estimated on the whole sample, but still statistically insignificant. On the other hand, impacts in the West Midlands and South West compare unfavourably with those estimated over the whole sample. Impacts on GCSE scores are not statistically different from zero, while we estimate a 12 percentage point difference in the probability of achieving C or higher between the treated and comparison group.

We also run the estimation by number of tutoring sessions attended. Results are intuitively plausible but come with some important caveats. Selected students who end up attending either no sessions or only 1 session achieve GCSE points that are no different from those among the comparison group. The estimated difference from the comparison group is around a statistically significant 2 points for students attending between 2 and 6 sessions and 3 points for those attending 7 or more sessions. There is a similar increasing pattern across the point estimates of the probability of achieving C or higher, but, as with the full sample, these are statistically insignificant. Importantly, however, the number of sessions attended is likely to be determined by unobserved factors (for example, motivation or commitment). As we cannot ensure these are balanced across treated and comparison groups, the relationship we identify between number of sessions and outcomes is likely to be a combination of the both any true impact of the programme and possibly different degrees of such unobservable characteristics across the comparison groups and each subgroup of treated students. As such, one should be cautious about making a causal statement between the number of sessions attended and outcomes.

Table 6 – GCSE Points - results by subgroup

Gender	*Male*	*Female*	
	2.053 ***	2.131 ***	
	[.534]	[.4]	
FSM	*No*	*Yes*	
	2.124 ***	2.057 ***	
	[.405]	[.519]	
Broad region	*London & SE*	*NW & Yorkshire & the humber*	*SW & W Midlands*
	3.057 ***	3.156 ***	-0.236
	[.451]	[.727]	[.616]
N# of sessions	*<=1 sessions*	*2-6 sessions*	*>=7 sessions*
	0.107	1.906 ***	3.299 ***
	[.833]	[.444]	[.484]

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

Table 7 - Probability of grade C or higher - results by subgroup

Gender	*Male*	*Female*	
	0.034	-0.018 ***	
	[.028]	[.022]	
FSM	*No*	*Yes*	
	0.011	-0.012 ***	
	[.022]	[.028]	
Broad region	*London & SE*	*NW & Yorkshire & the humber*	*SW & W Midlands*
	0.028	0.064	-0.124 ***
	[.023]	[.039]	[.037]
N# of sessions	*0-1 sessions*	*2-6 sessions*	*>=7 sessions*
	-0.061	-0.008	0.042
	[.04]	[.026]	[.028]

p-value<0.001 ***; p-value<0.01 **; p-value<0.1 *

References

Frölich, M., Huber, M., Wiesenfarth, M., 2015. The Finite Sample Performance of Semi- and Nonparametric Estimators for Treatment Effects and Policy Evaluation (IZA Discussion Paper No. 8756). Institute for the Study of Labor (IZA).

Rosenbaum, P.R., Rubin, D.B., 1985. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. Am. Stat. 39, 33. doi:10.2307/2683903

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55. doi:10.1093/biomet/70.1.41