Basic Maths Premium Pilot

Evaluation Report

February 2024

Molly Scott, Berenice Scandone, Julia Griggs, Emily Roberts, Tom Bristow, Eleanor Woolfe, Monica Dey, Andi Fugard

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

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

✉ info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Department for Education

What Works Network

Contents

About the evaluator

The project was independently evaluated by a team from the National Centre for Social Research (NatCen): Berenice Scandone, Julia Griggs, Emily Roberts, Tom Bristow, Eleanor Woolfe, Monica Dey, Andi Fugard, Molly Scott, Elena Cossu, Isaac Thornton, and Ben Stocker.

The lead evaluator was Berenice Scandone. The chief statistician was Andi Fugard.

Contact details

Andi Fugard
National Centre for Social Research
35 Northampton Square
London EC1V 0AX

Email: andi.fugard@natcen.ac.uk

Acknowledgements

We are grateful to the teachers and heads of maths for contributing their time and views to support this evaluation. We would like to thank Glenn Goodman and Paul Harrington at the Department for Education for their extremely helpful advice relating to preparing NPD requests, as well as Andrew Otty from the project team. We are grateful to former members of the Basic Maths Premium Pilot evaluation team who supported this project during earlier phases: Martina Vojtkova (former co-trial lead), Arnaud Vaganay (formerly research director and lead evaluator), 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), Robert Wishart (formerly research director and chief statistician for the trial), Kostas Papaioannou (senior researcher), Isaac Thornton (formerly researcher) and Ben Stocker (formerly researcher).

This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

This output request has been granted publication-level clearance (as of 21 December 2023).

Executive summary

The project

The Basic Maths Premium (BMP) is an intervention developed by the Department for Education (DfE) that aims to improve achievement in maths for 16- to 18-year-old students with a prior attainment of grade 3 or below in GCSE maths. This pilot aims to test the effects of providing additional financial support to post-16 settings to boost GCSE attainment and the relative effectiveness of different modes of provision.

The BMP was offered to all post-16 educational settings in Category 5 or 6 Achieving Excellence Areas (AEA). All eligible 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 could receive the additional funding. This pilot used a three-armed cluster-randomised trial with a quasi-experimentally defined comparison group to assess the impact of each funding model on the maths achievement of eligible students. The different funding models being tested were:

- Model A: guaranteed payment of £500 per eligible student enrolled by November 2018;
- Model B: guaranteed payment of £250 per eligible student enrolled by November 2018 and a further £250 in the 2020/2021 academic year for every student who achieved the required pass in maths by Summer 2020; and
- Model C: £500 in the academic year 2020/2021 for every student enrolled by November 2018 who went on to achieve in maths by summer 2020.

A complementary implementation and process evaluation (IPE) explored institutions' compliance with the intervention, fidelity to the intended model including levels of spending, usual practice to support GCSE maths resit students, the resources and activities funded through the BMP, settings' responsiveness and perceived outcomes, and costs of taking part.

The BMP was provided by the Education and Skills Funding Agency (EFSA), an executive agency of DfE. This pilot ran from autumn 2018 for two years but was disrupted by the Covid-19 pandemic. Due to concerns about the incomparability of exam results awarded in summer 2020, the evaluation focused on assessing the impact of BMP up to the November 2019 resit period.

Table 1 Key conclusions

Summary of key conclusions	
1.	Students in post-16 settings in receipt of BMP made the equivalent of 0 additional months' progress in level 2 maths resit attainment, on average, compared to a defined comparison group. This result has a low security rating.
2.	There was no evidence of a differential impact of BMP on level 2 maths resit attainment according to students' previous attainment, free school meals eligibility, institution size, or the amount of BMP funding spent by participating institutions. These results may have lower security than the overall findings.
3.	There was no evidence that BMP had an impact on the likelihood that eligible students sit a GCSE exam rather than a Functional Skills Level 2 exam or no maths resit at all.
4.	Perceived benefits for students included improved understanding of maths, confidence and motivation, increased attendance in lessons and attainment through additional teaching and contact hours, smaller classes, employing more experienced staff, and motivating staff through opportunities for development and progression. Perceived benefits for teaching staff included lower workload and pressure, improved opportunities for development and progression, and increased confidence.
5.	Models A and B were affected by severe implementation issues as 'guaranteed' payments were delivered late; ultimately not disbursed until March 2019. Settings in Model C reported that payment-by-results made it harder to finance support upfront and uncertainty about how much funding settings would receive made it difficult to allocate and spend resources.

EEF security rating

These findings have a low security rating. This was a well-powered, quasi-experimental comparison design with some observable selection issues. Implementation challenges, including Covid-19 disruption and delayed BMP payments, as well as pronounced area-level differences between BMP and non-BMP settings led to concerns about dilution of impact of the results. Forty-three percent of settings chose not to take up the BMP after randomisation, and a further 5% could not be included in the analysis for other reasons, but the peer reviewers did not consider this attrition to threaten internal validity as matching was conducted on the final enrolled sample.

Additional findings

Overall, this trial did not find evidence to support the premise that additional funding would translate into an improvement in level 2 maths attainment among students resitting these exams in post-16 education. Nor was any evidence found that BMP had an impact on the likelihood that eligible students sit a GCSE exam rather than a Functional Skills Level 2 exam or no maths resit. Although there was some variation by funding model, as with any study there is uncertainty around the results: the possible impact in each group is compatible with both negative and positive changes in likelihood of passing level 2 resits. The IPE provides mixed evidence of the success of the intervention.

Delays in disbursing the guaranteed portion of BMP payments raised concerns about the possibility of resources being used to support students entering post-16 education in 2019/2020 rather than the intended cohort entering in 2018/2019. As a result, the evaluation design was revised to additionally examine student outcomes in the 2019/2020 cohort, but the Covid-19 pandemic ultimately precluded this.

Survey findings showed that BMP funding was not used in full by all settings throughout the study period. Survey findings also indicated that funding was not exclusively used to support the expected target group of students: there were cases where the funding was spent to help increase overall achievement in maths across the board. Additionally, settings reported receiving funding other than the BMP to support resitters in the time period between 2018 and 2020.

In Models B and C, it was noted that the dependency of funding on student achievement for the payment by results component made it difficult to plan spending effectively and these models were perceived as unfair due to the associated financial risk. Furthermore, findings from the survey point to some evidence of ‘creaming and parking’, whereby funding support was prioritised towards students considered to be very close to getting a pass at the expense of other learners who were considered less likely to pass.

