PROJECT TITLE	Basic Maths Premium Pilot
DEVELOPER (INSTITUTION)	Department for Education
EVALUATOR (INSTITUTION)	NatCen Social Research
PRINCIPAL INVESTIGATOR(S)	Berenice Scandone
TRIAL (CHIEF) STATISTICIAN	Andi Fugard
STUDY PLAN AUTHOR(S)	Martina Vojtkova, Rodrigo Torres, Robert Wishart, Kostas Papaioannou, Molly Scott, Andi Fugard, Karl Ashworth, Josep Espasa, Arnaud Vaganay, Berenice Scandone, Phoebe Averill

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.0	12/09/2019	Original version
2.0	25/09/2020	• Removed analysis of combined cohorts (section 3.1, 3.3) in response to Covid-19 • Updated sections 3.2 and 3.4.1 to reflect new definition and measurement of the primary outcome • Removed secondary analysis assessing the impact of the intervention on whether a GCSE maths re-sit is attempted (previously RQ3) • Added dosage analysis (RQ4) • Updated IPE interview sampling and schedule • Updated section 8 (personnel) • Updated section 10 (timeline)
3.0	31/03/2021	• Removed analyses for the 2019/20 cohort of pupils • Updated design overview (3.2) • Updated 3.4.2 and 3.7 to add information on how we will ensure the comparison group only

		includes students who would have been eligible for BMP • Updated sections 3.8 (primary outcome analysis) • 3.10.2 clarified which funding periods will be considered for the dosage analysis • 4.4 Clarified why non-treatment schools were excluded from IPE interviews • Small copy-editing changes • Updated outcome measure for the 2018/19 cohort of pupils (section 3.4) • Updated dosage analysis section to remove second cohort analysis • Updated the effect size formula in 3.10.5 • Updated section 7.1 (GDPR) • Updated section 8 (personnel) • Updated section 10 (timeline) • Updated references to software packages to be used throughout.
4.0	6/07/2021	• Updated research questions (3.1), to reintroduce one of the

		original research questions that had been removed in subsequent versions of this Study Plan. • Updated information about the source of data that will be used to construct the primary outcome (4.1, 3.2). • Updated information on the analysis approach for the reintroduced research question (3.10.3)
5.0	5/10/2021	• Updated study timeline to reflect delays in the process of requesting outcomes data. • KS4 November 2019 resit results are reflected in 2018/2019 data rather than in 2019/20 KS5 data as previously advised – updated to reflect this

Table of contents

1. Intervention

1.1. Overview

The Basic Maths Premium (BMP) Pilot is an intervention developed by the Department for Education (DfE) that provides additional funding to support 16-18 year old students studying maths after GCSE. The premium, worth up to £500 per eligible student, will be made available to post-16 educational settings in the most disadvantaged areas of England. The pilot will run from autumn 2018 for two years.

The aim of the evaluation is to:

- assess the impact of different funding models on outcomes for students with prior attainment of a grade 3 or below in GCSE maths; and
- identify how the additional funding is used by institutions to understand the underlying mechanisms of change and to build up an evidence base on which activities lead to improvements in teaching and learning.

1.2. Why

Since 2014, 16-year-olds who don't get at least a grade '4' (roughly equivalent to a 'C') in their GCSE are required to keep on studying maths and English until they are 18 or secure a GCSE grade 4 or above in these subjects. However, supporting learners in post-16 institutions to secure these qualifications is challenging. Institutions have received no additional funding for these pupils, and achievement rates remain low: just one in six of those students eligible for free school meals who do not achieve the expected standard in English and maths at age 16 go on to gain those qualifications by age 19[1].

1.3. Who

The BMP is provided by the Education and Skills Funding Agency (EFSA), an executive agency of the DfE.

The intervention was offered to all eligible post-16 educational institutions in Category 5 and 6 Achieving Excellence Areas (AEA)[2]. All institutions that agreed to be part of the pilot (by signing a memorandum of understanding) and had at least one eligible student from cohort 2018/19 enrolled by November 2018 received the intervention and will be included in the evaluation.

[1] DfE: Level 2 and 3 attainment by young people aged 19 in 2016. Table 13b. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/603921/Tables_6_-_15_2017SFR_V2.xlsx?_ga=2.43900325.1492168197.1566900094-1719746700.1564482429

[2] The Achieving Excellence Areas categories are based on a combination of indicators which show current educational performance with indicators which show capacity to improve to define areas which are most in need of support. Further details available at: https://www.gov.uk/government/publications/defining-achieving-excellence-areas-methodology

Students enrolled on a 16 to 19 study programme for the first time without prior attainment of a maths GCSE grade 4 or above for the 2018 to 2019 academic year cohort are eligible for the additional funding attached to this pilot. Students who are exempt from the English and maths condition of funding are not eligible.

The intervention is expected to benefit:

- Directly:
  - Post-16 institutions in Category 5 and 6 AEAs that have signed up for the pilot and have at least one student eligible to receive BMP funding by November 2018.
- Indirectly:
  - Students without prior attainment of at least grade 4 in GCSE maths that enrol in their post-16 studies in the participating institutions in the 2018/2019 academic year (the "eligible students" for this pilot). These students are hereafter referred to as "Cohort 18/19".

For the purposes of the BMP pilot and its evaluation, both the 18/19 and the 19/20 cohort student groups (as defined above) are considered to be beneficiaries of the BMP.

1.4. What

Depending on the funding model, participating post-16 educational institutions will receive up to £500 of additional funding per eligible student to provide support to students' achievement in maths.

The definition of achievement in maths for the purpose of the pilot is as follows:

- For students with prior attainment of a grade 3 in GCSE maths, achievement will be defined as attainment of a grade 4 or above in GCSE maths, and;
- For students with prior attainment of a grade 2 or below in GCSE maths, achievement will be defined as attainment of a grade 4 or above in GCSE maths, or Functional Skills Level 2 in maths.

The BMP Pilot will test three alternative funding models:

- Model T1: Consists of a guaranteed payment of £500 for every student enrolled by November 2018 without a Grade 4 or above in GCSE maths (only for Cohort 18/19 funding-eligible students);
- Model T2: Consists of a guaranteed payment of £250 for every student enrolled by November 2018 without a Grade 4 or above in GCSE maths, and a further £250 in the academic year 2020 to 2021 for every student who achieves in maths by Summer 2020 (only for Cohort 18/19 funding-eligible students);
- Model T3: Consists of a payment of £500 in the academic year 2020 to 2021 for every student enrolled by November 2018 who goes on to achieve in maths by Summer 2020 (only for Cohort 18/19 funding-eligible students).

Institutions can choose how to utilise their additional funding. Regardless of how resources were assigned, institutions will be free to invest their resources in Cohort 18/19 and/or Cohort 19/20 students as they see fit. This implies that eligible students (Cohort 18/19) will not necessarily be the only beneficiaries of the intervention.

The DfE has informed participating settings that the additional funding can be used for approaches or programmes known to be effective for improving maths outcomes for students. For example, the DfE suggest that the funding could be used to provide more teaching hours, smaller class sizes or to invest in technology to support students without a Level 2 qualification to achieve this level. However, institutions and colleges will have flexibility over how they choose to allocate the funds.

1.5. Variation from original programme design impacting evaluation

Guaranteed payments for those institutions assigned to funding schemes T1 and T2 were made in March 2019, seven months after the beginning of the academic year 2018/2019. This raised concerns about the possibility of a large proportion of resources being spent in the next cohort of students aiming to re-sit maths GCSEs during the following academic years.

A survey of participating institutions in February and March 2019, confirmed that although a large proportion of institutions were already spending some of the resources before receiving them[3] (around 65% on average), a large majority of institutions were planning to spend part of them also on the following cohort of students expecting to re-sit their maths GCSE[4] (68% of institutions completing the survey). This justified estimating the impact of the intervention not only for those students in the original cohort (for which resources were initially allocated), but also for those in the following one, as they may also benefit from the resources.

In order to have a better depiction of all of the effects of this intervention (and as detailed in section 3.1.1), the study design was previously revised to estimate the impact of the programme for students in both cohorts 2018/19 and 2019/20 (see study plan v1.0). A combined analysis including students from both cohorts together was additionally planned. However, due to the disruption to education caused by the Covid-19 pandemic from March 2020, and the resultant non-comparability of the two cohorts, the analysis will focus only on the first cohort of students (cohort 18/19). References to the second cohort have been removed from the current study plan hereafter.

[3] More specifically, institutions were asked whether they already started spending the funding they expected to receive through the Basic Maths Premium.

[4] In this case institutions were asked whether part or all the allocated resources would be spent in academic year 19/20.

2. Study rationale and background

2.1. Policy background

Since September 2013, all 16-19 year-old students have been required to follow a 'study programme', which reflects their study and employment goals and their prior attainment. From August 2014, this study programme has included a requirement to continue the study of maths and English if the student does not already hold a GCSE grade A*-C (grade 9-4 in the reformed system) in maths and/or English. From August 2015, students starting a new study programme with a GCSE grade D (grade 3 in the reformed system) in maths and/or English must enrol on GCSE courses only. This requirement is a 'condition of funding'. Providers with students who do not meet this condition, i.e., are not enrolled on the relevant maths and English course, will lose funding in a future year. There are cases in which students may be exempt.

An adjustment to the condition of funding has been established for the academic year 2019 to 2020. The change means that students with a grade 2 or below can study towards a pass in Functional Skills Level 2 or they can still study towards a GCSE grade 4-9. Once they have achieved this there is no requirement to undertake further English or maths qualifications to meet the condition of funding. Those with a grade 3 must still study GCSE only. This is a change from academic year 2018 to 2019, where students with a grade 2 or below who passed functional skills level 2 would still need to continue to study towards a GCSE grade 9 to 4.

The latest data from The Office of Qualifications and Examinations Regulation (Ofqual) indicates that close to 60% of students achieved a grade 4 or above in 2018[5]. The requirement for some post-16 students to continue studying English and/or math has led to an increase in the number of entries to study GCSE maths or Functional Skills Level 2 at post-16 educational settings (Ofqual, 2017), with a recent report highlighting that around 37% of students require two or more re-take attempts to achieve the GCSE (Rodeiro, 2018).

Institutions have received no additional funding to support these students. According to the Education Policy Institute (Dominguez-Reig, G. & Robinson, 2019), the funding per student for 16-19 education has actually fallen 16 per cent in real terms between 2010/11 and 2018/19. This, combined with an increase in the number of students required to continue studying maths has led to multiple challenges associated with resource availability within post-16 education settings, including burden on teaching staff, larger class sizes, and insufficient funding to support students' achievement.

2.2. Theoretical background

2.2.1. Expected impact of additional funding on attainment

In 2017, DfE conducted research to understand the impact of changes in school funding on pupils' outcomes in England (Department for Education, 2017)[6]. Their findings suggest that additional school resources positively influence attainment, although the effects are modest at all key stages.

[5] Ofqual, GCSE outcomes in England. https://analytics.ofqual.gov.uk/apps/2018/GCSE/Outcomes/

[6] See review in the following link: https://www.gov.uk/government/publications/school-funding-and-pupil-outcomes-review.

The first part of the report, a literature review, references a few studies that are directly relevant to the evaluation of the BMP:

- One of these studies (Nicoletti & Rabe, 2012) quantifies the relationship between spending on education and test scores at 16. By comparing outcomes for siblings exposed to different levels of expenditure, they find that a £1,000 increase in expenditure per student raises achievement by about 0.2 GCSE points.
- International evidence from PISA (the OECD's Programme for International Student Assessment) shows the level of education spending can have an impact on a nation's educational performance (OECD, 2012). According to the OECD, levels of spending explain around a fifth of the variation in PISA results. However, the impact is much less pronounced for high-income countries.

The second part of the DfE report looks at the relationship between funding and attainment at KS4 in English maintained schools between 2010 and 2015. The authors find no statistically significant effect at KS4.

Other existing research suggests that investment of additional per student expenditure shows larger effects in disadvantaged institutions (Pugh et al., 2011). In a further review, additional resourcing improved GCSE attainment for all students, and this association was strongest for students with lowest attainment in Key Stage 2 (Jenkins et al., 2006). This study also identified a small positive relationship between funding and attainment in GCSE maths, though notably, this relationship only reached significance for the 40% of students with lowest prior attainment.

Other previous studies have pointed to the positive effects of the additional financial support for schools provided in the form of the Pupil Premium. However, most of these studies were qualitative and lacked a comparison group. The Committee of Public Accounts noted evidence that the attainment gap between disadvantaged pupils and their peers had started to narrow since the introduction of the Pupil Premium in 2011 (Committee of Public Accounts, 2015). Also, an association between the overall effectiveness of the school and the impact of the Pupil Premium was identified. It was found that good and outstanding schools showed higher level of commitment to close the attainment gap by targeting interventions and using robust tracking systems (Ofsted, 2014).

A large number of studies analyse the impact of cash transfer programmes on education outcomes. A recent systematic review of these studies found that cash transfer programmes improve the odds of being enrolled in and attending school (Baird et al., 2014).
However, the effect of cash transfer programmes on test scores was found to be small (<0.1 SD). The authors of the review conclude that, in the absence of complementing interventions, cash transfers are unlikely to improve student attainment substantively. However, another systematic review indicated that cash transfers were found to have statistically significant and positive effects on cognitive development test scores (Bastagli et al., 2016).

There is also an emerging yet inconclusive body of literature which seeks to evaluate the utility of 'payment by results' (PBR) as an alternative model of funding for public services. PBR is an approach to funding public service initiatives, whereby payments are rewarded to service deliverers contingent upon achieving specific outcomes (Webster, 2016). PBR has been increasingly adopted by policymakers in sectors such as crime and justice and health, due to the potential to commission specific outcomes from public sector initiatives, whilst reducing risk of expenditure on interventions which prove to be ineffective (Fox & Albertson, 2012). As such, evidence indicates that the conditional approach to funding can encourage innovation and efficiency within service delivery, as providers are incentivised to meet outcome targets (Department for CLG, 2015).

Nevertheless, the literature also outlines a number of challenges associated with PBR. There is concern that this funding modality may lead providers to neglect other important outcomes, in favour of achieving the outcome on which compensation is dependent (Fox & Albertson, 2012). Similarly, PBR may encourage ‘creaming and parking’, whereby only individuals most likely to reach the desired outcome are selected to receive the intervention. For example, in an employment programme, individuals experiencing greater barriers to finding sustained employment were given less support by staff (Work and Pensions Committee, 2011).

The National Audit Office identified the following ten success factors for PBR schemes (NAO, 2015):

- Clear and measurable objectives;
- Clearly identifiable cohort/target population;
- Ability to clearly attribute outcomes to provider interventions;
- Data available to set baseline;
- An appropriate counterfactual can be constructed;
- Services are non-essential and underperformance of failure can be tolerated;
- Providers exist or are prepared to take the contract at the price and risk;
- Providers are likely to respond to financial incentives;
- Sufficient evidence exists about what works to enable providers to estimate costs of delivering services; and
- A relatively short gap between provider intervention and evidence of outcome.