The quasi-experimental methodology was not successful at identifying a strong comparison group and, since BMP was targeted towards areas with higher need, was unable to overcome differences between BMP settings in more disadvantaged areas and non-BMP settings in less disadvantaged areas. The evaluation included additional analysis to directly compare the effectiveness of the randomly allocated BMP funding models (with no non-BMP comparison group). This analysis did not find evidence that any of the BMP funding models was more effective than another.

The results from this pilot do not confirm a link between education spending and attainment that has been demonstrated in other studies, such as Nicoletti and Rabe (2012) and the OECD analysis of PISA results (OECD, 2012). These results are more in line with findings from a recent systematic review of studies analysing the impact of cash transfer programmes on educational outcomes (Baird, et al., 2014), which only found a small impact of less than 0.1 standard deviations on test scores.

Cost

Cost to settings

Delivering BMP cost each setting, on average, £1,011 per student cohort, per year (£864, £1198 and £947, respectively, for Models A, B and C). The cost per eligible student was £8 per year estimated across three years. These figures reflect the cost to institutions of administering the BMP funding and do not include the cost of the BMP itself.

Cost of policy

The BMP represents a cost of up to £500 per student, with payments in Models B and C contingent on resit pass rates. In the context of this intervention, average per-student costs were estimated to be £500, £280, or £60, respectively, for the three funding models. This does not include any additional costs to policymakers for administering the scheme.

Impact

Table 2: Summary of impact on primary outcome(s)

Outcome/ Group	Risk ratio (95% confidence interval)	Estimated months’ progress	EEF security rating	No of students	P Value	EEF cost rating (Setting costs)*	EEF cost rating (BMP payments)
Likelihood of passing maths level 2 resit for all BMP	1.008 (0.834, 1.214)	0	🔒 🔓 🔓 🔓 🔓	47,310	0.94	**£** £ £ £ £	**£ £ £** £ £
Model A	1.025 (0.615, 1.709)	0	🔒 🔓 🔓 🔓 🔓	31,246	0.91	**£** £ £ £ £	**£ £ £** £ £
Model B	0.833 (0.542, 1.279)	-2	🔒 🔓 🔓 🔓 🔓	30,810	0.34	**£** £ £ £ £	**£ £ £** £ £**
Model C	1.071 (0.781, 1.468)	1	🔒 🔓 🔓 🔓 🔓	31,828	0.60	**£** £ £ £ £	**£** £ £ £ £**
All BMP (FSM)	0.991 (0.842, 1.167)	0	N/A	47,310***	0.92	N/A	N/A
Model A (FSM)	1.04 (0.758, 1.423)	0	N/A	31,246	0.78	N/A	N/A
Model B (FSM)	0.96	0	N/A	30,799	0.74	N/A	N/A

	(0.744, 1.249)						
Model C (FSM)	0.98 (0.715,1.334)	0	N/A	31,788	0.86	N/A	N/A

*Costs to settings to administer BMP payments

**Estimated cost of BMP payments, calculated by multiplying the number of eligible students and the respective payment amount. Average per-student costs in Models B and C will ultimately be contingent on resit pass rates.

***Total number of students in the regression model; number of FSM-eligible students in the sample=18,108.

Introduction

Background

Policy background

Since September 2013, all 16- to 19-year-old students have been required to follow a study programme that 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 or English if the student does not already hold a GCSE grade A*–C (grade 9–4 in the reformed system) in the subject. 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, that is, are not enrolled on the relevant maths and English course, will lose funding the following year. There are cases in which students may be exempt.[1]

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 9 to 4.[2] 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 2018/2019 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 maths in 2018.[3] The requirement for some post-16 students to continue studying English or maths 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 required GCSE grade (Rodeiro, 2018).

Institutions have received no additional funding to support these students. According to the Education Policy Institute (Dominguez-Reig and Robinson, 2019), the funding per student for 16–19 education has actually fallen by 16% in real terms between 2010/2011 and 2018/2019. 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.

Theoretical background

Expected impact of additional funding on attainment

In 2017, DfE conducted research to understand the impact of changes in school funding on students' outcomes in England (DfE, 2017). Its findings suggest that additional school resources positively influence attainment, although the effects are modest at all key stages.

The first part of the report, a literature review, references some studies that are directly relevant to the evaluation of the Basic Maths Premium (BMP).

- One of these studies (Nicoletti and Rabe, 2012) uses a rich administrative dataset covering approximately 93% of students in England to quantify 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 permanent £1,000 increase in expenditure per student raises achievement by about 0.02 standard deviations, which translated into 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,

[1] There are two exemptions to the condition of funding: (1) students with special educational needs and/or disabilities (SEND) who are assessed as not able to study either GCSE or stepping stone qualifications and (2) students with overseas qualifications that are established as equivalent to GCSE grade 4 or grade C.
[2] Functional Skills Level 2 maths is an alternative examination that is equivalent to a maths GCSE grade 4 (pass grade).
[3] Ofqual, GCSE outcomes in England: https://analytics.ofqual.gov.uk/apps/2018/GCSE/Outcomes/

2012).[4] According to the OECD, levels of spending explain around a fifth of the variation in overall PISA results—a sizeable amount. However, the impact is much less pronounced for high-income countries.[5]

The second part of the DfE report looks at the relationship between funding and attainment at Key Stage 4 (KS4) in English maintained schools between 2010 and 2015. The authors find no statistically significant effect of funding on attainment at KS4.

Other existing research suggests that investment of additional per-student expenditure shows larger effects in disadvantaged institutions (Pugh et al., 2011). In another U.K. review, additional resourcing improved GCSE attainment for all students and this association was strongest for students with lowest attainment in KS2 (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 were qualitative and lacked a comparison group. The Committee of Public Accounts noted evidence that the attainment gap between disadvantaged students 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 schools rated as 'good' and 'outstanding' showed higher levels of commitment to close the attainment gap by using targeted interventions and robust tracking systems (Ofsted, 2014).

A large number of studies have analysed the impact of cash transfer programmes on education outcomes. A systematic review of these found that such 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 (0.04 standard deviations for unconditional cash transfer programmes, and 0.08 standard deviations for conditional ones). The authors of the review conclude that, in the absence of complementing interventions, cash transfers are unlikely to improve student attainment substantively. In contrast, 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 that 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 awarded to service deliverers contingent upon achieving specific outcomes (Webster, 2016). For example, in an education context, a PBR instrument may award funding contingent on achieving certain grades or a pass mark in exams. 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 while reducing risk of expenditure on interventions that prove to be ineffective (Fox and Albertson, 2012). As such, evidence indicates that a conditional approach to funding can encourage innovation and efficiency within service delivery as providers are incentivised to meet outcome targets (DCLG, 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 and 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;

[4] Andreas Schleicher, the OECD's Deputy Director for Education, writing for Reform in 2012.

[5] The evidence presented shows that after a threshold of around 35,000 USD spent on educating a student from age six until the end of 15, further expenditure is unrelated to performance.

- an appropriate counterfactual can be constructed;
- services are non-essential and underperformance or 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.

Dosage and use of additional funds

A similar study by Nicoletti and 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, summarised in the DfE report (2017), suggest that:

- spending on teachers has a positive impact on attainment for most groups of students, although effects are small: a £1,000 increase in per-student spending on teachers is associated with between a 0.5% and 2.5% increase in standardised GCSE test scores in science, English, and maths;
- increased student-teacher ratios have a small negative impact on attainment; for most students a one-student increase in the student-teacher ratio reduces standardised GCSE test score by 1%, rising to 2% for the lowest attaining 10% of students;
- spending on education support staff was found to positively affect the attainment of students eligible for free school meals (FSM), those with English as an Additional Language (EAL), and Gifted and Talented students: a £1,000 increase in spending on education support staff would have increased Gifted and Talented test scores by 11%, FSM scores by 7%, and EAL scores by over 12%; and
- spending more on learning resources, such as computers and books, in most cases positively affected attainment: spending an extra £1,000 in this area would have boosted the test scores of students with Special Educational Needs and Disabilities (SEND) by 6.2%.

Study rationale

Overall, there is evidence to suggest that additional per student expenditure may improve students' GCSE outcomes. This is the case for individual student attainment (DfE, 2017; Nicoletti and Rabe, 2012) as well as national educational performance (OECD, 2012).[6] However, there is a gap in evidence in relation to the specific effects of providing additional financial support to schools to boost attainment and the effectiveness of different modes of provision. While existing studies have pointed to positive effects, most of the research was qualitative and lacked a comparison group.

The Education Endowment Foundation (EEF) commissioned the National Centre for Social Research (NatCen) to evaluate the BMP pilot. This evaluation uses a quasi-experimental research design (see below for further details) to assess the impact of additional funding for post-16 institutions on GCSE maths resit outcomes, how the funding was used, and the underlying mechanisms leading to any detected change or no change. It also measures the relative impact of three distinct funding models: guaranteed payment, payment by results, and half guaranteed payment, half by results.

Integrated evaluation design

The BMP pilot included an impact evaluation and an implementation and process evaluation (IPE). The impact evaluation sought to test and compare the impact of each funding model on the maths achievement of GCSE maths resit students in post-16 settings with prior attainment of a grade 3 or below in GCSE maths. It was designed as a three-armed cluster-randomised trial with a quasi-experimentally defined comparison group. It was not possible to have an experimentally assigned control group for this trial since DfE had committed to providing BMP funding to all eligible institutions.[7] The IPE complemented the impact evaluation. It explored:

[6] Andreas Schleicher, the OECD's Deputy Director for Education, writing for Reform in 2012.
[7]https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/Basic-Maths-Premium-Study-Plan-October-2021.pdf?v=1635158485

- institutions' compliance with the intervention;
- fidelity of implementation to the intended model including levels of spending;
- usual practice to support GCSE maths resit students in treatment group settings;
- the resources and activities funded through the BMP; and
- institutions' responsiveness, perceived outcomes, and costs of taking part (see Evaluation Objectives).

Intervention

Overview

The BMP is an intervention developed by DfE that provides additional funding to support 16- to 19-year-old students with prior attainment of a grade 3 or below in GCSE maths. The premium, worth up to £500 per eligible student, has been made available to post-16 educational settings in the most disadvantaged areas of England. The pilot ran from autumn 2018 for two years.

Why

Since 2014, 16-year-olds who do not achieve 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 receive no additional funding for these students 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.[8]

Who

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

The intervention was offered to all post-16 educational institutions in Category 5 or 6 Achieving Excellence Areas (AEA).[9] All institutions that agreed to be part of the pilot by signing a memorandum of understanding (MoU) and had at least one eligible student from cohort 2018/2019 enrolled by November 2018 received the intervention and were included in the evaluation.

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 were eligible for the additional funding attached to this pilot. Students who were exempt from the English and maths condition of funding were not eligible.

The intervention was expected to benefit:

- directly—
    - post-16 institutions in Categories 5 and 6 AEAs that signed up for the pilot and had at least one student eligible to receive BMP funding by November 2018; and
- indirectly—
- students without prior attainment of at least grade 4 in GCSE maths that enrolled in their post-16 studies in the participating institutions in the 2018/2019 academic year (the 'eligible students' for this pilot).

We refer to students who enrolled in their post-16 studies in the 2018/2019 academic year as 'Cohort 2018/2019, and to students who enrolled in the 2019/2020 academic year as 'Cohort 2019/2020'. These two cohorts are discussed further below.

[8] 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

[9] 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

What

Depending on the funding model, participating post-16 educational institutions received 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 was as follows:

- for students with prior attainment of a grade 3 in GCSE maths, achievement was 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 was defined as attainment of a grade 4 or above in GCSE maths, or Functional Skills Level 2 in maths.

The BMP Pilot tested three alternative funding models:

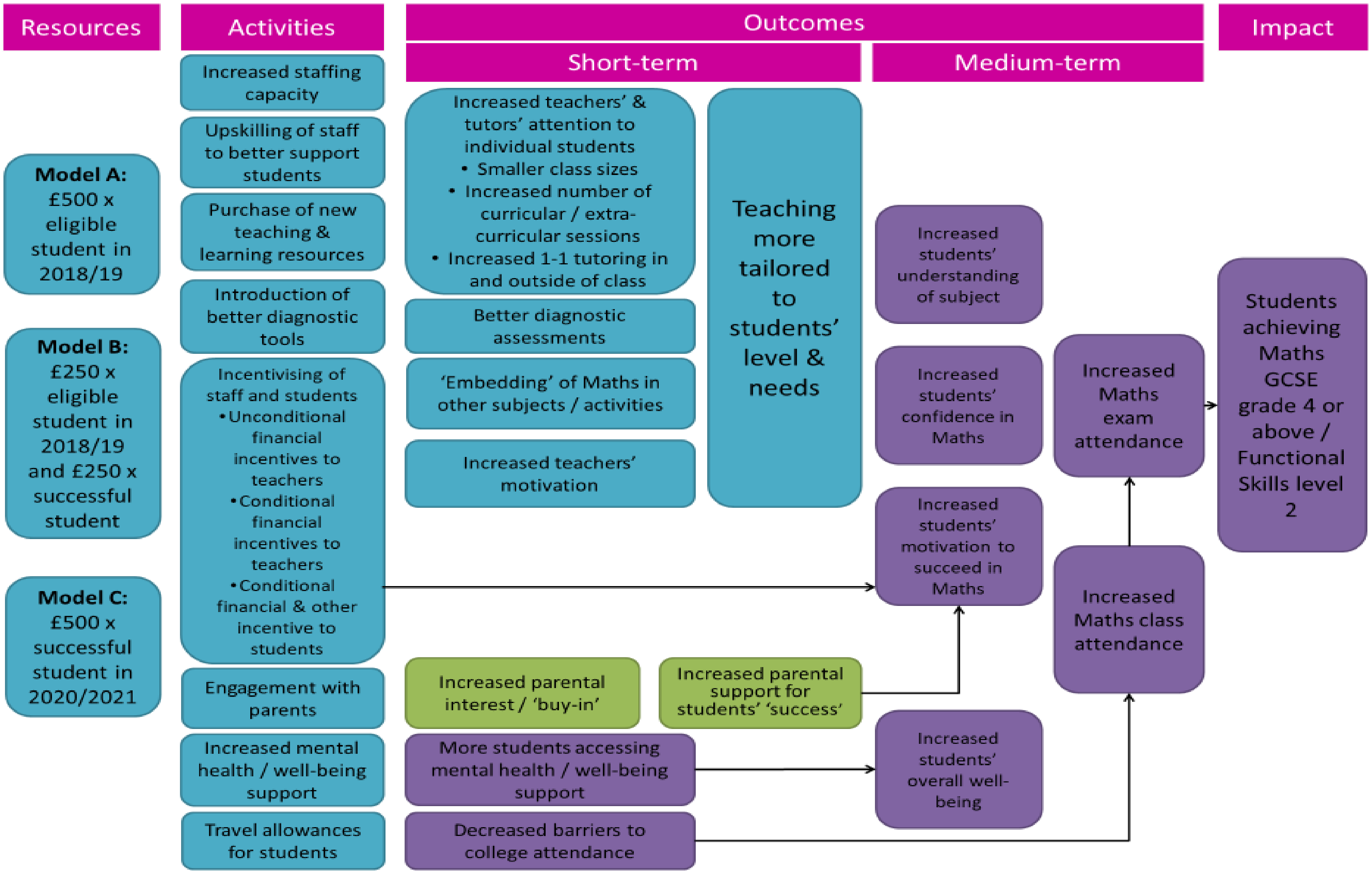
- Model A—consisted of a guaranteed payment of £500 for every student enrolled by November 2018 without a grade 4 or above in GCSE maths (only for cohort 2018/2019 funding-eligible students); payments were due to be paid at the start of the 2018/2019 academic year;
- Model B—consisted of a guaranteed payment of £250 for every student enrolled by November 2018 without a grade 4 or above in GCSE maths and a further £250 in 2020/2021 for every student who achieved the required pass in maths by summer 2020 (only for cohort 2018/2019 funding-eligible students); and
- Model C—consisted of a payment of £500 in 2020/2021 for every student enrolled by November 2018 who went on to achieve in maths by summer 2020 (only for cohort 2018/2019 funding-eligible students).

The DfE informed participating settings that the additional funding could be used for approaches or programmes known to be effective for improving maths outcomes for students—such as 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 had flexibility over how they chose to allocate the funds. Regardless of how resources were assigned, institutions were free to invest their resources in cohort 2018/2019 and/or Cohort 2019/2020 students as they saw fit (and did not have to formally document what the funding was spent on). This implies that eligible students (cohort 2018/2019) would not necessarily be the only beneficiaries of the intervention.

Intervention logic model

The BMP logic model (Figure 1) was designed in collaboration with DfE staff at the start of the pilot evaluation in June 2018. The logic model outlines the funding provided to institutions in each funding model and the activities expected to be implemented by institutions across the three models. It sets out the intended long term impact of the programme and the short term and medium term outcomes for teachers, students, and parents that are expected to lead to these impacts. No modifications were made to the logic model during the trial.

Figure 1 BMP logic model


Resources
Activities
Outcomes
Short-term
Medium-term
Impact
Model A: £500 x eligible student in 2018/19
Model B: £250 x eligible student in 2018/19 and £250 x successful student
Model C: £500 x successful student in 2020/2021
Increased staffing capacity
Upskilling of staff to better support students
Purchase of new teaching & learning resources
Introduction of better diagnostic tools
Incentivising of staff and students • Unconditional financial incentives to teachers • Conditional financial incentives to teachers • Conditional financial & other incentive to students
Engagement with parents
Increased mental health / well-being support
Travel allowances for students
Increased teachers' & tutors' attention to individual students • Smaller class sizes • Increased number of curricular / extra-curricular sessions • Increased 1-1 tutoring in and outside of class
Better diagnostic assessments
'Embedding' of Maths in other subjects / activities
Increased teachers' motivation
Teaching more tailored to students' level & needs
Increased parental interest / 'buy-in'
Increased parental support for students' 'success'
More students accessing mental health / well-being support
Decreased barriers to college attendance
Increased students' understanding of subject
Increased students' confidence in Maths
Increased students' motivation to succeed in Maths
Increased students' overall well-being
Increased Maths exam attendance
Increased Maths class attendance
Students achieving Maths GCSE grade 4 or above / Functional Skills level 2

Underlying assumptions

For all funding models, underlying assumptions as to why the intervention will achieve the intended outcomes include that:

- financial incentives will increase motivation for staff and students;
- increased motivation from staff and students has an impact on practices; and
- institutions have knowledge of, or can access, guidance on how to effectively utilise resources to improve students' success in maths.

For Models A and B, assumptions also include that:

- guaranteed funding is enough to implement changes leading to students' success in maths;
- institutions are able to accurately estimate the number of eligible students for the BMP;
- institutions are able to plan investments and activities in advance; and
- the extent of upfront investment and activities introduced will depend on the current level of institutional resources, the size of the educational provider, and an estimate of the number of eligible students.

For Model C, these include that:

- the extent of upfront investment and activities introduced will depend on the current level of institutional resources, the size of the educational provider, an estimate of the number of eligible students, and an estimate of the number of successful students.