2.2.2. Dosage and use of additional funds

A similar study by Nicoletti & Rabe (2013) also examined the effect of different types of expenditure on different groups in secondary school, rather than looking simply at the effect of additional spending. Their results, again summarised in the DfE report (2017), suggest that:

- Spending on teachers has a positive impact on attainment for most groups of pupils, although effects are small: A £1,000 increase in per-pupil spending on teachers is associated with between a 0.5% and 2.5% increase in standardised GCSE test scores in Science, English and Mathematics.
- Increased pupil-teacher ratios have a small negative impact on attainment. For most pupils a one-pupil increase in the pupil-teacher ratio reduces standardised GCSE test score by 1%, rising to 2% for the lowest attaining 10% of pupils.
- Spending on education support staff was found to positively affect the attainment of Free School Meals (FSM), English as an Additional Language (EAL), and Gifted and Talented pupils. A £1,000 increase in spending on education support staff would have increased Gifted and Talented scores by 11%, FSM scores by 7% and EAL test scores by over 12%.
- Spending more on learning resources (e.g., computers, books) in most cases positively affected attainment. Spending an extra £1,000 would have boosted the test scores of SEN pupils by 6.2%.

3. Impact evaluation

3.1 Research questions

The impact evaluation will answer the following research questions (RQs):

RQ1	What is the impact of the different funding models on maths achievement of eligible students in post-16 settings in England compared to business as usual?
RQ2	What is the effectiveness of the different funding models relative to each other?
RQ3	How do these effects differ by student disadvantage status, prior attainment in GCSE maths and number of eligible students per institution?
RQ4	What is the impact of the total additional funding spent by post-16 settings in England on eligible students' maths achievement?
RQ5	What are the effects of the different funding models on the likelihood of pupils attempting a GCSE maths exam retake, for students with a prior achievement of grade 2 and below?[7]

3.1.1 Hypotheses

Each research question is related to a set of specific hypotheses. In particular, this study design incorporates one primary research question with some additional exploratory analyses, and two secondary research questions. For all research questions we will be reporting results using the first cohort of students (cohort 18/19) as the main results.

3.1.1.1 Impact of the intervention compared to business as usual (RQ1)

These hypotheses correspond to the study's main goal: assessing the effectiveness of the different funding schemes (T1, T2 & T3) on individuals' educational achievement in maths.

Hypothesis H1 is that students in participating institutions will on average have, at the end of the intervention, better maths outcomes than comparison students in non-participating institutions. More specifically:

- H1a: Cohort 18/19 students in participating institutions that have been randomly assigned to model T1 will on average have better maths outcomes than comparison students in non-participating institutions.
- H1b: Cohort 18/19 students in participating institutions that have been randomly assigned to model T2 will on average have better maths outcomes than comparison students in non-participating institutions.
- H1c: Cohort 18/19 students in participating institutions that have been randomly assigned to Model T3 will on average have better maths outcomes than comparison students in non-participating institutions.

3.1.1.2 Relative effectiveness of different funding models (RQ2)

The hypotheses under research question 2 are:

- H2a: There will be a different (higher or lower) impact on maths outcomes of Cohort 18/19 students in institutions assigned to model T1 compared to those in institutions assigned to model T3.

[7] RQ5 was first specified in the original study plan (v1.0), where it appeared as 'RQ3'. It was subsequently removed in later versions of the study plan and replaced with the current RQ4. This is because it was thought that the question could not be analysed using available data. Further investigation has confirmed that we can obtain data to answer the question, so it has been re-introduced into the Study Plan under RQ5. For more information see Section 3.1.2.

- H2b: There will be a different (higher or lower) impact on maths outcomes of Cohort 18/19 students in institutions assigned to model T2 compared to those in institutions assigned to model T3.
- H2b: There will be a different (higher or lower) impact on maths outcomes of Cohort 18/19 students in institutions assigned to model T1 compared to those in institutions assigned to model T2.

3.1.1.3 Subgroup impacts (RQ3)

The hypotheses under research question 3 are:

- H3a: The effect of the T1, T2, and T3 combined[8] on student maths outcomes for Cohort 18/19 will differ (be higher or lower) for those students that have ever been eligible for free school meals compared to comparison students in non-participating institutions.
- H3b: The effect of the T1, T2 and T3 combined on student maths outcomes for Cohort 18/19 participants will differ (be higher or lower) depending on the number of eligible students that institutions have.
- H3c: The effect of the T1, T2 and T3 combined on student math outcomes for Cohort 18/19 participants will differ (be higher or lower) depending on student prior attainment in GCSE maths.

3.1.1.4 Dosage analysis (RQ4)

Analysis will be conducted to estimate the impact of the total additional funding spent by post-16 settings in England on eligible students' maths achievement. This analysis will be conducted using the pooled sample from the three treatment groups.

- H4: T1, T2 and T3 Students in post-16 settings that spend a greater amount of additional funding will be more likely to achieve a grade 4 or above in maths than post-16 settings spending less additional funding.

3.1.1.5 Impact of the intervention on the likelihood of pupils attempting a GCSE maths exam retake for students with prior achievement of grade 2 and below (RQ5)

The hypotheses under research question 5 are:

- H5a: Being assigned to model T1 will have a different (higher or lower) impact on the chances of Cohort 18/19 students attempting a GCSE maths re-take exam compared to comparison students in non-participating institutions.
- H5b: Being assigned to model T2 will have a different (higher or lower) impact on the chances of Cohort 18/19 students attempting a GCSE maths re-take exam compared to comparison students in non-participating institutions.
- H5c: Being assigned to model T3 will have a different (higher or lower) impact on the chances of Cohort 18/19 students attempting a GCSE maths re-take exam compared to comparison students in non-participating institutions.

[8] The main reason why all treatments arms are combined in one is to preserve statistical power, given that the sub-group analysis will involve partitioning the sample into smaller sub-samples. As part of the sensitivity analysis, we will also estimate this per each arm separately. A discussion of this issue can be found in the subgroup analyses section.

We will accompany these analyses with descriptive presentation of the proportion of students attempting a GCSE maths resit in the BMP and non-BMP groups, alongside those attempting a Functional Skills Level 2 exam and those who don't take any re-sit. This will help to contextualise the findings under this research question and understand what has changed in terms of the kinds of exam that pupils are entered for.

3.1.2 Background information about the research questions

There have been some changes to the research questions outlined in Section 3.1 over different iterations of this Study Plan. We briefly summarise those changes here.

RQ5 was part of the original study plan (v1.0), where it was called RQ3. This question was subsequently removed in later versions of the study plan and replaced by RQ4. The question has now been reintroduced as RQ5.

The original motivation for RQ5 was to help understand whether exposure to the BMP affects the likelihood of pupils attempting GCSE maths re-sits, rather than Functional Skills level 2 (or not taking any exam at all). The reason for removing this question in later versions of the study plan was related to the project team's prior understanding of the structure of the data that would be available to us to examine outcomes. Originally, we had planned to construct our outcome measures using the Young People's Matched Administrative Database (YPMAD) dataset within the National Pupil Database (NPD). However, the way that this dataset is structured means that it would not have been possibly to identify all retake exams that students may have taken. The YPMAD is compiled annually and the exam information that it contains is constructed by looking across different datasets to identify the best qualification that students could have received. This means that, had a student happened to re-sit both GCSE maths and Functional Skills Level 2 in a given academic year, but only passed the functional Skills Level 2 exam, we would not have been able to observe from this dataset that they had attempted a GCSE maths retake as well.

We therefore determined that it was not possible to assess RQ5 using this data, and this led to the question being replaced by RQ4. RQ4 was thought to be important given that schools received the unconditional portion of funding later in the year than they had expected and thus may have spent less than anticipated. This could have affected the support students received and hence their outcome.

The evaluation team has now re-introduced RQ5. This is because we have now discovered that the question can in fact be answered using data available from the NPD. Discussions with the Data Sharing and Data Platform Services Unit have confirmed that there is an exam-level extract available through the Key Stage 4 dataset that includes data from students at all relevant post-16 educational settings and that will enable us to reliably identify all re-sit exams attempted by students in our target cohort, thereby allowing us to address RQ5.

3.2 Design overview

Design type		Three-armed cluster-randomised trial with a matched comparison group
Unit of analysis **(institution, pupils)**		Students without a grade 4 or above in GCSE maths in post-16 educational institutions
Number of Units to be included in analysis **(Intervention, Comparison)**		• Intervention: 436 post-16 educational institutions expected (141 for treatment arm 1, 145 for treatment arm 1 and 150 for treatment arm 3) • Comparison: More than 436 post-16 educational institutions expected (depending on the methodological approach)
Outcomes	**Primary**	Achievement in maths as defined by the pilot: • GCSE maths grade 4 or higher for students with prior attainment of grade 3 in GCSE maths; and • GCSE maths grade 4 or above or maths Functional Skills Level 2 attainment for students with prior attainment of grade 2 or below in GCSE maths (binary outcome)
	Secondary	N/A
Outcome sources **(instruments, datasets)**	**Primary**	National Pupil Database (Key Stage 5 Exam table) (Derived from KS5_GRADE). We will be using formal exam results and not teacher-assessed grades introduced due to Covid-19.
	Secondary	N/A

The BMP pilot will be evaluated using a quasi-experimental evaluation design. The design combines elements of a three-armed cluster-randomised trial with propensity score estimation and regression analysis to establish a comparison group.

824 eligible post-16 institutions in Category 5 and 6 AEAs were randomly assigned to one of three treatment arms:

- (T1) Institutions receiving a guaranteed additional £500 for each eligible student in Cohort 2018/19 enrolled by November 2018.
- (T2) Institutions receiving a guaranteed additional £250 for each eligible student in Cohort 18/19 enrolled by November 2018 and a further £250 in September 2020 for each eligible student in Cohort 18/19 enrolled by November 2018 that has achieved in maths by September 2020.
- (T3) Institutions receiving an additional £500 in September 2020 for each eligible student in Cohort 2018/19 enrolled by November 2018 that has achieved in maths by September 2020.

Following randomisation, all eligible institutions were invited to take part in the pilot and the study.[9] Of these, 469 institutions (53%) agreed to participate. These institutions were assessed against the remaining eligibility criterion (a minimum of one eligible student enrolled by November 2018) in January 2019. From those 469 institutions 436 remain in the trial. These institutions will represent the final sample of institutions participating in the pilot (the treatment group institutions).

Given DfE's commitment to offer all institutions in the pilot areas some form of intervention, it was not possible to randomly assign a proportion of the institutions in the pilot areas to a comparison condition of business as usual. We will therefore construct a suitable comparison group for the evaluation using a quasi-experimental design (QED). Our comparison group will be drawn from a pool of institutions in Achieving Excellence Areas (AEA) Categories 4 or below. These areas were not offered the programme but have some institutions that could act as a suitable comparison group provided that they show similar characteristics to institutions in the three intervention arms comprising the treatment group.

We propose to use the doubly-robust estimation approach (Funk et al., 2011) as our primary estimation technique. This method combines propensity score estimation with regression analysis to estimate causal impact[10]. First, propensity scores are derived across the sample, representing the conditional probability of being exposed to the BMP given a set of observable covariates. Second, a regression model is estimated to model the outcome variable as a function of exposure to the intervention and covariates. The doubly-robust estimator brings together the propensity scores and estimated coefficients from the regression step to identify the causal impact of the intervention.

An advantage of the doubly-robust approach relative to two-stage propensity score matching (which we use for our robustness check, as described in section 3.9) is that it relaxes some restrictions on the range of possible institutions from which individuals can be selected for the comparison group. This reduces the risk of selecting institutions that do not contain sufficient numbers of individuals to form the comparison group. A second advantage of this approach is that it is robust to misspecification of either the outcome regression or the model estimating propensity scores.

As a robustness check, we propose to use a two-stage matching procedure, where we initially match institutions based on their observed characteristics, and then individuals within matched institutions, making use of individual level characteristics. Accounting for both individual and institution-level characteristics in the matching approach is particularly important in an educational context, as both influence educational performance. The advantage of this two-stage approach is that individuals selected in the comparison group will belong to institutions that are very similar to those in the treatment group, as well as exhibiting similar individual-level characteristics. However, this approach is much more demanding in terms of data and runs the risk that there are insufficient numbers of comparable individuals within comparison

[9] Treatment assignment was not disclosed to the eligible institutions at this stage.

[10] A difference in difference approach was discarded since it would not be possible to assess the plausibility of the parallel trends assumption. This would require the existence of a series of pre-intervention observations of the outcome variable for students in both the treatment and comparison group.

group institutions to be matched to treatment group individuals[11]. We will use two different calliper sizes on the propensity score for our preferred matching approach (radius matching) under this robustness check, to assess the stability of our results.[12]

The robustness check will assess the relative effectiveness of the different funding mechanisms by making pairwise comparisons between individuals in each of the treatment groups with individuals belonging to the matched comparison group. Clustered standard errors will be used, given that observations are clustered within institutions. In simple terms, the effect will be estimated as the difference between the average observed outcome across all individuals in the treatment group and that of their matched sample of individuals in the comparison group following the intervention.

Both the doubly-robust method and two-stage propensity score matching analyses will be carried out under intent-to-treat (ITT). This captures the impact of the intervention for all settings that agreed to participate in the study and were eligible for it, regardless of whether they ended up spending any money provided by the intervention.

3.3 Participants

The intervention was offered to all eligible post-16 educational institutions in Category 5 and 6 Achieving Excellence Areas (AEA). All institutions that agreed to be part of the pilot (by signing a memorandum of understanding) and had at least one funding-eligible student from cohort 18/19 enrolled by November 2018 were exposed to the intervention and will be included in the evaluation.

Given this, we will consider the following to be the study participants: Students enrolled on a 16 to 19 study programme for the first time without prior attainment of a maths GCSE grade 4 or above for the 2018/2019 academic year cohort by November 2018 eligible for the additional funding attached to this pilot (the “eligible students” for this pilot). Students who are exempt from the English and maths condition of funding are not eligible. Students that have enrolled following November of each academic year will be excluded from the evaluation.

All eligible institutions were randomly allocated to one of the three treatment conditions (T1, T2, T3 as outlined above) using stratified blocked randomisation. The randomisation used two stratifying factors: institution type[13] and number of Condition of Funding (CoF) students in 2016/2017.[14] Institutions were initially grouped by institution type. Within each institution type stratum, institutions were ordered by number of CoF students, from largest to smallest, and

[11] In case the initial approach for matching is not successful; that is, if we don’t find enough matched individuals for those in the treatment group, the two-stage approach will be replaced by a stratification procedure. In that case, instead of matching one or more institutions per each institution in the treatment group, we will classify institutions into deciles according to their propensity score, and subsequently perform regression analysis utilizing student-level data to estimate the effect of the treatment. More details on this approach are included in the section 3.9.4 of this plan.

[12] Given the fact that we are also proposing stratification matching, we have decided to implement only radius matching on the propensity score, using two different calliper sizes (0.1 and 0.2 standard deviations on the propensity score).

[13] Using a standard classification into 16-19 provider, Academy and Sixth-Form provided by the DfE.

[14] This variable was used as a proxy for the likely number of students within each institution that will meet the eligibility criteria for the pilot funding in 2018/2019.

divided into blocks of three based on contiguous ordering of size. Within each block, the ordered list of institutions was associated with a randomly ordered list of the three treatment assignments. Institutions where the number of CoF students in 2016/17 was unknown were divided into blocks of three and each block assigned a randomly ordered list of the treatment assignments. As a result, a last replication correction procedure, as outlined in Alferes (2012) was applied to incomplete blocks within each stratum[15].

The randomisation was carried out independently by NatCen Social Research at the end of May 2018 prior to eligible institutions being informed about the opportunity to participate in the pilot. Once randomisation was complete, all eligible institutions were contacted by the DfE and offered the opportunity to participate in the pilot on the understanding that they will be informed by NatCen which treatment arm they have been assigned to upon agreement to participate in the pilot and evaluation.

No baseline testing or data collection was conducted prior to randomisation and treatment assignment, as the evaluation will rely on baseline data from administrative sources such as the NPD.

Randomisation to each of the three treatment groups was carried out in R (Version 3.4.1, using random number generation (*runif*) available from the *base* package). The randomisation was conducted by an independent analyst within the evaluation team.

Below in table 1, we provide a brief breakdown of numbers randomised by institution type and trial arm.

Table 1: Allocated settings by provider type and treatment group

Treatment group	16-19 provider	Academy	6th form	Total
A	58	167	47	272
B	60	164	47	271
C	64	171	46	281
Total	182	502	140	824

3.3.1 Recruited institutions and expected number of eligible students

Of the 824 institutions randomly assigned to treatment groups, 469 agreed to participate in the project by the end of 2018. From those 469 institutions, 436 remain after accounting for those institutions dropping out from the sample[16].

After this attrition, the 436 institutions that have agreed to participate remain approximately equally distributed across the treatment groups (participation rate is between 52% and 54%

[15] In short, for those blocks where the number of units is below 3, we randomly assign those settings to one of the three treatment arms.

[16] In total, 33 institutions withdrew for the following reasons: Institution with no eligible 2018/19 students (25 institutions); institution lost its only student eligible for funding before funds were transferred (1); institution was merged with another institution in the trial (1), institution decided to withdraw (5); duplicate institution randomised twice (1).

for each treatment group). Table 2 shows that when looking at participation rates by type of provider, however, 16-19 providers were much more likely to agree to participate than other providers; presumably reflecting the greater relevance of the programme to their institutions, given the potential number of eligible students (see table 4).

Table 2: Proportion of Respondents by institution and treatment group

Treatment group	16-19 provider	Academy	6th form	Total	Proportion of participants
A	42	79	20	141	52%
B	46	76	23	145	54%
C	44	83	23	150	53%
Total	132	238	66	436	53%
Proportion of allocated institutions	73%	47%	47%	53%	

As discussed in sections 1.3 and 1.4, while the eligibility for the BMP funding and the amount of resources allocated per institution was based on the number of eligible students in 2018, in practice the effects of the intervention are likely to affect both the 18/19 and the 19/20 student cohorts. This is because participating institutions have not been constrained to which cohort the funding should be spent on. Indeed, as discussed in section 1.4, a survey of participating institutions indicated that the resources that have been allocated to date are likely to be spent on both the 18/19 and 19/20 cohort.

The evaluation originally planned to analyse the impact of the intervention on both cohorts separately as well as the overall effect of the intervention for both cohorts together. However, given the disruption to schooling and outcome measurement due to Covid-19, which affects both cohorts differently (see section 3.4), we no longer consider combining cohorts to be a robust option. The analysis will therefore be conducted only for the 18/19 cohort. This cohort consists of students enrolled on a 16 to 19 study programme for the first time without prior attainment of a maths GCSE grade 4 or above for the 2018/2019 academic year cohort by November 2018 are eligible for the additional funding attached to this pilot (the “eligible students” for this pilot). Students who are exempt from the English and maths condition of funding are not eligible. Students that have enrolled following November of each academic year will be excluded from the evaluation.

In Table 3 we report the number of Cohort 18/19 students who were eligible for funding per type of institution and treatment arm. Eligibility was based on students meeting the ESFA condition of funding (data for this cohort has become available in January 2019).

Table 3: Number of participants eligible for funding by treatment group and type of institution (Cohort 18/19)

Treatment group	16-19 provider	Academy	6th form	Total
A	15,065	1,429	407	16,901
B	13,909	1,548	335	15,792
C	14,972	1,635	375	16,982
Total	43,946	4,612	1,117	49,675

When putting the information from Table 2 and Table 4 (below) together we see that a large majority of eligible students for cohort 18/19 (approximately 88%) attended 16-19 providers. However, these represent a lower proportion (around 30%) of the institutions participating in the study. This implies that we expect an unequal distribution of eligible students across institutions according to their institution type, with a larger concentration of these in 16-19 providers.

Table 4: Average number of participants per institution eligible for funding by type of institution (Cohort 18/19)

Treatment group	16-19 provider	Academy	6th form	Average
A	359 (min:1; max:1,169)	18 (min:1; max:200)	20 (min:1; max: 74)	120
B	302 (min:1; max:1,200)	20 (min:1; max:148)	15 (min:4; max:65)	109
C	340 (min:1; max:1,324)	20 (min:1; max:155)	16 (min:1; max:47)	113
Total	333	19	17	114

3.3.2 Balance at randomisation

Once we have gathered individual-level data from the NPD, we will produce institution-level measures checking how balanced the samples for the different arms are in terms of institutions observed characteristics, in order to review how successful the randomisation process was.

Amongst other variables, we will be reporting standardized differences in means across institutions in different treatment groups on aggregated measures of previous academic attainment,[17] institution type, number of students eligible for the intervention[18] and previous free school meals eligibility,[19] reporting whether these differences are statistically significant.[20] We will also be reporting whether these differences were over 0.1 standard deviations. The balance checks will also include kernel density plots or histograms on measures of previous academic achievement and FSM eligibility.

3.4 Data

3.4.1. Primary outcome variables

The primary outcome of interest will be achievement in maths as defined by the pilot. Specifically, for students with prior attainment of a grade 3 in GCSE maths, achievement will be defined as attainment of a grade 4 or above in GCSE maths, and for students with prior attainment of a grade 2 or below in GCSE maths, achievement will be defined as attainment of a grade 4 or above in GCSE maths, or Functional Skills Level 2 in maths.

Attainment outcomes were intended to be measured at the end of the 2019/20 academic year. In response to the Covid-19 school closures, outcome assessment of GCSEs was non-standard in the 2019/20 academic year; Centre assessed grades have been used to assess pupil attainment (Department for Education, 2020). This introduces some inconsistency in outcome measurement across years, as pupils were more likely to be awarded higher grades under centre assessment (e.g., 77.2% of students aged 16 obtained grade 4 or above in 2020 versus 71.5% in 2019; Ofqual, 2020). We have decided to therefore amend the outcome measure, to the highest grade awarded by November 2019. This will therefore include grades up to and including re-sits in November 2019 for cohort 2018/19 (i.e., 6 months earlier than originally planned). This means that grades for cohort 2018/19 will not use the centre-assessed grades.

We understand that the primary concern of the study is whether or not participating students attain a grade 4 or above in GCSE maths or a Level 2 qualification (depending on the student's prior attainment). In view of this, we propose that for each participating student, we code achievement in maths as a binary variable. That is, depending on student's prior attainment as set out above, we will code attainment of either Level 2 Functional Skills in maths and/or GCSE maths of grade 4 or higher as a pass,[21] and all else as a fail.

Specifically, for cohort 18/19 students with prior attainment of grade 3 in GCSE maths, the most up to date GCSE maths attainment scores achieved by November 2019, prior to the Covid-19 pandemic, will be taken from the NPD and coded in binary format: a pass will be defined as attaining grade 4 or above; a fail will be defined as reaching grades 1 to 3 or a U.

For cohort 18/19 participating students with prior attainment of grade 2 or below in GCSE maths, we define the outcome variable as follows. For those students that only sit a Level 2 Functional Skills exam during the evaluation period, we will record those that have attained

[17] Utilizing Key Stage 2 maths results.

[18] Institution's number of Condition of Funding (CoF) students in 2018/2019.

[19] Proportion of students that have been eligible for free school meals at any time in the past five years (ages 10 to 15). This measure will be reported as effect sizes (standardized values, measured in standard deviations of the outcome).

[20] For a complete list of institution-level variables, please refer to section 3.4.2 (other sources of data)

[21] While there are multiple pass grades for GCSE maths, Level 2 Functional Skills is awarded only on a pass/fail basis. Consequently, a binomial pass/fail measure is more appropriate than a continuous outcome for the primary outcome of interest.

the qualification by November 2019 as a pass, and those that have not attained the qualification as a fail. For students that sit both Level 2 Functional Skills and GCSE maths during the evaluation period, we will define a pass as attaining either a Level 2 Functional Skills qualification or a grade 4 or above in GCSE maths.

We will classify eligible students who neither sit a Level 2 Functional Skills exam nor re-take GCSE maths exam during the evaluation period, including those that have discontinued their studies at that institution,[22] as a fail, as this approach aligns with the primary concern of the study.[23]

In practice, the outcome will be an individual level binary variable after the treatment took place, reporting whether the individual met or did not meet the defined thresholds according to the previous criteria.

For participating students in cohort 18/19,[24] Level 2 attainment in maths measures will be taken from the NPD for all students that started their post-16 studies in the 18/19 academic year. We expect to receive this data in autumn 2021.

Outcome data will be obtained by matching pupils who started post-16 studies at the start of the 18/19 academic year in the study to their outcome data from the NPD Key Stage 5 Exam table by the end of the 19/20 academic year. This is a "long" dataset with one row per student and each exam taken, along with the season when the exam was attempted. We will use this data to trace all the Maths GCSE or Level 2 Functional skills exams taken by students in our sample in either 2018-19 or 2019-20, along with the grade received in the exam and the season (winter or summer) when they took the exam. We can then exclude results from any re-sits taken in summer 2020 (when grades were teacher-assessed) in order to construct the outcome. The data includes the Level 2 attainment results (including re-sits) of students who have moved into post-16 education and covers all post-16 institutions in England. It will therefore capture our required sample and is suitable to construct the outcome.

3.4.2 Other sources of data

The estimation of propensity scores and outcome regressions will draw on both individual-level and institution-level data. These data will come from NPD administrative datasets (mainly the Young Person's Matched Administrative Dataset, YPMAD), containing information about individuals undertaking post-16 studies and institutions (aggregating individuals' data at the institution-level when necessary). We will also utilize local area level variables used by the DfE to produce the AEA categories.

More information about how these data will be used in the analysis is contained in section 3.8.1.

[22] We expect students dropping out from their studies to be a non-negligible proportion of those enrolled in post-16 institutions. National figures show retention rates for these institutions to be between 80% and 95% according to the type of qualification they are pursuing. However, for the group of students failing their Math GCSEs we expect this figure to be substantially higher.

[23] The primary concern of the study is whether the additional funding increases the number of students achieving in maths as defined above. Consequently, we believe that treating students that have not re-sat their GCSE maths or have not sat their Functional Skills Level 2 qualification as having failed (rather than as missing values) is appropriate in this instance.

[24] Students in academic year 2018/19 enrolled in participating post-16 institutions by November 2018, without prior attainment of at least a grade 4 GCSE in maths.

The institution-level variables we will include are:

- Mean number of students enrolled in the institution at baseline and two years prior[25] (institution size)
- Number of students in 2016/17 meeting the eligibility criteria for the programme[26]
- Institution's proportion of students meeting the eligibility criteria in 2016/17[27]
- Type of institution[28]
- Institution geographical location (region)[29]
- Institution's proportion of students who were eligible for FSM between age 10 and 15 (as a measure of social disadvantage)[30]
- Institution's proportion of students achieving Level 2 or higher in maths and English at Key Stage 4 prior to enrolment (as a proxy variable for academic attainment of previous cohorts)[31]

Achieving Excellence Area Indicators (available from the DfE):[32] [33]

- Access to a good secondary school index,[34]
- Achievement and progress indicators,[35]
- System leader coverage indicator,[36]
- Initial teacher training provider coverage index,[37]

[25] We will produce this variable taking the mean of all individuals per institutions present in the YPMAD dataset for academic years 2016/17, 2017/18, and 2018/19.

[26] Institution's number of Condition of Funding (CoF) students in 2016/2017.

[27] This variable will be estimated from estimated institution size and number of students under CoF for the 2016/17 academic year.

[28] As defined in the variable YPMAD_main_inst_type_lookup in the YPMAD dataset

[29] Variable YPMAD_Region_A23_main in the YPMAD dataset.

[30] Variables YPMAD_EverFSMage10to15 in the YPMAD dataset:

[31] Refers to results for the first-time students sit the examination. The variable used to construct this measure is in the is L2plusFuncAc_Eng_Maths, as well as institution size number of students.

[32] These indicators are measured at the local area level, and they work as a proxy for the availability of high-quality institutions in the area of interest. Since they were used to define the 6 categories of the AEAs and they are likely to be related to academic outcomes, we will be including them in our analysis.

[33] As the Achieving Excellence Areas (AEA) were based on two composite measures, each one of them constructed making use of other indicators, we will aim to recover the original data used to construct those measures. As the Achieving Excellence Areas were defined by the DfE, we expect to be able to obtain the original indicators used by the DfE to construct these measures. Should this not be possible, we will not be able to include variables used in the construction of the AEA categories in our analysis. Since some of the indicators were estimated using several sources of data, it would not be possible for the evaluator to reconstruct these variables in a timely manner and within the resources available for this evaluation.

[34] Proportion of secondary institution pupils in a given geographical area with access to a good school (school with a good or outstanding Ofsted rating). The district level outputs are produced by combining data held within DfE on pupil address with published school address and Ofsted rating data.

[35] For primary education, the DfE made use of the schools' 2015 Average Point Score in Reading, Writing and Mathematics as the attainment indicator, and the 2015 Key Stage 1 to Key Stage 2 value added score as the progress indicator. For secondary education the measures used were the Attainment 8 and Progress 8 indicators, respectively, also using 2015 data.

[36] The purpose of this indicator is to consider disparities in the capacity of areas to improve through school-to-school support. The indicator estimates the support available from system leaders within a 20km circle of an underperforming school, after all other demands on those leaders have also been factored in (technically termed 'the two-step floating catchment area method').

[37] The purpose of this indicator is to consider the capacity of areas to recruit new teachers. The indicator estimates the local supply of teacher trainees based on the location of providers and number of trainees they recruited and trained in 2015/16. The sources of data to produce these indicators are: Current System Leaders and their schools (DfE/NCTL); School quality and performance data (DfE/Ofsted); School locations and their pupil numbers (DfE). DfE/NCTL Initial teacher training census 2015-2016, DfE, Edubase data (all schools) and DfE, SFR 16/2015: Schools, pupils and their characteristics.

- Quality of leadership indicator,[38] and
- Academy sponsor coverage.[39]

The individual-level variables we will use in the model are:

- Student's Condition of Funding[40]
- Students' age[41]
- Gender[42]
- Ethnic group[43]
- Eligibility for free school meals[44]
- Special Education need status (SEN)[45]
- Prior attainment at KS2 in English and maths[46]

The covariates will be drawn from data from the academic year 18/19 (using 2016/17 data to produce some of the institutional characteristics). Note that the choice of these covariates is unchanged by the disruption caused by Covid-19. However, these covariates may be slightly adjusted in the process of checking and refining the propensity score matching model, which we describe in section 3.8.1.3 below.

In addition to the covariate data described above, we will also draw on data requested from the pupil-level School Census dataset held within the NPD, and the Individualised Learner Record (ILR) data. Variables from these datasets will be used to restrict the comparison group sample to exclude individuals who would have been ineligible for the BMP, if they had been part of a participating school in Category 5 and 6 AEA areas. To determine pupil eligibility, we will draw on the following variables:

- Maths condition of funding (available in the ILR)
- English condition of funding (available in the ILR)
- Pupil funding line (available in the ILR)
- Pupil enrolment status (available in the School Census)
- National Curriculum Year (available in the School Census)
- Maths and English GCSE highest prior attainment (available in the School Census)
- Maths and English GCSE funding exemption (available in the School Census)

Note that the ILR and School Census datasets cover different institutions in our sample, which is why both datasets are required. They also contain slightly different variables, so a slightly different selection will be used to derive pupil eligibility from each dataset. We have

[38] The measure uses Ofsted judgements on Leadership and Management to calculate the proportion of pupils in schools which have good or outstanding leadership. It is calculated separately for primary and secondary schools.