Differences between models

Model A is expected to be more likely than Model B and much more likely than Model C to enable institutions with all levels of resources to introduce activities that require upfront investment.

Model B is expected to be more likely than Model A to result in the introduction of performance-based incentives for staff and students, in a diversion of resources from other areas to maths and in 'creaming and parking'.[10] Model B is also expected to be more likely than Model C to result in upfront investment for institutions with all levels of resources.

Model C is expected to be more likely than Model B—and especially more likely than Model A—to result in performance-based incentives, in a diversion of resources from other areas to maths and in 'creaming and parking'.

Issues that occurred during the project

Payment timing

Due to delays in disbursing the funds by DfE, guaranteed payments for institutions assigned to Models A and B were made in March 2019, seven months after the beginning of the 2018/2019 academic year. This raised concerns about the possibility of a large proportion of resources being spent for the next cohort of students aiming to resit maths in 2019/2020. A survey of participating institutions in February and March 2019 confirmed that a large majority (68%) were planning to spend part of the funding on the following cohort of students expecting to resit their maths GCSE.

Due to this, evaluation initially planned to examine the impact of the funding for students in cohort 2019/2020 as well as those in cohort 2018/2019. This plan was subsequently changed due to the impact of Covid-19 (discussed below).

Covid-19

During the Covid-19 pandemic, schools partially closed in March 2020. This resulted in the cancellation of GCSE examinations and grades being awarded to students on the basis of their teachers' assessment. Consequently, the results from cohort 2018/2019 and cohort 2019/2020 became incomparable and it was no longer appropriate to include

[10] 'Creaming and parking' is where only individuals most likely to reach the desired outcome are selected to receive the intervention. This means those students who are on the border of achieving a pass grade may receive more support, resources, and attention. This may be at the detriment of those students who are further away from achieving a pass grade.

both groups in the planned analysis. The analysis now focuses entirely on cohort 2018/2019 and cohort 2019/2020 was dropped. In addition, it was agreed with the EEF that the analysis would only measure outcomes up to November 2019 and not include the teacher-assessed grades awarded in summer 2020.

Evaluation objectives

Impact evaluation

The impact evaluation addressed the following questions:[11]

1. 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?
2. What is the effectiveness of the different funding models relative to each other?
3. How do these effects differ by student disadvantage status, prior attainment in GCSE maths, and number of eligible students per institution?
4. What is the impact of the total additional funding spent by post-16 settings in England on eligible students' maths achievement?
5. What are the effects of the different funding models on the likelihood of students attempting a GCSE maths exam retake for students with a prior achievement of grade 2 and below?

Implementation and process evaluation

Based on the nature of the intervention—a fund not earmarked for any specific activity—and the literature reviewed in the Background section, the IPE focused on the following dimensions and respective research questions:

Differentiation, that is, the determination of:

1. What activities and tools are currently available in pilot institutions to support GCSE maths re-sitters?
2. Of these activities and tools, which were not available before 2018/2019?
3. Which activities and tools were phased out after 2018/2019?
4. To what extent were the amount of support and the type of activities tailored to the needs of individual students?
5. To what extent did pilot institutions use the guidance on the DfE website?

Dosage, that is, the proportion of the additional funds each institution invested in support activities:

6. How many contact hours on average did eligible students receive?
7. What was the average cost of the intervention to institutions?

Responsiveness, that is, the degree to which institutions engaged with the intervention:

8. How clear were the objectives of the trial and the funding arrangements?
9. To what extent is the intervention aligned with the objectives and priorities of each school/college?
10. Does the maximum amount (£500 per student) seem sufficient to address the problem?
11. What funding schemes, other than the intervention, have been used to support GCSE maths re-sitters?
12. How well was the information cascaded to teachers or heads of maths?
13. To what extent did the intervention help teachers focus on outcomes? How pressured to succeed did they feel?
14. How concerned were institutions that the funding received from DfE would not cover their expenses?

[11] https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/Basic-Maths-Premium-Study-Plan-October-2021.pdf?v=1635158485

15. Did implementers understand how ‘success’ was measured? Did they find it fair?
16. Was the time lag between the intervention, the outcome, and the payment acceptable?

Evidence of promise, that is, the extent to which the intervention delivered ‘soft’ outcomes at the pilot stage:

17. What are the perceived effects on students’ soft outcomes?[12]
18. What are the unintended effects of the intervention (for example, evidence of creaming or resources being diverted from other subjects)?
19. To what extent did the intervention push teachers to innovate?

Ethics and trial registration

Ethical approval for the impact evaluation and IPE was obtained from NatCen’s Research Ethics Committee on 11 May 2018.

DfE recruited institutions to take part in the BMP pilot. Participation in the evaluation was a condition for receiving the funding. A memorandum of understanding (MoU) was signed by each institution expressing their agreement to take part in the pilot and all evaluation tasks (see Technical Appendices - Appendix A). The MoU set out the research aims, the roles and responsibilities of the three parties (EEF, DfE, and NatCen) and formalised the institution’s commitment to take part in the trial. Participants were informed of what participation in the intervention and evaluation would involve, how their data would be collected and processed, and their rights to withdraw (see Data Protection).

NatCen then invited the project leads at institutions who had consented to participate in the pilot to take part in three survey waves and two rounds of interviews (see Methods). Participants were sent an invitation email, a link to the study webpage, and the pilot privacy notice (see Technical Appendices – Appendix B). The webpage and privacy policy informed participants of the research aims and activities, the voluntary nature of participation, and confidentiality; they also explained how their data would be used and their rights to withdraw.

The trial was registered with International Standard Randomised Controlled Trial Number (ISRCTN) on the 28 October 2019. The trial number is ISRCTN59217701.[13]

Data protection

The EEF commissioned NatCen to carry out this evaluation of the DfE BMP pilot. The data collected as part of this evaluation was used for research purposes only and was processed in line with the U.K. General Data Protection Regulation (GDPR). NatCen was the data controller. The legal basis for processing personal data was ‘legitimate interest’.[14] This means NatCen has processed this data for the legitimate purpose of conducting the evaluation of the BMP. Special categories of personal data, including ethnicity, were processed using the ‘research exemption’.[15] This means data processing was necessary for a research purpose with appropriate safeguards for protecting the rights and freedoms of data subjects.