[39] The indicator is calculated using a similar approach as described above for System Leaders. It estimates the coverage by lead or outstanding schools in an academy sponsor chain, of maintained schools which are either below floor standards or, are inadequate or require improvement. The coverage is restricted to a range of 10km in urban areas and 20km in rural settings. These distances have been informed by the average geographic spread of schools in an academy chain.

[40] This variable is not currently available in NPD datasets. It is expected that it will be available from year 2018/19 in the YPMAD dataset. Otherwise, we will seek to obtain this from the DfE for both cohorts of interest.

[41] This variable will be derived utilizing the variable YPMAD_DOB in the YPMAD dataset.

[42] Variable YPMAD_gender in the YPMAD dataset.

[43] Variable YPMAD_EthnicGroupPLASC15 in the YPMAD dataset.

[44] At age 15, variable YPMAD_FSMEligibilityage15 in the YPMAD dataset.

[45] Variables YPMAD_SENStatusage15 in the YPMAD dataset.

[46] This corresponds to composite score available from the KS2 dataset (variables KS2_MATMRK for maths and the following 2 variables for language: KS2_GPSMRK and KS2_ENGWRITMRK)

discussed with the BMP team the best means of replicating the calculations that they did to determine eligibility in the treatment group, using these variables.

3.5. Sample size calculations

		Study Plan		
		Per treatment arm	Overall: All treatment arms	FSM: All treatment arms
MDES		0.21	0.12	0.29
MDE[47]		0.038	0.022	0.032
Pre-test/ post-test correlations	**level 1 (pupil)**			
	level 2 (class)			
	level 3 (institution)			
Intracluster correlations (ICCs)	**level 2 (class)**	N/A	N/A	N/A
	level 3 (institution)	0.20	0.20	0.20
Alpha[48]		0.017	0.017	0.017
Power		0.8	0.8	0.8
One-sided or two-sided?		2	2	2
Average cluster size[49]		7.96	7.96	1.91
Number of institutions	**Intervention**	145	436	436
	Comparison	145	436	436
	Total	290	872	872
Number of pupils	**Intervention**	16,520	49,675	11,922
	Comparison	16,520	49,675	11,922
	Total	33,040	99,350	23,844

The primary study outcome has been defined as a binary variable: whether or not the student achieves in maths as defined by the pilot. Arguably, the sample size calculation should be based upon the effect size estimated through the logistic model framework. Unfortunately, as noted by Demidenko (2007) ‘there is no consensus on the approach to compute the power and sample size with logistic regression’. Moreover, McConnell & Vera-Hernandez (2015), following Schochet (2013), argue that modelling differences in probabilities is more intuitive. Given the lack of consensus in dealing with power analysis for binary outcomes and, particularly, the lack of an agreed approach for complex logistic models with clustered designs, we have followed standard practice for differences in outcome probabilities, using the standard binomial variance [$p(1 – p)$] as a measure for the expected variance of the outcome[50] (e.g., McConell and Vera-Hernandez, 2015).

[47] Measured as a probability (so 0.05 is equivalent to 5 percentage points) assuming 24% pass rate on average and 17% for ever FSM students in a business as usual scenario (control group).

[48] Assuming Bonferroni correction for multiple hypothesis testing for 3 treatment arms.

[49] This corresponds to the harmonic mean of the sample. The simple average across institutions is 113.93 students per institution. For FSM students we assume a 24% prevalence rate.

[50] In this case p is the estimated probability of success, and the variance of the outcome is the variance in the outcome for the pooled sample (including observations in treatment and control groups).

We have chosen to estimate indicative minimum detectable effect sizes (MDES) using standard power calculation approaches. In these, we assume equal sizes of the treatment group and control group under a balanced experimental design.[51]

From Maynard and Dong (2013, p. 51), we have the following formula for a two-level hierarchical linear model with cluster-randomised assignment:

$$(1)\quad MDES = M_{J-g^{*}-2}\sqrt{\frac{\rho}{P(1-P)J}+\frac{(1-\rho)}{P(1-P)Jn}}$$

The MDES is the standardised minimum detectable effect size, i.e., the minimum impact we would expect the experiment to detect with the given sample size and design. Where $M_{J-g^{*}-2}$ is the sum of the t values corresponding to the appropriate levels of the Types I and II errors, with the appropriate degrees of freedom; ρ is the unconditional intra-cluster correlation at the institution level, J is the number of level 2 units (institutions), n is the average number of students per institution and P is the proportion of units assigned to treatment.

From Bloom (2006), the standardized mean difference effect size is:

$$(2)\quad ES = \frac{\bar{Y}_t-\bar{Y}_c}{\sigma}$$

Where σ^2 is the pooled outcome variance in the estimation sample. We can simply multiply the MDES, calculated through PowerUp, by σ to get the MDE, i.e., the minimum impact effect stated as a difference using the scale of a proportion. We know σ from the binomial nature of the data, i.e.,

$$(3)\quad \sigma^2 = \theta(1-\theta)$$

where we will take θ to be the proportion anticipated achieving success in the control group (business as usual).

Consequently, we calculate the MDE as follows:

$$(4)\quad MDE = \sqrt{\theta * (1-\theta)} * MDES$$

The table above presents the intention-to-treat minimum detectable effect size (MDES) as well as the minimum detectable effect (MDE): defined as the minimum detectable difference in the probability of achieving in maths (GCSE grade 4 or above, or Function Skills Level 2 depending on the student's prior attainment) between the treatment group and the comparison group, assuming a comparison group achievement rate of 24 per cent.

We also provide an MDES and MDE for young disadvantaged people, defined as those that have received free institution meals in the last six years (EverFSM-6), assuming a comparison group achievement rate in this sub-group of 17 per cent.

[51] With acknowledgement that the number of required comparison units may be notably higher based on an expectation that a considerable number of comparison units may be trimmed to achieve common support in propensity scores.

Our assumptions in calculating the MDE are as follows:

- 2-level multi-level model with allocation to treatment modelled as a fixed effect at level 2 (institutions), with institutions (level 2) allowed to vary randomly with respect to student outcomes (i.e. r_{ij}~(0,σ^2)).
- The principal outcome is a test of each of the three treatment groups against a matched control group on the proportion of students achieving in maths as defined by the pilot.
- Achievement in maths outcome will be defined as attainment of either grade 4 or above in GCSE maths for those with prior attainment of grade 3 in GCSE maths or attainment of Level 2 Functional Skills or GCSE maths grade 4 or above for those with prior attainment of grade 2 or below in GCSE maths, with information about both qualifications combined into a single outcome variable.
- One or more individuals in the control group per each individual in the treatment group, with acknowledgement that the pool of comparison cases available may need to be notably larger to maximise our ability to identify suitable units to work as a counterfactual for all treatment cases.
- MDES for a single treatment and control comparison. There was no important difference between treatment groups in terms of numbers of institutions or numbers of students per institution.[52] Consequently, a power analysis for one treatment group is equally relevant to all three treatment groups.
- We anticipate approximately 436 institutions participating of the intervention, equating to 145 institutions per treatment arm.
- We calculate and use a harmonic mean of 7.96 students for the average number of students per institution in the MDES calculation, which we assume to be constant across treatment arms.[53]
- We anticipate a control group achievement rate of 24 per cent (θ = 0.24) for the full sample (of all eligible students) based on GCSE attainment data for prior cohorts (EEF, 2019).
- We have three separate significance tests for the primary outcome, with each of the three treatment groups separately compared to the comparison group. However, to account for the three pairwise comparisons we assume a nominal Type 1 error rate of 0.05 with a two-tailed test, amended using a Bonferroni adjustment for three separate independent tests, to a value of 0.017 for each test. The Type II error rate is set at 0.2, giving power of 0.8.
- We assume an intra-cluster correlation coefficient (ρ) of 0.20[54].
- No level 1 and level 2 baseline covariates have been included in the MDES calculation in order to better align with the proposed non-experimental design and analysis approach and remain conservative in our estimation of the MDES[55].
- We assume no attrition in the design given that we expect to be able to retrieve data on all institutions that have opted in to participate in the research and eligible students within these from the NPD.

For FSM sub-group analysis:

- We present power calculations for disadvantaged students, defined as those students that have been eligible for free school meals at any time in the past six years before

[52] The maximum difference in the estimated MDES across treatment groups is less than 1 percentage point of a standard deviation.

[53] In practice it is 6.64 students for T1, 10.23 for T2 and 7.73 for treatment 3

[54] We expect a high intra-cluster correlation in the outcome given the fact that eligible individuals show similar scores before the intervention.

[55] As eligible individuals are only those students showing level 3 or below in their GCSEs before the intervention, we also did not expect much of the variation in the outcome to be explained by differences amongst individuals in their pre-intervention scores.

entering their post-16 institutions (based on the 'Ever-6 FSM' NPD variable). The MDES calculation combines all three treatment arms and assumes a 24 per cent prevalence of ever-FSM students per institution (EEF, 2019).

- We approximate the harmonic mean number of FSM students by estimating the expected number of FSM students in each institution (using the prevalence rate of 24% noted above) and calculating the harmonic mean using these estimated values.
- We anticipate a control group achievement rate of 17 per cent (θ = 0.17) among eligible FSM students based on GCSE attainment data for prior cohorts (EEF, 2019).

The presented estimates are preliminary and subject to change depending on the actual number of participating institutions and students that meet the intervention and evaluation criteria.

3.6 Selection mechanism

Understanding the factors that are related to whether institutions decide to participate in the programme is crucial when looking to establish a valid counterfactual for this evaluation. These factors are important to consider when selecting covariates for the analysis, which is described further in section 3.8.1.

We hypothesise that selection into treatment mainly depends on four main institutional characteristics: institution's location, number of eligible students, type and size of the institution.

3.6.1 Location: The Achieving Excellence Areas

The first requirement to be eligible for the programme is meeting the location criteria. That is, the institution of interest must be in a geographical area defined as Category 5 or 6 Achieving Excellence Area to be offered the Basic Maths Premium benefit.

The Achieving Excellence Areas (AEA) were established making use of several indicators related to school quality[56] (see data section for more detail on this measure). The main goal was to produce a measure that is a proxy of the level of availability of "good quality" schools for each geographical area in England.

Given the structure of this indicator (composite score), and the number and type of variables involved, we hypothesise that institutions located in Achieving Excellence Areas category 5 or 6, might not be substantially different to some institutions in AEAs category 4, for example. Since institutions in AEA categories 4 to 1 were not offered the intervention, they could provide a good counterfactual for this evaluation, as soon as they are similar to those institutions selecting into the treatment. Hence, the information utilized to construct these categories becomes crucial to establish a counterfactual for those institutions participating in the programme.

[56] Access to a good secondary school index, achievement and progress indicators, system leader coverage indicator, initial teacher training provider coverage index, quality of leadership indicator and academy sponsor coverage.

3.6.2 Number of eligible students

Since all institutions were offered the treatment in eligible areas, and many institutions rejected the offer (around 47% of sixth-form colleges and academies, and 27% of post-16 institutions), we hypothesise that there is a second important criterion when choosing to participate in the programme: the importance of the economic benefit.

Given the fact that 16-19 providers have on average a much larger number of eligible individuals[57] (those students meeting the Condition of Funding criteria) and that, at the same time, they were much more likely to take the intervention; we believe that institutions may have assessed whether they would have a large pool of individuals who could be eligible for the programme and made a decision based on that.

Even though the benefit is assigned per eligible student, having a larger amount of resources could influence the type and range of activities that could be implemented by institutions to improve the outcomes of interest.

Although this is a reasonable assumption, the fact that some institutions with a small number of individuals decided to take part of the intervention in all type of institutions, suggests that other reasons may also be important when taking this decision.

In this case, and to find a valid counterfactual, we should be looking for institutions showing a similar number of individuals under Condition of Funding in the control group.

3.6.3 Type of institution

The type of institution signals in a way the educational needs and characteristics of the student body, and therefore, also becomes influential when deciding whether the programme is a priority. Different types of institutions not only serve different subpopulations of students, but also show different organizational goals and capabilities that may affect their decisions about taking part in the programme. In this regard, we believe this institutional diversity can be partially captured through this characteristic.

3.6.4 Size of the institution

Institution size might also be indicative of the institutions' capabilities of running activities related to the implementation of the programme. Despite the number of eligible students, the existing staff and organizational structure could affect the decision to participate in the programme. We believe that larger institutions are more likely to take up the programme, even after taking into account the number of eligible individuals.

3.6.5 Other institutional characteristics: Institutions' disadvantage and previous academic attainment

Institutions' strategic priorities and expectations can be shaped by the level of social disadvantage of their communities as well as by their relative previous performance in standardized examinations. In this regard, we believe that both, institutions' average previous

[57] Based on our figures for academic year 2018/2019.

attainment as well as their proportion of students having received Free School Meals in the past, can be good proxy indicators for these two characteristics.

3.7 Selection of the comparison group and identification assumptions

Since all institutions in Category 5 and 6 AEAs were invited to participate in the programme, we will aim to find a comparison group of individuals attending institutions in AEA Category 4 (expanding to Category 3 AEAs and so on, if necessary) who are enrolled on a relevant course and not exempt from the English and maths condition of funding.

The offer of treatment to all institutions in Category 5 and 6 AEAs, and the possibility of rejection of such offer, immediately raises issues about the potential selection bias due to differences between treatment and comparison groups. This could affect internal validity, since units in the control group would not be strictly comparable, as unobserved characteristics affecting the outcome could differ across groups.

Specifying the mechanisms that explain selection into treatment can be challenging in non-experimental designs. A misspecification could affect the reliability of the results of this study. The first challenge is the possibility that individuals attending institutions in the potential comparison group are not necessarily a good counterfactual for the individuals in the treatment group. This could happen, for instance, if institutions in the comparison group show higher or lower scores on the outcome variable before the treatment takes place, or if institutions in this group show specific distinctive unobserved characteristics that could be correlated with the outcome of interest. In this case the treatment effect could vary according to those institutional characteristics. If this were the case the selection model would not be well specified, and hence the propensity score would not correctly account for the selection process. This could also happen if these unobserved characteristics are important in explaining the outcome of interest and are not included in the regression model (second stage of the doubly-robust estimation). The likely result is a biased estimator of the treatment effect.

Utilizing the propensity score is an effective way of constructing a counterfactual as soon as some key assumptions are met. In particular, the conditional independence assumption. The conditional independence assumption implies that selection into the treatment is solely based on observable characteristics, and that all variables that influence treatment assignment and potential outcomes simultaneously are observed. If this assumption holds, and after choosing the appropriate variables, the estimated propensity score removes bias when estimating the impact of the intervention under a doubly-robust method approach.

This assumption is especially strong in this trial, since institutions in the treatment group self-select into the treatment, and around 50% of institutions decided not to participate of the trial[58]. To reduce the probability of bias, we will make use of extensive information at the institution level, aiming to include most institution-level characteristics that are likely to influence both the self-selection process and the effectiveness of the intervention.

Of benefit to the evaluation is the fact that within AEAs, institutional characteristics will vary for the variables used to produce the defined areas. While we cannot ensure that the institutions in the treatment groups which voluntarily decided to participate in the trial are not

[58] Nonetheless, in 2016/17 around 95% of students in category 5 & 6 areas were in institutions taking part of the intervention.

systematically different in any unobserved way to those institutions in the comparison group, we anticipate that we will find suitable institutions in the control group that could work as a counterfactual among institutions in AEAs classified in categories other than 5 and 6.

Similarly, for the regression model to be well specified, it requires that both the variables involved in the selection process as well as those determining the effectiveness of the intervention are included in the model. Technically this means that conditional on the set of observed covariates included in the regression model, the assignment of the treatment between treatment and control groups can be assumed as random, and no relevant unobserved individual or institutional-level characteristics have been omitted from the model. If the model is well specified, the coefficient for the variable of interest (treatment indicator) is unbiased and represents the true impact of the programme in the outcome of interest.
In this specific study we are assuming that this is the case, so the comparison group will be strictly comparable to those in the treatment groups conditional on the observed covariates.[59] In that case, a valid comparison group can be used to estimate the impact of the intervention.

Note that since treatment groups were not revealed to institutions until they decided to participate in the trial, we assume that the selection mechanism (self-selection into treatment) taking place for all institutions was similar for institutions in all treatment groups. This implies that a single comparison group can be drawn for all treatment groups.

Alongside specifying the factors associated with the decision to participate in the trial, it is also necessary to restrict the comparison group to only individuals who *would have been* eligible for the BMP funding if it had been offered. If we include students in the comparison group who would have actually been ineligible for the BMP had they been offered it, this risks introducing bias into the impact estimation. This means that we need to construct a measure of hypothetical BMP eligibility for individuals in Category 4 AEA areas (and 3, if the comparison group sample extends to this) that exactly matches the actual eligibility criteria for those in the treatment group. Individuals in the Category 5 and 6 AEA areas were eligible for the BMP if they had not yet passed Maths GCSE by the end of the 2017-18 academic year and were required to re-sit. Individuals were required to re-sit if they were exempt from the English and Maths condition of funding.

Information about whether or not students had passed Maths GCSE by the 2017-18 academic year is held in the NPD datasets. Information about the condition of funding status can be derived from two other sources: the pupil-level School Census dataset that is part of the NPD, and ILR data. These two datasets contain the variables that were initially used by the BMP team to determine eligibility for the treatment group. We need to use both because they cover different samples of institutions. While the School Census covers schools and academics only, the ILR contains data from other institution types, including Further Education colleges, sixth form colleges and others.

[59] In this evaluation, it is expected that those institutions chosen as the comparison group from Category 4 areas will not greatly differ in relevant variables from those in Categories 5 and 6 which are part of the pilot, so they can act as an effective counterfactual. The plausibility of this assumption will be examined as part of the propensity score estimation, where institution level and aggregated student-level characteristics will be compared between institutions and individuals in the control groups and those in the treatment groups. As the definition of AEAs is based on 11 different indicators, some of them unlikely to affect the outcome of interest, we believe it is possible to achieve balance in those characteristics affecting the outcomes between institutions in Category 4 areas and those in categories 5 or 6. Although there could be common support issues affecting the external validity of our findings, we expect that at least for a sample of institutions we will be able to find equivalence in observed pre-treatment institution- and individual-level characteristics across groups.

3.8 Primary outcome analysis (RQ1)

Our primary analysis will estimate the impact of the intervention using the doubly-robust method, as detailed in Funk et al. (2011). This method combines propensity score estimation with regression analysis to derive the doubly-robust impact estimator. A key advantage of this method is that it is robust to potential misspecification of either the propensity scores or the outcome regression model, providing a safeguard against potential bias of the impact estimate.

We now describe each of the steps involved in this method in turn.

3.8.1 Propensity score estimation

3.8.1.1 Propensity score model

The first step in constructing the doubly robust estimator is to estimate propensity scores for all individuals in the sample. Propensity scores were first proposed by Rosenbaum and Rubin (1983) and have a wide variety of applications in causal inference. Formally, propensity scores represent the conditional probability of being exposed to the BMP intervention, given a set of observable characteristics. Another way to interpret propensity scores is as a 'balancing score'. The central idea (and one of the core assumptions behind the theory) is that conditional on having similar propensity scores, there should be no systematic differences between individuals in the treatment and control groups that would otherwise lead to bias in impact estimation. The assumptions underlying propensity score theory are set out in Rosenbaum and Rubin (1983).

We will estimate the propensity score at the individual level, with a binomial outcome taking the value one for individuals in the treatment group and zero otherwise.[60] The propensity score will take the following form:

$$\text{(5)} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta_1 X_{ij} + \beta_2 S_j$$

Where π_{ij} is the probability that individual *i* in setting *j* is exposed to the BMP intervention, $\boldsymbol{S_j}$ are institution level observed characteristics and $\boldsymbol{X_{ij}}$ are individual level observed characteristics. The predicted values from this model for each individual in the sample form the propensity score.

3.8.1.2 Specifying covariates

We will include a range of covariates in the propensity score model. Suitable covariates should be those that would be potential confounders in the impact estimation if not suitably accounted for. In simple terms these include variables that are hypothesized to determine both the probability of being exposed to the BMP and the probability of passing the GCSE resit (or equivalent). It is also important that the covariates could not themselves be influenced by the intervention. We therefore focus on variables that were defined prior to the start of the BMP (baseline covariates), or that are otherwise expected to be unaffected by the intervention.

[60] We assume that all individuals from all treatment arms together will make up the treatment group during the propensity score estimation stage.

We will use KS4 results, data from the Young Person's Matched Administrative Dataset (YPMAD), public school census data, and Achieving Excellence Area indicators from the DfE as the source of matching covariates to construct the propensity score model. The list of institution and individual-level covariates that we will use is given in section 3.4.2. The choice of covariates has been informed by our theoretical understanding of the characteristics that are related to both the likelihood of being exposed to the BMP and to the outcome. This includes the institutional characteristics associated with selection into treatment set out in section 3.6 (including the factors used to determine the Achieving Excellent Area categories). We have included both institution and individual-level covariates since both could serve as confounders when estimating the intervention effect.

3.8.1.3 Propensity model diagnostics

Having defined propensity scores, the next steps are to use these to define the comparison group and carry out diagnostic checks. The doubly-robust approach is equivalent to an inverse-probability of treatment weighting (IPTW) estimator augmented by a regression model (Funk et al., 2011). Checks that apply to IPTW estimators also apply to our approach and can be carried out before we obtain access to outcomes data.

If the propensity score model has been well defined (and satisfies some required assumptions), we will expect to observe minimal observed differences in the distribution of covariates between the intervention group and the comparison group after weighting by the IPTW. We will carry out the following checks to understand whether this is the case, and the propensity modelling approach has been successful:

1. Check covariate balance after matching

    - We will check the comparability of the treatment and control group after weighting, in terms of the observable covariates.
    - Given the large sample size, even miniscule difference between groups would be “statistically significant”, so we will do this by assessing the standardized mean differences in each covariate between the treatment and control group after matching. We will consider a standardised mean difference of 0.1 or above to indicate that there is a remaining imbalance in the distribution of the covariate between the treatment and control group after matching.
    - We will retain tolerance for some standardised differences in our selected matching model – that is, we will not require a less than 0.1 difference across all covariates. However, we will expect to find an overall reduction in standardised differences after matching, and if half or more of differences over 0.1 remain after matching then we will adjust the model.

2. Assess common support

    - One of the assumptions underpinning propensity score modelling is that all observations in the created control sample should have a non-zero *a priori* probability of being exposed to the intervention or control. This assumption can be intuitively thought of as meaning that the two groups need to be sufficiently comparable – i.e., that there must be some ‘overlap’ between them, which is known as common support.
    - We will check common support firstly through a visual representation of the distribution of propensity scores between the intervention group and control group, to check that there are not sizeable differences between them. We can also count how many observations are found to be ‘off support’ after matching.

3. Assess the pseudo-R^2
   - The pseudo-R^2 is a measure of the extent to which the logit model estimating propensity score has predictive value for explaining exposure to treatment. Sianesi (2004) recommends that propensity scores are re-estimated on the matched sample, in such a way that the pseudo-R^2 can be compared both before and after matching. The intuition is that if matching has worked well, there should be few differences in the distribution of covariates between the treatment and control groups, and therefore the pseudo-R^2 will be low. We can do the same using a model weighted by the IPTWs.
   - We will check whether there has been a reduction in the pseudo-R^2 after matching. A value of close to 0 after in the matched sample is a good indication of good balance between the treatment and control groups.

If we observe a high pseudo-R^2 , a large number of units 'off support', or considerable standardized differences between the treatment and comparison groups remaining over a range of covariates after weighting, this could indicate that the propensity score model has not been specified correctly. We will therefore adjust the model to increase the ability of the propensity score approach to select a suitable counterfactual that is well-comparable to the treatment group.

We will make this adjustment in two stages. First, we will expand the set of possible units from which the comparison group may be created. Our first model will use institutions from Category 4 Achieving Excellence Areas as the source of the potential comparison group. We will expand this to include Category 3 institutions in the first instance if the success of matching appears poor.

If we still have concerns with the success of matching after expanding to the Category 3 Achieving Excellence Areas, we will make further adjustments as follows:

- Re-running the matching model using trimming options to drop observations with extreme propensity scores, to improve common support if this is an issue.
- Re-visiting the set of covariates to assess whether using transformations or interaction terms in the model improves balance and re-assessing the theoretical case for inclusion of each covariate to understand if there are any key factors missing from our model, or superfluous variables included.

We will maintain a record of all the propensity modelling specifications used, the results of the checks to the success of the matching, and a record of what decisions were taken to adjust the matching model if this is needed, and why. Since these propensity score models will be fitted before we have access to the outcomes data, there is no possibility of "data dredging" to influence the estimated intervention effect.

3.8.2 Regression estimation

After estimating the propensity scores, we will run regressions to model the relationship between the outcome and intervention exposure, conditional on the covariates. We will run three separate regression models, i.e., each treatment group compared separately to the comparison group, where the significance level of the test will be conducted using the adjusted Bonferroni level[61].

We propose to use a binomial multilevel logistic regression model of the outcome π_{ij}, with students at Level 1 with observed characteristics X_{ij} nested within institutions at Level 2 with observed characteristics S_j. These student-level and institution-level characteristics will be the same as those set out in section 3.4.2, which are used in the model of propensity scores.

Institutions will be allowed to vary randomly (with u_j as a random effect). The regression residual (e_{ij}) will have a zero mean with a Gaussian distribution.

The basic form of the model is[62]:

$$(6)\quad \boldsymbol{logit(\pi)_{ij} = \beta_0 + \beta_1 treatment_j + \beta_2 X_{ij} + \beta_3 S_j + \mathbf{u}_j + \mathbf{e}_{ij}}$$

We will make use of the estimated coefficients $\widehat{\beta_\iota}$ to estimate the expected value of the outcome ($\widehat{\pi_\iota}$) for all individuals in the control sample.

To check whether they have a high predictive value explaining the outcome[63], we will also explore whether polynomials (up to third degree) of these variables and/or interaction terms could be included in an alternative specification.

3.8.3 Doubly-robust estimator

Finally, the ITT estimate for each treatment group is captured by estimating an average of the estimated impact of the intervention for each individual in the sample ($\hat{\tau}_{DR}$ in the equation below).

The general specification for the doubly-robust estimator can be found in equation 7 (from Lunceford and Davidian, 2004), where A_i signals exposure to the programme (1 or 0) for individual i, and Y_i indicates the outcome of interest for individual i (observed outcome); $m_1(\underline{X})$ accounts for the predicted outcome under exposure to the treatment, $m_0(\underline{X})$ the predicted outcome under no exposure to the treatment; and $\widehat{p_\iota}$ represents the estimated propensity score for individual i.

[61] The evaluation involves three randomised treatment groups and a matched comparison. An aim of the evaluation is to test each treatment separately against the control group to determine the efficacy of each treatment. Consequently, we will run three paired treatment-control tests. In order to maintain a five per cent significance level for each test, a standard Bonferroni adjustment was applied to account for the three tests in each of the families. This resulted in a value of 0.0167, i.e., α/3, and only tests showing a significance level below this threshold will be accepted as statistically significant at alpha = 0.05.

[62] We will run separate models for each treatment group as follows:

$$\boldsymbol{logit(\pi)_{ij} = \beta_0 + \beta_1 treatment1_j + \beta_2 X_{ij} + \beta_3 S_j + \mathbf{u}_j + \mathbf{e}_{ij}}$$

$$\boldsymbol{logit(\pi)_{ij} = \beta_0 + \beta_1 treatment2_j + \beta_2 X_{ij} + \beta_3 S_j + \mathbf{u}_j + \mathbf{e}_{ij}}$$

$$\boldsymbol{logit(\pi)_{ij} = \beta_0 + \beta_1 treatment3_j + \beta_2 X_{ij} + \beta_3 S_j + \mathbf{u}_j + \mathbf{e}_{ij}}$$

[63] High predictive value in this context means explaining a sizable proportion of the variance in the outcome of interest and/or impacting importantly on the estimated ITT.

(7) $$\widehat{\tau}_{\mathrm{DR}} = \frac{1}{N}\sum_{i=1}^{N} \frac{A_iY_i - (A_i - \widehat{p}_i)m_1(\underline{X}_i)}{\widehat{p}_i} - \frac{1}{N}\sum_{i=1}^{N} \frac{(1 - A_i)Y_i + (A_i - \widehat{p}_i)m_0(\underline{X}_i)}{1 - \widehat{p}_i}$$

We will estimate bootstrapped standard errors for this estimator, as recommended in Funk at al. (2011). The model will be estimated using the *teffects aipw* command in Stata.

3.8.4 Analysis of relative effectiveness (RQ2)

In addition to the analysis of the impact of each intervention when compared to a control group, we will evaluate whether there are differences in the relative effectiveness of the three treatment conditions to address Research Question 2:[64]

We propose to use the estimated coefficients for the impact of the intervention for each treatment arm and their standard errors, analysing whether differences between arms are statistically significant and at what significance level.

In order to implement this analysis, we will estimate a single regression model including a dummy indicator per each treatment arm. We will report the estimated ITT (coefficient) for each arm[65] and their standard errors. We will also report predicted values (probabilities) for each treatment arm when all the covariates in the regression model take their mean values.

We will then compare the estimated impact per treatment arm (ITT coefficients) and their respective confidence intervals. Confidence intervals will be estimated using the adjusted Bonferroni level[66] for multiple hypothesis testing. We will report the estimated ITT differences between treatment arms[67] and whether they are statistically significant at 95% confidence level after accounting for the Bonferroni correction.

3.9. Robustness checks

To check for the validity of our results under different methodological approaches, we will implement a two-stage propensity score matching to estimate the effect of the intervention for all three treatment arms.

We will initially match institutions in the treatment groups to one or more institutions in the control group[68] (using a radius matching with replacement approach, with a radius of 0.1 standard deviations on the propensity score).[69] After institutions are matched, we will perform matching at the student level, matching each student in the treatment group to similar students

[64] RQ2: What is the effectiveness of the different funding models relative to each other?

[65] In practice this means estimating the predicted values when each treatment indicator takes the value one while the other treatment indicators take the value zero.

[66] In order to control for family-wise error-rate a standard Bonferroni adjustment will be applied to account for the interaction terms. This results in a value of 0.017, i.e., α/2, and only treatment coefficients showing a significance level below this threshold will be accepted as statistically significant at alpha = 0.05.

[67] Treatment 1 vs. Treatment 2; Treatment 2 vs. Treatment 3; and Treatment 1 vs. Treatment 3.

[68] In this case, and as for the main analysis, we will only be using institution- and individual-level data from the academic year 2018/19 to construct the control group.

[69] Radius matching is a variant of calliper matching. The basic idea is that by using the propensity score we utilize all observations in the control group which are within a certain distance from the matched unit. As opposed to the nearest neighbour with calliper method, the idea is to use not only the nearest neighbour within each calliper, but all of the comparison members within it. During matching, matched units in the control group are replaced in order to ensure the biggest possible number of institutions within each calliper.

from the group of matched institutions to each specific institution (also using a radius matching with replacement approach utilizing the same radius size).