Research participation for learners involved the processing of their personal data held by the National Pupil Database (NPD). Learners were not required to take part in any other data collection activities. Attainment data from the NPD was used to inform our impact evaluation. Students were informed that their institution would receive the intervention and were given the opportunity to withdraw their data from processing. All assessment data was pseudonymised before being analysed.

All data collected for this study was stored securely. We safeguarded the anonymity of all participants and no institution, staff member, or student has been or will be named in any outputs of reports. Participants were able to 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 was anonymised and archived with the ONS Secure Research Service. At this stage, the EEF became the data

[12] ‘Soft outcomes’ in this case refers to outcome areas that are subjective or less tangible than more readily quantifiable outcomes such as exam grades. For example, this includes participants’ perceptions of the outcomes of BMP and aspects such as student confidence, motivation, engagement, and enjoyment of maths.

[13] **http://www.isrctn.com** or

[14] Point (f) of paragraph 1 of Article 6 of the GDPR: **https://gdpr-info.eu/art-6-gdpr/**

[15] Point (j) of paragraph 2 of Article 9 of the GDPR: **https://gdpr-info.eu/art-9-gdpr/**

controller. All personal information, and any other data held on this project, will be securely deleted once the project is complete in July 2022.

Project team

The BMP intervention was delivered by DfE.

Table 3: DfE delivery team

Delivery team	
Andrew Otty	Team leader
Glenn Goodman	Statistician

The evaluation was led by NatCen's Children and Families team, who worked closely with impact evaluation experts in NatCen's Centre for Evaluation.

Table 4: NatCen personnel

Children and Families Team		
Dr Berenice Scandone	Research director	Principal investigator. Senior oversight of and final report
Dr Julia Griggs	Research director	Former principal investigator and senior oversight of IPE
Dr Arnaud Vaganay	Former research director	Former Principal Investigator and senior oversight of IPE
Emily Roberts	Researcher	IPE data collection, analysis and final report
Tom Bristow	Researcher	
Phoebe Averill	Former researcher	IPE data collection and analysis
Adam Gilbert	Former research assistant	IPE data collection and analysis
Eleanor Woolfe	Researcher	IPE analysis and final report
Monica Dey	Researcher	Final report
Centre for Evaluation		
Daniel Phillips	Former director of the Centre for Evaluation	Senior oversight of trial
Martina Vojtkova	Former director of the Policy Research Centre and former head of Evaluation	Trial design
Dr Andi Fugard	Co-director of the Centre for Evaluation and former research director	Chief statistician
Molly Scott	Senior researcher	Impact evaluation analysis
Elena Cossu	Senior researcher	Impact evaluation analysis
Isaac Thornton	Former senior researcher	Impact evaluation analysis
Ben Stocker	Former researcher	Impact evaluation analysis
Dr Kostas Papaioannou	Senior researcher	Trial design

Dr Rodrigo Torres	Former research director	Former chief statistician
Robert Wishart	Former research director	Former chief statistician
Josep Espasa	Former senior researcher	Trial design
Anysia Nguyen	Former researcher	Trial design

Methods

Trial design

The design for the impact evaluation component of this evaluation combines elements of a cluster randomised controlled trial with a quasi-experimental design (QED) to estimate impact. Eligible settings were randomly allocated to one of three BMP funding models (outlined in Intervention above) in May 2018. Due to DfE's commitment to offer all institutions in pilot areas some form of intervention, it was not possible to experimentally generate a business-as-usual control group. The impact evaluation therefore uses a QED approach to develop a suitable comparison group for estimating impact. This constructed comparison group constitutes a 'business-as-usual' condition—settings that did not receive any form of intervention. In other words, the aim of the QED is to identify a comparison group that was not offered any form of BMP funding but was in all other respects as similar as possible to the BMP group.

The primary outcome for the impact evaluation is achieving a pass in maths level 2 by November 2019. As a secondary outcome, the trial also assesses the impact of the pilot on whether students resit a GCSE maths exam (as opposed to Functional Skills Level 2 or are not entered for any qualification).

We used the following sources of data for the statistical analysis:

- NPD data: this was used to measure the primary and secondary outcome and to derive covariates for the analysis. We requested two separate extracts of NPD data for the analysis: an initial extract containing data required to measure covariates only and a second containing data required to derive the final outcome and some extra covariate information.
- Individualised Learner Record (ILR) and school census data: we requested data from the ILR and school census to help define student eligibility for the trial analysis.
- BMP sample file data: this is a dataset of all participating BMP institutions. We used this dataset to identify which treatment group each setting was randomised to as well as to remove institutions from the dataset that were marked as having withdrawn from the trial.
- Survey data: we linked data collected from three waves of a survey conducted with participating BMP schools. This data was used to carry out the dosage analysis as part of the trial.
- Institution-level information drawn from the following publicly available sources: AEA data, Get Information about Schools, and the Further Education Skills Dataset. These datasets were used to derive additional setting-level covariates.

The data requested from the NPD and ILR covered all students who received a grade 3 or less in maths GCSE in 2017-18 and then enrolled in any post-16 setting in 2018/2019. We requested exam results from the KS4 data table from the 2018/2019 academic year, plus the November 2019 resit period.

There have been some changes to the trial design during the implementation of this pilot. The most significant of these were as a result of the onset of the Covid-19 pandemic. This led to the decision to shorten the time available for the impact evaluation to exclude resits taken during the summer 2020 exam period. This is because grades awarded in summer 2020 were teacher assessed due to Covid-19 and were not thought to be comparable to the grades awarded in previous exam seasons. In tandem with this change, the trial was also modified to remove an additional assessment of the effect of BMP on the cohort of students entering post-16 education in 2019/2020. Both of these changes were already discussed and agreed with the EEF prior to the finalisation of the most recent study plan.[16]

The research team also made some further small changes to the impact evaluation approach since the study plan was published. These are described below in the Statistical Analysis section.

The methodology for the impact evaluation is summarised in Table 5. The remainder of this section sets out the research methodology in more detail.

[16] See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/basic-maths-premium-pilot

Table 5: Trial design

Trial design, including number of arms		Three-armed cluster randomised trial with a quasi-experimentally defined comparison group.
Unit of randomisation		Setting.
Stratification variable (s) (if applicable)		Stratified blocked design using institution type and number of condition of funding students in 2016/2017.
Primary outcome	Variable	Achievement in maths level 2 resit.
	Measure (instrument, scale, source)	Achievement is a binary variable, taking the value 1 as follows: • 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). This is defined using NPD data.
Secondary outcome(s)	Variable(s)	Whether a student resits maths GCSE.
	Measure(s) (instrument, scale, source)	This is a binary outcome that takes the value 1 if a student with a prior attainment of grade 2 and below attempted a GCSE maths exam up to November 2019. The outcome takes the value 0 if the student only attempted Functional Skills Level 2 in maths or did not resit any maths exam. This is defined using NPD data.
Baseline for primary outcome	Variable	Prior attainment level at maths KS2.
	Measure (instrument, scale, source)	KS4_KS2MAT24P_PTQ_EE, drawn from the NPD This is defined as KS2 maths prior attainment level for the KS2-KS4 progression measure (2011 methodology). It takes values of between 0–6 according to the level of attainment achieved: level 0–level 6.
Baseline for secondary outcome(s)	Variable	Prior attainment level at maths KS2.
	Measure (instrument, scale, source)	KS4_KS2MAT24P_PTQ_EE, drawn from the NPD. This is defined as KS2 maths prior attainment level for the KS2-KS4 progression measure (2011 methodology). It takes values of between 0–6 according to the level of attainment achieved: level 0–level 6.

Participant selection

Intervention group selection

BMP funding was awarded to all post-16 settings in AEA Category 5 or 6 areas that had at least one eligible student enrolled in the 2018/2019 academic year and had signed an MOU agreeing to be part of the pilot. At the student level, students were eligible for BMP funding if they were 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. Students who were exempt from the English and maths condition of funding were not eligible, neither were students that enrolled for a post-16 education setting after November 2018.

Comparison group selection

This trial does not include an experimentally assigned group that did not receive BMP. The purpose of the quasi-experimental analysis is to construct a comparison group that is as similar as possible to the BMP group in all respects apart from exposure to the BMP funding. In the first instance, our comparison group is drawn from students in AEA Category 4. The premise underlying the analysis approach is that there should be at least some settings and students in AEA Category 4 that bear similarity to those in BMP institutions in AEA Categories 5 and 6. The study plan also provides the option to expand to include students from AEA Category 3 if needed to achieve a suitable comparison group.

Within these comparison areas, we restrict the sample to only students who *would have been* eligible for the BMP funding had it been offered. To do this, we constructed a measure of hypothetical BMP eligibility for students outside of the areas where BMP was funded. We consulted with the DfE to understand the calculations that the ESFA carried out to identify BMP eligible students in the intervention group, and see whether we could replicate these for our comparison group pool. The EFSA calculations drew on some funding files and lines of code that were not available to the evaluation team, so we worked together to establish how to develop the best possible proxy using available data from the NPD, school census, and ILR.

Table 6 reports how many students are found to meet the eligibility criteria for BMP using this proxy. Overall, we find that around 24.8% of students in Year 12 in 2018/2019 (across all of England) met the programme's individual-level eligibility criteria. As a cross check on our derived BMP eligibility variable, we checked how well this aligned with the student-level BMP eligibility variable in our intervention group. We find that 93.2% of BMP students in our data are confirmed to have met the eligibility criteria, according to our proxy variable. This is a high degree of alignment, indicating that our estimated eligibility condition matches up well with actual BMP allocation.

Outside of the AEA category areas where BMP was funded, our comparison group contains only those students estimated to have met the eligibility criteria for the BMP according to this derived variable. These students make up around 22.4% of Year 12s outside AEA Categories 5 and 6 and form the basis for the comparison group we use for our analysis.

Table 6: Estimating student-level BMP eligibility (comparison group proxy)

	Proportion	n
Overall: proportion of all Year 12s in 2018/2019 who meet the individual-level BMP eligibility criteria according to our proxy calculation.	24.8%	495,961
Intervention group: proportion of those tagged as being BMP students in AEA Categories 5 and 6 who meet the eligibility criteria according to our proxy calculation.	93.2%	44,971
Comparison group: proportion of students in AEA Categories 1–4 who meet the BMP eligibility criteria according to our proxy calculation.	22.4%	335,670

Source: NPD, School Census, ILR and BMP sample file data.

Final analysis sample

Our final analysis sample is defined as follows:

- in AEA Category 5 and 6 areas, the sample consists of all settings that received BMP funding; within those settings, the sample includes all students who entered Year 12 in 2018/2019 and did not have prior achievement of maths level 2; and
- in other AEA category areas, the sample consists of all students who entered Year 12 in 2018/2019 and did not have prior achievement of maths level 2; we removed from the sample any additional students who were estimated to be ineligible for BMP using the proxy definition described above.

Defining BMP exposure for analysis: two definitions

Two slightly different definitions of BMP exposure were available to the evaluation team.

For the main analysis we use an indicator of BMP exposure that is defined at the setting level. This comes from the BMP sample file that was maintained during the trial and includes all settings that received funding and did not withdraw from the trial. Within these settings, we treat all students who were enrolled in post-16 education after November 2018 and did not have prior attainment of a maths level 2 qualification as being part of the BMP intervention group. This 'setting-level' indicator of intervention exposure coheres with the cluster randomised design of the overall trial, since funding was indeed allocated at level of the setting rather than at student level. By defining BMP exposure in this way, the analysis presumes that all eligible students who were part of settings that received BMP could have potentially benefited from it. For this definition of BMP exposure, the comparison group consists of all students in AEA Category 4 (or 3) areas who *would have been eligible* for BMP had it been available in their area (according to our proxy eligibility definition, described above).

A second indicator of BMP eligibility was also added to our NPD extract by the DfE team. This is a student-level measure that indicates whether students were individually identified on their systems as being BMP beneficiaries.[17] We have conducted a post-hoc, exploratory impact analysis using this individual-level BMP exposure measure.

In principle there should be minimal differences between the two ways of defining the BMP treatment group in our analysis sample since this is already restricted to all students estimated to meet the eligibility criteria for the BMP (whether they were in a school receiving the BMP funding offer or not). But there are some minor discrepancies, as shown in Table 7. Table 7 compares the two BMP exposure indicators for students in our main analysis sample. That is, students in BMP settings are predominantly flagged as individual BMP beneficiaries and those in non-BMP settings are flagged as non-BMP students. As expected, there are no BMP settings in AEA Category 1–4 areas, and our analysis sample does not include any 'non-BMP' settings in AEA Category 5–6 areas. We do observe a small disparity in the two definitions: 2,536 students in BMP settings are not individually flagged as BMP beneficiaries and 321 students in AEA Category 1–4 areas are flagged as BMP beneficiaries. This implies that 6.2% of settings in comparison areas are recorded as containing any students that are individually identified as belonging to BMP (compared with 99.5% of BMP settings). The reason for discrepancies between the two variables could be due to students meeting the BMP funding eligibility criteria switching providers after the date at which the individual-level BMP exposure variable was constructed. Our exploratory analysis aimed to examine whether these discrepancies in the two treatment group definitions make a difference for final impact estimates.