We will be reporting the number of times that institutions in the control group were matched, in order to investigate whether matching without replacement is a more suitable approach.[70]

The rationale for the student level matching is twofold. First, it is possible that important institution level data may not be available for matching and any such omission may be partially compensated for through student level data. Second, any random allocation of the treatment can by chance result in covariate imbalance between treatment groups with any given sample.

Matching at the student level can help by ameliorating any such chance occurrences of student level covariate imbalance, which may happen even in the presence of acceptable institution level covariate balance. The two-stage level matching also ensures that those matched individuals in the control group belong to institutions similar to those where individuals in the treatment group are enrolled.

3.9.1 First stage: Estimating the propensity score at the institution-level

To select the group of institutions that will be part of the control group under this robustness check, we will first estimate a propensity score for all institutions in the sample[71]. By comparing the distribution of the propensity score of institutions from category 4 AEAs institutions to that of the treatment group (from category 5 and 6 AEAs) we will assess the extent to which we obtain common support.[72]

To estimate the propensity score for the matching at the institution level, and for each treatment group separately, we will use a logistic model to estimate the probability (P) of each institution being in each treatment group, conditional upon their scores across the set of institution level matching covariates. As participation in the treatment group can be modelled through a binary outcome, we initially assume that participation in the treatment group (T_j=1) is strongly associated with certain observed institution level characteristics (S_j). Once an appropriate logistic model has been estimated using the observed data (retrieving the regressions coefficients), the propensity score is calculated as:[73]

(8) $\pi_j = P\left(T_j = 1 \middle| S_j\right)$

Where the logistic model is:

(9) $\log\left(\frac{\pi_j}{1-\pi_j}\right) = \alpha + \beta S_j$

To estimate the propensity score, we will make use of a range of variables; including, but not limited to those used in constructing the AEA groups (where appropriate) and student-level

[70] This in case a large proportion of institutions were used more than once and concerns about the validity of the estimated standard errors arise.

[71] Initially including all institutions in category 4, as well as institutions in categories 5 & 6.

[72] This will also allow us to identify for each institution in the treatment groups, a set of institutions in category 4 AEAs (or below) that could work as a counterfactual.

[73] In this case, after the model is estimated, and for each institution separately, we make use of the fitted values for the coefficients in the logistic regression to estimate the propensity score. We would estimate one separate propensity score per each treatment arm.

aggregated measures of academic ability, socioeconomic status, and other demographic characteristics. A comprehensive list of the institution level variables used for the implementation of matching can be found in section 3.4.2.

We will adopt the same checks as those outlined in section 3.8.1.3 to assess the success of the institution-level matching and include a record of the different specifications used during the iterative process, including a short description of the rationale behind the decisions made, reporting them in an appendix of the report.

If common support is still not achieved, the primary analysis will proceed with the exclusion of treatment group institutions outside the common support area. In such a case, we will clearly report how this restricts the interpretation of the effect estimate and generalisability of the findings.

If the institution common support looks appropriate[74] at this stage (achieving a good balance in the variables of interest between treatment and control group institutions) we will proceed to the next stage: matching institutions based on the propensity score.

3.9.2 First-stage matching process

A wide variety of matching techniques are available, both for estimating the propensity score and then, subsequently, for undertaking the matching based on the propensity score.[75][76] As we are interested in finding more than one institution (or individual) that could work as a counterfactual, we will be using radius matching.

We propose to use radius matching since it is a straightforward approach, shows certain levels of flexibility by allowing the researcher to define a calliper size, and usually shows similar results to other robust matching techniques, such as Mahalanobis matching (Rubin & Thomas, 2000). Radius matching is also easier to understand, and as opposed to kernel matching, takes a shorter time to be implemented. We dismissed nearest-neighbour matching since previous literature has shown that it is prone to produce less efficient estimates of the effect (Caliendo & Kopening, 2005).

We will be using radius matching estimating the same model using two different callipers.[77] The size of the calliper should not be larger than 0.25 standard deviations in the propensity

74 We define common support as appropriate if the differences in means in observed characteristics (listed in section 3.4.2) between individuals in the treatment and control groups before the intervention are not larger than 0.1 standard deviations for at least 50% of the variables under analysis. We also expect that at least a range of 50% in the propensity score is covered by the common support area.

75 We use the term propensity score broadly here, and other distance metrics (e.g. Mahalanobis distance) could be substituted.

76 It is common practice to explore more than one matching approach, but it is not practically feasible to attempt all combinations of estimation and matching approaches. For our preferred approach we will limit the analysis to one estimation technique (radius matching), although we propose an alternative approach depending on the matching results (stratification matching).

[77] To estimate the group of matched institutions per treated unit, we will define a calliper or bandwidth for the propensity score from which institutions in the control group can be selected. Selecting more than one unit will allow a bigger chance of finding matched units in the case of institutions. It will also allow for more than one unit in the individual level matching to be part of the matched sample, allowing for more efficient estimates of the ITT later on.

score as recommended by Lunt (2013).[78] We will use descriptive statistics to assess how many of the institutions in the treatment groups were actually matched to institutions in the control group, also assessing balance in the covariates used to estimate the propensity score.[79]

After the propensity score for the matching at the institution level is identified we will match institutions in the treatment group to institutions in the comparison group, by using a 1-n radius matching methodology. Following Arpino and Cannas (2016) notation, let I_1 and I_0 denote the set of treated and comparison units, respectively, and let A_{rj} indicate the set of comparison units matched to unit $j \in I_1$ (note that j indicates a generic cluster j while kj indicates a generic untreated cluster k linked to cluster j):

(10) $A_{rj} = \{kj \in I_0 : \hat{e}_{kj} = |\hat{e}_{kj} - \hat{e}_j| < \mathrm{C}\}$

Where C is the defined capiler during the institution matching process. Then the matched sample M of institutions will be represented by all units in the following group.

(11) $M = \{rj: A_{rj} \neq \emptyset\} \cup \{\bigcup_{rj} A_{rj}\}$

3.9.3 Second-stage matching process

In the second stage, and once we have an appropriate quality match at the institution level, we will proceed to match eligible individuals in treatment institutions to eligible students in the comparison group. We will estimate the propensity score for all eligible students in the treatment and comparison groups. For each student in each treatment group institution, we will aim to find one or more individuals in their set of matched institutions that could work as a counterfactual.

To perform this matching, we will make use of the propensity score at the individual level, with a similar specification to that of the institution level. For individual i in institution j, the propensity score is the following[80].

(12) $\pi_{ij} = P(T_{ij} = 1|X_{ij})$

Where X_{ij} are individual level observed characteristics. In this case we will be comparing individuals in the treatment institutions (T_{ij}=1) to individuals in matched institutions who belong to the matched sample A_{rj} for institution *i*.

[78] We propose an initial calliper of 0.1 standard deviations on the propensity score, with a second estimation using a calliper of 0.2 standard deviations. Depending on the PSM results (and the number of matched institutions per institution in the treatment groups), some flexibility on the calliper could be allowed, as soon as covariate balance is achieved (differences between individuals in the treatment and control groups in the matched sample are not larger than 0.1 standard deviations for at least 50% of the variables under analysis.

[79] We propose to estimate means for all institution level variables listed in section 3.4.2. Variables will be standardized and subsequently we will check differences between treatment and control group and their statistical significance. As we will be estimating the intention to treat (ITT), all randomised institutions will be included in the analysis, irrespective of whether they complied with the interventions.

[80] Where the logistic model is the following: $\log\left(\frac{\pi_j}{1-\pi_j}\right) = \alpha + \beta X_{ij}$

The information used for matching at the individual level will correspond to student-level variables taken from NPD datasets. We propose using individual level variables listed in section 3.5.3. Once we have performed the matching, we will assess covariate balance both at institution and student level[81], especially focusing on whether or not the inclusion of individual level variables in the second stage of the matching process improves the balance between treatment and control group samples.

Estimation of the ITT effect

Using weights from the matching stages, the intention-to-treat effect can be estimated as follows:

$$\text{(13)} \qquad \widehat{ITT} = \frac{1}{card\ (I_1)} \sum\nolimits_{i\,\epsilon\,I_1} \left(Y_{ij} - \sum\nolimits_{ki\,\epsilon\,I_0 \cap M} Y_{kij} * w(ij, ki) \right)$$

This ITT estimates the effect of each one of the three treatments on the outcome of interest, corresponding to research questions (RQ1) of the hypotheses section[82] where Y_{ij} is the outcome of interest for unit *i* in institution *j*, and Y_{kij} is the outcome for the *k*th matched unit (at the individual level) in the comparison group to unit i in institution j of the treatment group. The weight, *w(ij, ki)*, is the weight assigned to the comparison unit ki in the estimation of the unobserved potential outcome, Y(0), for the treated unit i in institution j.

In principle, each unit in the treatment group is matched to one or more units in the comparison group within the defined calliper. For the case of radius matching, the initial selected weights will be 1/Nij, where Nij is the number of selected units for each treatment unit i in institution j.

To take into account the clustered structure of the data, we propose to estimate the standard errors for the effect using a clustered version of the nonparametric bootstrap, proposed by Abadie and Spiess (2016).[83].

3.9.4 Alternative approach: Stratification Matching

Although a two-stage matching approach is arguably one of the best ways to find a suitable counterfactual for each individual in the treatment groups, as it matches comparable institutions and then individuals within matched institutions (see Zubizarreta & Keele, 2017); it is possible that matching will not be successful in terms of the number of units in each treatment institution with matched units in institutions in the comparison group.

[81] We will estimate means and standard deviations for all institution and individual level variables listed in section 3.4.2 for the final sample of matched units. Variables will be standardized and subsequently we will check differences between treatment and control group and their statistical significance.

[82] In this case we will be using one sub-sample of comparison group observations for each treatment arm after the matching process. The matched samples should differ slightly across comparison groups, despite the fact that institutions were randomly selected to the different treatment arms.

[83] According to Lechner (2002) estimates there is no important difference in the estimated standard errors using bootstrapping or an analytical formula accounting for repeated units when implementing matching.

If after an assessment the number of matched institutions and individuals is below a defined threshold[84] and/or the number of variables[85] with differences in means over 0.1 standard deviations is greater than 50%, we will implement stratification on the propensity score as an alternative approach to account for institution level characteristics when matching individuals.

For the stratification on the propensity score we will make use of an approach similar to that of Hong & Raudenbush (2006). In this case initially we estimate the propensity score at the institution level following exactly the same procedure as the one detailed in equations (8) and (9). Once the propensity score is estimated, we eliminate units (institutions) outside the common support area following the criteria detailed in section 3.9 (first stage matching). We will estimate deciles[86] on the institution-level propensity score and check balance in institution and individual level characteristics. Once this is finished, we will save the corresponding institution strata for each individual in the sample.

On a second stage, we will estimate the propensity score at the individual level utilizing individual level and institution level covariates[87]. The individual-level matching will take the form:

$$(14) \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta_1 X_{ij} + \beta_2 S_j$$

Where π_{ij} is the propensity score for individual *i* in institution *j*. S_j are institution level observed characteristics and X_{ij} are individual level observed characteristics described in section 3.5.3. In this case, again we will drop units (individuals) outside common support. We will save the propensity score for all remaining individuals in the sample, aiming to use it during the ITT estimation.

Finally, and following a specification similar to that of Hong and Raudenbush (2006), we will estimate the effect by running a logistic regression where individuals in the treatment groups are compared to those in the control group, conditional on their strata on the institution-level propensity score. The specification also includes the logit of the propensity score at the individual level as a covariate, in order to account for possible imbalance at the individual level within strata. The preferred specification is as follows.

$$(15) \qquad \boldsymbol{logit(\pi)_{ij} = \beta_0 + \beta_1 treatment_j + S_j + \sum_{k=1}^{10} \beta_k Q_k + Logit\left(\hat{e}_{ij}\right) + u_{ij}}$$

Where S_j is an institution random effect, Q_k represents dummies for each strata on the propensity score at the institution level, and $\boldsymbol{\hat{e}_{ij}}$ represents the propensity score at the individual-level estimated on the second stage. In this case the ITT will be retrieved from the coefficient β_1. Robust standard errors will be estimated accounting for clustering at the institution level. We will run separate models for each treatment group.[88]

[84] We propose to use a threshold of 70% of institutions and 50% of students as soon as there is balance on covariates after matching.

[85] This corresponds to the variables used to estimate the propensity score.

[86] This could be lowered to quintiles in case the number of institutions per decile is too small and balance on individual and institution covariates is not affected importantly.

[87] The specification for the propensity score at the individual level will be exactly the same as the one used in the robustness checks stage.

[88] Treatment groups 1, 2 and 3.

3.10 Further Analyses

3.10.1 Subgroup analyses (RQ3)

To address Research Question 3, we will perform an analysis similar to that of our main approach (primary analysis). We will investigate whether the estimated effect of the intervention on the primary outcome varies according to the following individual/institution level characteristics: social disadvantage (measured at student level), number of eligible students (at the institution level), and GCSEs previous attainment (at the student level).

The criteria utilized to define these variables are detailed below:

(1) Social disadvantage (defined as those funding-eligible students that have been eligible for free school meals at any time in the past five years before entering their post-16 institution. Based on “FSM’ NPD variable).[89]

(2) Number of students receiving funding through the basic math premium scheme (this measure reflects the amount of resources available, or potentially available, to implement the activities related to the programme in each institution).[90]

(3) Student GCSE attainment at KS4, to explore whether the intervention and the more stringent definition of achievement for students with a prior achievement of grade 3 in GCSE maths have differential effects on student outcomes compared to impacts on students with a prior achievement of grade 2 or below in GCSE maths.[91]

We will be estimating the effect of the intervention for all treatment groups together (in order to gain statistical power). However, and as a robustness check (detailed below), we will check whether our results vary according to treatment group, learning whether any identified effect is driven by individuals who belong to one of the treatment arms. This analysis will be included as part of the subgroup impacts section after the main analysis (all treatment together).

As for the primary analysis, we will be estimating and reporting the impact of the intervention for the three different samples of students described in section 3.3 (Cohort 18/19 funding-eligible students).

We will use the same techniques than those detailed in the primary analysis (doubly-robust method), estimating the propensity score for individuals in all treatment groups together (the treatment group in this case); as well as for individuals attending institutions in the pool of potential control units (in Category 4 AEA areas or below, utilizing cohort 18/19 data to produce the counterfactual).

Once we have estimated the propensity score, we will implement a regression analysis. Apart from an additional interaction term (detailed below), we will specify exactly the same model as for the primary analysis, including the same explanatory variables (see equation 15). For instance, the estimated model for the FSM subgroup analysis will have the following specification:

[89] Variable YPMAD_EverFSMage10to15 in the YPMAD dataset.

[90] Institution’s number of Condition of Funding (CoF) students in 2016/2017. This variable was made available to the evaluator by DfE.

[91] This variable can be obtained recoding the variable YPMAD_GCSE_GRADE_MATHS from the YPMAD dataset. It could take a binary value or a continuous value.

(16) $$\boldsymbol{logit(\pi)_{ij} = \beta_0 + \beta_1 treatment_j + \beta_2 treatment_j * FSM_{ij} + \beta_3 X_{ij} + \beta_4 S_j + u_j + e_{ij}}$$

In this case β_2 represents the coefficient for an interaction term between the treatment and student disadvantage status (FSM). We will estimate this coefficient and its standard error making use of the doubly-robust method.

A similar analysis will be performed for institutions' number of students eligible for funding, and individuals' previous GCSE results as detailed above.

3.10.1.1 Robustness check

Since our main sub-group analysis will combine all three treatment arms into an overall treatment group for all estimations; as a robustness check, we will be evaluating the existence of a differential impact according to the institution/individual level characteristics of interest, but for each one of the treatment arms separately.

For all three variables of interest (student FSM status, student previous GCSE results and institution-wide number of students eligible for funding) we will estimate a model similar to that specified in equation 18. However, in this case we will only incorporate units belonging to the respective treatment group (A, B or C).

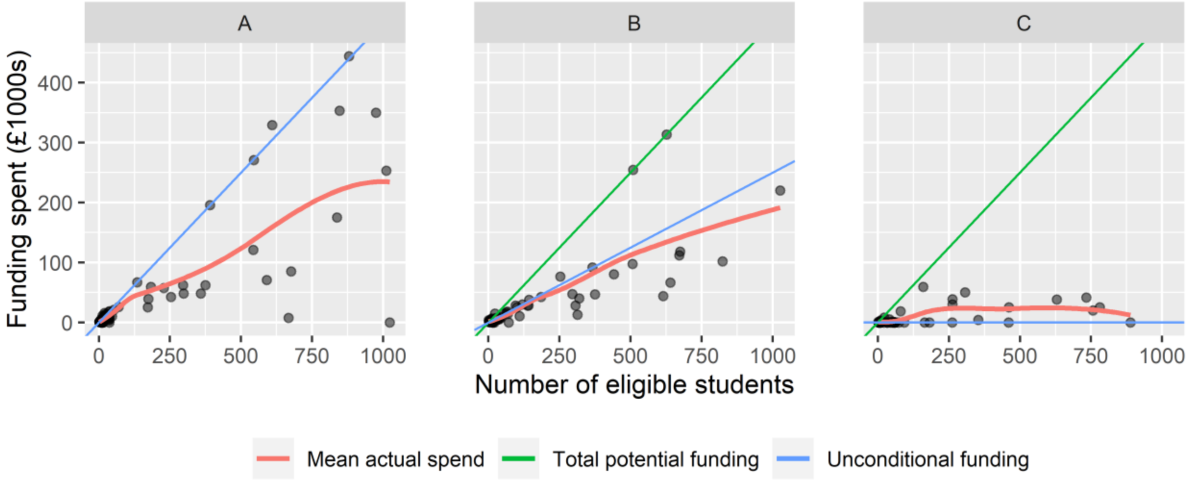
Initially, and for each treatment group, we will re-estimate the propensity score, to account for likely differences in observables characteristics between subgroups. As the assignment to the treatment was random, we do not expect important differences in institution observed characteristics across treatment arms. However, it is likely that samples in the control group will vary slightly after estimating the propensity score used for the estimation of the ITT.

After the propensity score is estimated we will estimate the regression model per each treatment group separately, reporting effect sizes as relative risks, as well as confidence intervals. It is anticipated that given the smaller sample size for each treatment arm, some statistical power issues may arise when interpreting our results. In that case, results may prove to be only indicative.

3.10.2 Dosage analysis (RQ4)

Data on the amount spent by post-16 settings is being collected at two time points, using surveys with school staff. Data on funding spent to date collected from the second survey (conducted in November 2019) has been plotted against the total amount of funding available (both conditional and unconditional) in Figure 1. As indicated by Figure 1, the amount spent by post-16 institutions varies both by the number of pupils eligible for the Basic Maths Premium, and by the funding model they are allocated to.

Figure 1: Total amount spent (£000s) by post-16 settings by treatment allocation



It is possible that the effectiveness of the Basic Maths Premium may depend on the total amount of additional resources available for institutions to spend (as opposed to the exact amount provided per eligible pupil). For example, an institution may wish to hire an additional teaching assistant to support maths tuition. However, only institutions with a sufficient number of eligible pupils will have the necessary additional funding to be able to fund additional staff. This may restrict the choice of schools with fewer eligible pupils.

To assess whether the amount of money spent in the 2018/19 and 2019/20 academic years impacts on attainment by November 2019 we will conduct two analyses; a regression analysis and marginal effects. For both analyses, intervention groups will be combined as the funding mechanisms will likely be a strong determinant of the amount of money spent within institutions.

The regression analysis will be performed using the matched sample, incorporating all the covariates from the primary analysis (the “doubly-robust” covariates) and a categorical variable for bands of income spent. The primary analysis will focus on income spent during 2018/19 as this covers the period before the first maths exam attempt and any resit in November 2019. We will additionally explore whether any additional spend for 2019/20 is associated with outcomes since, from IPE findings, we know that over half of institutions said they make their spending decisions before or at the start of the academic year; this spend might influence resit results. We will therefore be assessing the relative effectiveness of the additional funding, based on the total money spent. The categorical bands will be determined based on quartiles of the total money spent in the final distribution of spending. To account for multiple hypothesis testing, the p-values for the family of three coefficients (as there are four categories, with one as a reference) will be Bonferroni adjusted.

The second analysis will assess the impact of the marginal pound spent on the probability of achieving a greater than 4 pass in GCSE maths or passing Level 2 Functional Skills. This will be estimated using the *margins dydx* command in Stata.

This model will be estimated for the pooled treatment group for the 2018/19 cohort. As spending data is collected through two waves of school surveys, only schools that complete these surveys can be included in the analysis.

Initially, it was planned to conduct the analysis for the 2019/20 cohort as well, using all funding spent over 2018/19 and 2019/20. However, given the disruption to schooling and outcome measurement due to Covid-19, which affects both cohorts differently, we no longer consider conducting the analysis for 2019/20 cohort to be a robust option.

3.10.3 BMP impact on GCSE re-sits (RQ5)

The impact of the intervention on GCSE resits (RQ5) can be estimated in the same way as our primary analysis, detailed in equation (7). In this case, however, the outcome of interest takes the value of 1 if individuals attempted a GCSE re-sit and 0 otherwise.

As for the primary analysis the logistic model is:

$$(17)\ \boldsymbol{logit(\pi)_{ij} = \beta_0 + \beta_1 treatment_j + \beta_2 X_{ij} + \beta_3 S_j + \mathrm{u_j} + \mathrm{e_{ij}}}$$

Where X_{ij} are individual level observed characteristics, S_j are institution level observed characteristics (both described in Section 3.4.2) and u_j institution random effect. The impact for each of the three treatment arms will be estimated separately. As for the primary outcome, a Bonferroni correction will be used to account for multiple hypothesis testing.

In this case we will be using exactly the same variables as those used in our preferred specification for the primary outcome analysis, implementing the doubly-robust method. The units involved (institutions and individuals) are the same[92], as well as the estimated propensity scores for all individuals in the sample. The ITT corresponds to the effect of the treatment on the probability of retaking GCSEs for individuals in the treatment groups. More specifically we will be reporting relative risks for this measure.

3.10.4 Treatment effects in the presence of non-compliance

Given the nature of the intervention, we do not expect non-compliance to be a concern. All eligible participating institutions will receive additional funding from the DfE as per the treatment condition that each institution has been randomly assigned to. All institutions have been given access to a useful resource page that they can use to guide the spending of the additional resources. However, there is no restriction on how participating institutions can spend the additional funding. We therefore do not anticipate any behaviour by participating institutions to be considered as non-compliant with the treatment. Consequently, we do not present approaches to address non-compliance in this study plan.

3.10.5 Missing Data

Based on past experience we anticipate missingness to be relatively low for the outcome variables. However, it is highly likely that we will have missing data in the variables used during the estimation stage. Missing data is of concern primarily if the covariates showing missing data are related to the outcome and if the pattern of missingness differs between treatment and control groups. If variables with missing values are unrelated to outcomes, then there is

[92] Assuming no missing data on this outcome.

no need to adjust for them. Such an adjustment will not reduce bias but will likely decrease precision through increasing the variance of the impact estimator, either through weighting or imputation variance adjustments.

If variables with missingness relate to outcomes, but the pattern of missingness is the same between treatment and control groups, there is no need to adjust for missingness. Consequently, to adjust for any missingness in the outcome variables we want first to establish whether variables used to predict missingness are associated with the outcome variable. Once we have identified relevant variables with missingness the second step is to assess whether the pattern of missingness differs between treatment and control groups.

To account for missing data in our models, we will make use of inverse probability weighting (IPW) techniques. For the first step we will identify the association between our suite of variables in the NPD extract and our main outcomes. Secondly, variables that are correlated with these outcomes will be entered as predictors into a logistic regression model where the dependent variable is a missing/non-missing indicator on the outcome variable. This will inform us of which predictor variables are related to missingness in the outcomes. The next step will be to enter the treatment indicator and a series of interaction terms between the treatment indicator and those variables with a significant main effect association with missingness. Finally, assuming significant associations are found for some of the interaction terms, we will run a reduced form of the model including only significant interaction terms and their associated main effects from the previous model. From this we will construct a missingness weight as the inverse of the propensity score from the reduced model. We will then re-run previous models detailed above to explore the adjustment on the magnitude of the impact estimator. Stata's complex survey suite of models will be used (*svyset*) in order to adjust appropriately for the weight on the variance estimator of the impact effect.

3.10.6 Effect size calculation

We will estimate the impact of the intervention versus control group on the probability of the outcome taking the value 1 (individual achieving Level 2 maths). As the outcome is binary, the effect size will be estimated as a risk-ratio, with 95% confidence intervals by taking the ratio of predicted marginal probability of passing the GCSE for treatment and control while holding all other covariates constant at their mean. The risk ratio will be estimated using the *mlogit, rrr* command in Stata.

3.11 Statistical analysis programme

Analyses will be conducted using the Stata analysis package version 16.1 SE.

4. Implementation and process evaluation

4.1. IPE dimensions

The nature of the intervention (a fund not earmarked for any specific activity), as well as the literature reviewed in section 4, suggest that the IPE should focus on the following dimensions:

- Differentiation, i.e., the determination of:
    - What changed after the introduction of the BMP in pilot institutions;
    - What distinguishes activities led by pilot institutions from activities led by control institutions after the introduction of the BMP;
    - What distinguishes institution-led activities within each treatment arm; and
    - What distinguishes institution-led activities across treatment arms.
- Dosage, i.e., the proportion of the additional funds each institution invested in support activities.
- Responsiveness, i.e., the degree to which institutions engaged with the intervention. Particular attention will be paid to the following issues:
    - Clarity of the intervention;
    - Perception of risks and rewards;
    - Response to the financial incentive provided;
    - Usefulness of the guidance provided;
    - Perceived duration of the gap between intervention, evidence of outcome and reward and implications.
- Evidence of promise, i.e., the extent to which the intervention delivered ‘soft’ outcomes at the pilot stage, including:
    - Students’ interest, engagement and self-confidence;
    - The development of innovative and cost-effective teaching support activities;
    - Unintended effects, including ‘creaming’.

4.2. IPE questions

The research questions to be answered in the IPE are listed below, as well as the corresponding Sources of Evidence (SE).

IPE dimension	RQ#	Research question	Source of Evidence
Differentiation	RQ1	What activities and tools are currently available in pilot and control institutions to support GCSE maths re-sitters?	SE2
	RQ2	Of these activities and tools, which were not available before 2018-2019?	SE2
	RQ3	Which activities and tools were phased out after 2018-2019?	SE2
	RQ4	To what extent were the amount of support and the type of activities tailored to the needs of individual students?	SE2
	RQ5	To what extent did pilot and control institutions use the guidance on the DfE website?	SE2
Dosage	RQ6	How many contact hours on average did eligible students receive?	SE3(3)

	RQ7	What was the average cost of the intervention to institutions?	SE3(3)
Responsiveness	RQ8	How clear were the objective of the trial and the funding arrangement?	SE3(1)*
	RQ9	To what extent is the intervention aligned with the objectives and priorities of each school/college?	SE3(1)
	RQ10	Does the maximum amount (£500/student) seem sufficient to address the problem?	SE3(1)
	RQ11	What funding schemes, other than the intervention, have been used to support GCSE maths re-sitters?	SE3(3)
	RQ12	How well was the information cascaded to teachers / Heads of Maths?	SE2
	RQ13	To what extent did the intervention help teachers focus on outcomes? How pressured to succeed did they feel?	SE2
	RQ14	How concerned were institutions that the funding received from DfE would not cover their expenses?	SE3(2)
	RQ15	Did implementers understand how ‘success’ was measured? Did they find it fair?	SE3(2)
	RQ16	Was the time lag between the intervention, the outcome and the payment acceptable?	SE3(3)
Evidence of promise	RQ17	What are the perceived effects on students’ soft outcomes?	SE2
	RQ18	What are the unintended effects of the intervention (e.g., evidence of creaming; resources being diverted from other subjects)?	SE2
	RQ19	To what extent did the intervention push teachers to innovate	SE2

(*) SE3(1) refers to the first wave of the post-16 institutions survey (see section below for a description). SE3(2) refers to the second wave, etc.

4.3. Sources of evidence

The IPE will be based the following sources of evidence (SE):

SE#	Method	Description
SE1	Developer interviews	Face-to-face interviews with DfE staff will be conducted to develop a Theory of Change (ToC). This ToC will be instrumental to: • Clarify the short-term, medium-term and long-term effects of each funding model; • Understand the pathways to impact and their key assumptions; and • Identify key research questions for the IPE.
SE2	Head of Maths interviews	Telephone interviews with Heads of Maths will be conducted to explore decision-making processes and implementation in greater depth. Topics will include:

		• Planning and decision-making: reasons for interest in the pilot, how decisions were made to spend the fund, who was involved, sources consulted; • Range of support for re-sitters; • Sources of funding to enable support; • Impact of the additional funds on standard practice; • Perceived effectiveness. The first round of interviews will be carried out in Summer 2019. Follow-up post-intervention interviews will be conducted in Autumn 2020.
SE3	Post-16 Institution surveys	We propose to invite all the pilot institutions to take part in short surveys at three time-points to monitor the planning and implementation of the funding. • Survey SE3(1) will take place in Spring 2019. It will ask about the clarity of the aims of the pilot and of the conditions for participation; and how institutions intend to support students. • Survey SE3(2) will take place in September/October of 2019. It will explore institutions' responsiveness to the intervention. • Survey SE3(3) will take place in September/October 2020. It will ask institutions how funds were used and how much was actually spent to support students. All the surveys will be online with the capability for completion by more than one member of staff. The invitation and reminder emails will be sent to named leads identified at the recruitment stage through the MOU signing process.

4.4. Sampling plan

Unit of analysis	Number	Sampled	Rationale
Post-16 Institution leads	430	430	All participating institutions will be surveyed, regardless of the funding model.
Head of Maths	unknown	36	We anticipate that 8 interviews in each intervention group should be sufficient to capture the range of experiences, totalling 24 interviews.[93] Follow-up interviews will be conducted post-intervention with 4 of these institutions in each group. We propose to include institutions from across the treatment arms, selected to achieve range across the following criteria: • Treatment status (T1, T2, T3) • Number of students eligible for BMP • Level of BMP spending (based on first survey) • Activities planned (based on first survey)

[93] Eight interviews were originally planned in comparison group settings as well; however, this activity was cancelled as it would have provided little additional insight relative to the research burden involved for the institutions. Instead, we assessed business as usual retrospectively by investigating what intervention groups did prior to the introduction of Basic Maths Premium.

			• Institution type (e.g., sixth form college, FE college, academy, maintained school)

5. Cost evaluation

Cost evaluation will be conducted to compare the average cost of the intervention per funding model, bearing in mind that (i) participating institutions are free to use the funds as they see fit; and (ii) models T2 and T3 entail a certain amount of financial risk.

Costs will be evaluated in accordance with EEF guidelines. Cost data will be collected for all providers in the three treatment groups through school/college surveys. Relevant costs include:

- Direct, marginal costs;
- Relevant pre-requisites;
- School staff time.

A per-pupil, per-year cost will be estimated, including:

- Fees for services (e.g., sessions of tuition, Continuing Professional Development (CPD), ongoing support and monitoring provided by programme deliverer).
- Purchasing (or printing/photocopying) resources, materials and equipment. This might include textbooks, handouts, or digital technology.
- Travel/subsistence.

The average cost per year of the intervention will be repeated over three years to highlight the proportions of start-up vs. running costs.

6. Ethics

6.1. Informing students

Students will be informed that their institution is taking part in the intervention, and about the intended data linkage with information held in the National Pupil Database (NPD). As the legal basis for processing data is ‘legitimate interests’, consent will not be sought from students to participate. Students will however be given the opportunity to return a withdrawal form to indicate their wish to withdraw their data from data processing as part of the evaluation. Participants may request to have their data deleted at any point prior to the submission of a draft report to the EEF.

6.2. Ethical approval

NatCen’s Research Ethics Committee (REC) reviewed and approved the research proposal for this project on 11th May 2018. The committee consists primarily of senior NatCen staff. The guidance and recommendations provided by the REC have been incorporated in this Analysis Plan.

7. Data protection

7.1. GDPR

The Department for Education (DfE) is the **data controller** and the National Centre for Social Research (NatCen) and the ESFA (an executive agency of the DfE) are the **data processors** for this project. This means that the DfE is responsible for deciding the purpose and lawful basis for processing data. From May 2018, the legal basis for processing data is "legitimate interests". NatCen will process the data for the legitimate purpose of conducting the evaluation of the Basic Maths Premium Pilot. Special categories of personal data (required for matching analysis) will be processed using the research exemption. This means that data processing is necessary for research purposes, with appropriate safeguards for protecting the rights and freedoms of the data subjects. A privacy notice for the study has been published on the NatCen website, and this was issued to all participating institutions in June 2018. It can be found here: http://natcen.ac.uk/taking-part/studies-in-field/maths-premium-pilot/our-pledge-to-you/

All data collected for this study will be kept securely. We will safeguard the anonymity of all participants and no institution, staff member or student will be named in any outputs of reports. Students will be informed that their institution is receiving the intervention and will be given the opportunity to return a withdrawal form to indicate their wish to withdraw their data from data processing. Participants may request to have their data deleted at any point prior to the submission of a draft report to the EEF.

At the end of the research, all participant data will be anonymised before being archived. Once the data is archived, EEF will take on the responsibility of data controller. All personal information, and any other data held on the project, will be securely deleted once the project is complete in July 2022.

7.2. Secure data handling

NatCen has a range of policies and practices in place to ensure secure data handling. These are summarised below.

We categorise all data and files to 5 different levels, dictating how they are stored, handled and transmitted. The sample data for this study is Level 3 - 'Respondent Confidential'. Only those who carry out research tasks and those who need to check or process the data will have access to names and addresses. Our confidentiality measures for Level 3 data include:

- Encryption: All staff and freelancer laptops that hold Level 3 respondent confidential data have a hard drive encrypted using PGP Whole Disk Encryption by Symantec. This means that should the laptop be lost or stolen, the data contained on the hard drive is inaccessible. The encryption used by PGP is certified to FIPS 140-2 standards. We also use encrypted digital recorders for qualitative interviews.
- Password Policy for office-based staff:
  - Complex passwords, change every 30 days
  - 10 password history automatically enforced
  - Account locked out after 5 wrong attempts
- Access control:
  - Access to project data is managed via compliant segregation
  - Strict access control policy, limited to named authorised individuals
  - Unique serial numbers assigned to avoid use of personal information.
- Data Security Plans:

    - Project data security plan detailing data security procedures.
    - Rights of access recorded before granted.
- File Systems Auditing:
    - File System Auditor used to monitor activities logging what was created, updated, moved, renamed and deleted and when.

8. Personnel

Project team at DfE
The project team at DfE consists of Andrew Otty and Glenn Goodman

Evaluation team at NatCen
The project is managed in the Children and Families team at NatCen. The trial is co-led by Berenice Scandone (Research Director) and Martina Vojtkova (Head of Evaluation) with support from:

- Julia Griggs (Research Director), Emily Roberts (Researcher) and Tom Bristow (Researcher) for the IPE.
- Andi Fugard (Research Director and Chief Statistician for the trial), Kostas Papaioannou (Senior Researcher), Molly Scott (Senior Researcher) and Ben Stocker (Researcher) for the impact evaluation.

The Basic Maths Premium Pilot evaluation team would like to thank Arnaud Vaganay (formerly Research Director), Rodrigo Torres (formerly Research Director and Chief Statistician for the trial), Josep Espasa (formerly Senior Researcher), Anysia Nguyen (formerly Researcher) for the impact analysis, Phoebe Averill (formerly Researcher on the IPE), and Robert Wishart (formerly Research Director and Chief Statistician for the trial).

9. Risks

Non-participation in post-intervention surveys and interviews. Institutions may be unwilling to participate in the post-intervention evaluation activities. This risk may be particularly pertinent for the comparison group, who will receive no funding, and Treatment Model 1, who will have received all of their allocated funding by this point and may have little impetus to participate in the research tasks. This risk will be addressed by setting out clearly the requirements for the trial in the MoU and providing institutions with clear instructions at the start of the project on what needs to be done and when.

Sample attrition. There is a risk of sample attrition from both intervention institutions and comparison group institutions. Students will be informed that their institution is taking part in the pilot and will have the right to withdraw from their data being processed for the evaluation. This may lead to high levels of sample attrition, as a result of students opting to withdraw from the evaluation, or due to dropping-out from their education programmes. The evaluation team assumes that only a small number of students will return these forms based upon previous EEF trials, however, this will be monitored closely.

Access to NPD data. The new GDPR compliant procedures and processes that are being implemented by NPD and the DfE are likely to slow down the process of obtaining an NPD data extract for carrying out the impact analysis. Exact timelines may be impacted by the time taken for DfE to roll out virtual labs, in place of physical labs within the Office of National

Statistics Secure Research Service (SRS). To mitigate the impact of this risk, NatCen will submit an application to NPD as early as possible, once the eligible sample has been confirmed.

10. Timeline

Date	Activity	Staff responsible/ leading
May 2018	Finalise setting and student eligibility criteria, outcome measures, data collection procedures, ethical approval and key evaluation design aspects	NatCen
May 2018	Stratified cluster-level randomisation	NatCen
May – Jul 2018	Recruit post-16 settings, sign MOUs, inform about random allocation	NatCen
Jun 2018 – Sept 2019	Develop theory of change, protocol and SAP	NatCen, DfE
Nov 2018-Jan 2019	Student enumeration and student and setting data collection, informing students about data processing,	NatCen, DfE
Mar 2019	First online survey of settings	NatCen
Jan 2019 – Jun 2019	Intervention – additional funding for post-16 settings disbursed in pilot areas	DfE
Oct 2019	Baseline data collection from NPD and school and college performance tables	NatCen, DfE
Jun-Oct 2019	Case study interviews in 32 settings	NatCen
Sept-Oct 2019	Second online survey of treatment settings	NatCen
Feb 2021	Revision of SAP and NPD data request (outcomes data)	NatCen
March 2021	Access baseline data for matching	NatCen
End-Dec 2021	Expected receipt of outcomes data for impact analysis	DfE
Jan 2022-March 2022	Impact analysis and Draft report	NatCen
July 2022	Peer review, Final report	EEF, NatCen, DfE

11. References

Abadie, Alberto and Javier Gardeazabal (2003). The economic costs of conflict. A case study of the Basque Country. The American Economic Review 93(1): 113-132.

Abadie, Alberto; Alexis Diamond; and, Jens Hainmueller (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): 493-505.

Abadie, Alberto; Alexis Diamond; and, Jens Hainmueller (2015). Comparative politics and the synthetic control method. American Journal of Political Science 59(2): 495-510.

Abadie, A., & Spiess, J. (2016). Robust post-matching inference. In Unpublished Paper). MIT and Harvard University. Retrieved from https://editorialexpress. com/cgi-bin/conference/down load. cgi.

Alferes, V.R., (2012) Methods of Randomization in Experimental Design, Quantitative Applications in the Social Sciences 171, SAGE Publications.

Angrist, J. D., & Pischke, J. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton: Princeton University Press

Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in medicine*, *35*(12), 2074-2091

Austin, Peter (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46(3): 399-424,

Austin, Peter (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Statistics 10: 150-161

Austin, Peter (2014). A comparison of 12 algorithms for matching on the propensity score. Statistics in Medicine 33: 1057-1069.

Austin, Peter; P. Grootendorst and G. Anderson (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. Statistics in Medicine 26(4):734-753.

Baird, S., Ferreira F., Özler B. & Woolcock M. (2014) Conditional, unconditional and everything in between: a systematic review of the effects of cash transfer programmes on schooling outcomes, Journal of Development Effectiveness, 6:1, 1-43.

Baser, Onur (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. Value in Health 9(6): 2006

Bastagli, F., Hagen-Zanker, J., Harman, L., Barca, V., Sturge, G. and Schmidt, T. (2016) 'Cash transfers: what does the evidence say? A rigorous review of programme impact and the role of design and implementation features'.

Caliendo, Marco and Sabine Kopeinig (2005). Some practical guidance for the implementation of propensity score matching. DIW Discussion Papers No. 485

Committee of Public Accounts (2015) Funding for disadvantaged pupils: Third Report of Session 2015–16. House of Commons. Available from: https://publications.parliament.uk/pa/cm201516/cmselect/cmpubacc/327/327.pdf

Cooper, H., Hedges, L. and Valentine, J. (2007) The handbook of research synthesis and meta-analysis.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187-199.

Demidenko, E. (2007) Sample Size Determination for Logistic Regression Revisited. Statistics in Medicine, 26, pp. 3385-3397, available at: http://dx.doi.org/10.1002/sim.2771

Department for Business, Innovation and Skills (BIS). (2014). Further education workforce strategy: the government's strategy to support workforce excellence in further education. London: BIS

Department for Communities Local Government (2015) Qualitative evaluation of the London homelessness social impact bond: second interim report. London: DCLG

Department for Education (2017): School funding and pupil outcomes: a literature review and regression analysis. London: DfE

Department for Education (2020) GCSE and A level students to receive centre assessment grades, https://www.gov.uk/government/news/gcse-and-a-level-students-to-receive-centre-assessment-grades [Accessed 23/09/2020]

Dominguez-Reig, G. & Robinson, D. (2019) Education Policy Institute: 16-19 education funding Trends and Implications. Available from https://epi.org.uk/wp-content/uploads/2019/05/16-19-Funding_EPI-_2019.pdf

Dong, N. and Maynard, R. A. (2013) PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. Journal of Research on Educational Effectiveness, 6(1), pp. 24-67, doi: 10.1080/19345747.2012.673143.

EEF (2019) The attainment gap 2017; available at: https://educationendowmentfoundation.org.uk/public/files/Annual_Reports/EEF_Attainment_Gap_Report_2018.pdf)

Fox, C., & Albertson, K. (2012). Is payment by results the most efficient way to address the challenges faced by the criminal justice sector?. Probation Journal, 59(4), 355-373

Funk, M.J, Westreich, D. , Wiesen C., Stürmer T., Brookhart, A., Davidian, M. (2011) Doubly Robust Estimation of Causal Effects, *American Journal of Epidemiology*, 173 (7), 761–767.

Hong, G. & Raudenbush, S.W. (2006) Evaluating Kindergarten Retention Policy. *Journal of the American Statistical Association*, 101 (475), 901-910.

House of Commons Committee of Public Accounts, Funding for disadvantaged pupils, 2015. https://publications.parliament.uk/pa/cm201516/cmselect/cmpubacc/327/327.pdf

Jenkins, A., Levačić, R., Vignoles, A., & Levacic, R. (2006). Estimating the relationship between school resources and pupil attainment at GCSE. Department for Education & Skills/Institute of Education.

King, Gary; C. Lucas; and, R. Nielsen (2017). The Balance-Sample Size Frontier in Matching Methods for Causal Inference. American Journal of Political Science 61(2):473-489.

Lechner M. (2002) Some practical issues in the evaluation of heterogenous labour market programmes by matching methods, *Journal of the Royal Statistical Society, A*, 165, 59-82.

Lee, David and Thomas Lemieux (2010). Regression Discontinuity Designs in Economics. Journal of Economic Literature 48: 281-355.

Li, Fan, Alan M. Zaslavsky, and Mary Beth Landrum. "Propensity score weighting with multilevel data." *Statistics in medicine* 32.19 (2013): 3373-3387.

Linde, Ariel and Paul Yarnold (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. Journal of Evaluation in Clinical Practice 22(6): 868-874.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, *23*(19), 2937-2960.

Lunt, M. (2013). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American journal of epidemiology*, *179*(2), 226-235.

McClelland, Robert and Sarah Gault (2017). The synthetic control method as a tool to understand state policy. The Urban Institute research report.

McConnell & Vera-Hernandez (2015) Going beyond simple sample size calculations: a practitioner's guide https://www.ifs.org.uk/publications/7844

McCrary, Justin (2008). Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics 142(2): 698-714.

NAO (2015). Outcome-based payment schemes: government's use of payment by results. HC 86, 2015-16.

Nicoletti, C. & Rabe, B. (2012). The effect of school resources on test scores in England. Discussion Papers 12/19, Department of Economics, University of York.

Nicoletti C., Rabe, B. (2013), 'Productivity of school expenditure: Differences across pupils from diverse backgrounds', Institute for Economic and Social Research: http://www.iwaee.org/papers%20sito%202013/Rabe_c.pdf

Ofqual (2017). Provisional summer 2017 exam entries: GCSEs, AS and A levels. Coventry: Office of Qualifications and Examinations Regulation

Ofqual (2020). Results tables for GCSE, AS and A level results in England, 2020. Available from https://www.gov.uk/government/publications/results-tables-for-gcse-as-and-a-level-results-in-england-2020

Ofsted (2014) The pupil premium: an update. Available from: https://www.gov.uk/government/publications/the-pupil-premium-an-update

Porter, N. (2015). Crossing the line. Improving success rates among students retaking English and Mathematics GCSEs. A Policy Exchange Policy Bite. London: Policy Exchange.

Pugh, G., Mangan, J., & Gray, J. (2011). Do increased resources increase educational attainment during a period of rising expenditure? Evidence from English secondary schools using a dynamic panel analysis. British Educational Research Journal, 37(1), 163-189

Rodeiro, C. V. (2018). Which students benefit from retaking Mathematics and English GCSEs post-16?. Research Papers in Education, 20(3), 245-270.

Rubin, D., & Thomas, N. (2000). Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association, 95*(450), 573-585.

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

Sianesi, B. (2004) An evaluation of the Swedish system of active labour market programmes in the 1990s. Review of Economics and Statistics 86(1): 133–155.

Stuart, Elizabeth (2010). Matching methods for causal inference: A review and a look forward. Stat Sci 25(1):1-21.

Schochet (2013) Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference. http://journals.sagepub.com/doi/pdf/10.3102/1076998611432176

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514-543.

Webster, R. (2016). Payment by results: Lessons from the literature. Retrieved from: http://russellwebster.com/PbRlitreview.pdf

Work and Pensions Committee. (2011). Work Programme: providers and contracting arrangements. London: The Stationery Office

Zubizarreta J.R. & Keele L. (2017) Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System. *Journal of the American Statistical Association*, 112(518), 547-560.