Table 7: Comparing BMP eligibility variables within our sample

	AEA Category 5 and 6 areas		**AEA Category 3 and 4 areas**	
	Non-BMP student	BMP student	Non-BMP student	BMP student
Non-BMP setting	0	0	44,240	321

[17] The individual-level measure of BMP exposure was derived by the DfE using data from the R04 ILR return of November 2018 for all providers (apart from one provider that did not complete this return: for this provider the R06 return was used instead). ILR returns data was combined with information from the Autumn 2018 Census (reference date October 2018).

BMP setting	2,536	41,916	0	0
Total	2,536	41,916	44,240	321

This table compares two alternative definitions of BMP exposure: one 'setting-level' indicator that is based on information from the BMP sample file, and another 'student-level' indicator that was added to our NPD extract by the DfE team. The sample is our final analysis sample—that is, all students in Year 12 in 2018/2019 that were estimated to meet the eligibility criteria for BMP.

Outcome measures

Baseline measures

For this evaluation, the baseline period refers to the period before randomisation took place at the end of 2018. No baseline testing or data collection was conducted prior to randomisation and treatment assignment as the evaluation relies on baseline data from the NPD, ILR, and school census. Baseline data drawn from these administrative datasets was used to construct relevant covariates for the statistical analysis. This includes prior attainment at KS2 English and maths as well as a range of other pre-intervention outcomes that are correlated with outcomes. Covariates used in our analysis are discussed in Table 9 below.

Primary outcome

The primary outcome for the impact evaluation is a binary variable reflecting whether a student achieved a pass in maths level 2 during the 2018/2019 academic year or in the November 2019 resit period. The definition of achieving a suitable pass varies depending on student's prior attainment and is defined as follows.

- For students with prior attainment of a grade 3 in GCSE maths, achievement is defined as attainment of a grade 4 or above in GCSE maths. We consider the student to have failed if the highest GCSE maths grade received up to November 2019 is grades 1 to 3 or a U, or if they did not resit GCSE maths during this period.
- 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. We consider the student to have failed if the highest maths grade received up to November 2019 was a grade 1 to 3 or a U in GCSE maths, or a 'fail' in Functional Skills Level 2 in maths.

In the first draft of the study plan, we planned to measure attainment outcomes up to the summer of 2020. As discussed in the Trial Design section above, due to the onset of Covid-19 we subsequently agreed with the EEF to remove the summer 2020 exam period from our assessment and instead measure outcomes up to November 2019.

Maths attainment is measured using the KS4 exam table in the NPD. This contains a record for each exam taken by a given student, along with the exam type (from which we can identify GCSE maths and Functional Skills Level 2 maths exams), date of the exam, and outcome. We used a DfE-provided identifier (PupilMatchingRefAnonymous) to link this exam-level extract back to our main sample of eligible students (that is, students who had received a grade 3 or less in maths GCSE in 2017/2018 and enrolled in any post-16 setting in 2018/2019). Linking back to the main student data allowed us to identify eligible students in our sample who did not have an outcome for any GCSE maths or Functional Skills Level 2 maths exams recorded in the exam-level data covering the 2018/2019 academic year and the November 2019 resit period. We recorded these students as not having achieved the required pass when we defined our main outcome measure.

Secondary outcomes

To address Research Question 5, we define a secondary outcome that captures whether an eligible student re-sat GCSE maths during the evaluation time horizon. This is a binary outcome that takes the value 1 if a student with a prior attainment of grade 2 and below attempted a GCSE maths exam up to November 2019. The outcome takes the value 0 if the student attempted Functional Skills Level 2 in maths or did not resit any maths exam. This measure therefore captures whether or not students chose to resit a GCSE exam rather than a Functional Skills Level 2, if they were eligible to resit either exam.

Note that this outcome is only defined for students who received a grade 2 or below in their previous maths level 2 attempt since students who received a grade 3 are required to retake a GCSE exam unless they are exempt from the condition of funding. This outcome is also measured using the KS4 exam table from the NPD, described above.

Randomisation

The unit of randomisation for this trial is the setting; 824 eligible post-16 institutions in AEA Categories 5 and 6 were randomised into one of three possible intervention arms (A, B, or C) at the end of May 2018.

Randomisation was carried out using a stratified blocked design with two stratifying factors: institution type and number of CoF students in 2016/2017. The institution type stratifier classified institutions into three types: (1) 16–19 provider, (2) Academy, or (3) sixth-form, according to data provided by the DfE. The number of CoF students in 2016/2017 was used as a proxy for the likely number of students per institution that would meet the eligibility criteria for BMP funding in 2018/2019.

Stratified randomisation was carried out as follows:

- first, institutions were grouped by institution type;
- within each institution type, institutions were ordered by the number of condition of funding students, from largest to smallest, and divided into blocks of three based on the 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 condition of funding students in 2016/2017 was unknown were divided into blocks of three and each block assigned a randomly ordered list of the treatment assignments; then
- a last replication correction procedure, as outlined in Alferes (2012), was applied to incomplete blocks within each stratum; for blocks where the number of units is below three this entailed randomly assigning those settings to one of the three treatment arms.

Randomisation was carried out in R (Version 3.4.1, using a random number generator, `runif`, in the `base` package). The randomisation was conducted by an independent analyst within the evaluation team prior to eligible institutions being informed about the opportunity to participate in the pilot.

Following randomisation, all eligible institutions were invited to take part in the pilot and the study. Of these, 469 (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 434 remained in the trial at the time the study plan was written. This is summarised in Table 8, which shows that the remaining institutions were approximately equally distributed across the treatment groups. The sample sizes remaining in the trial at the time of analysis are summarised in the Participant Flow Including Losses and Exclusions section below.

Table 8: Randomisation breakdown by institution type

	Initially randomised				**Retained in the trial as of October 2021 study plan**				
Treatment group	16–19 provider	Academy	6th form	Total	16–19 provider	Academy	6th form	Total	Proportion of those originally randomised
A	58	167	47	272	41	79	20	140	51%
B	60	164	47	271	46	76	23	145	54%
C	64	171	46	281	44	83	22	149	53%
Total	182	502	140	824	131	238	65	434	53%

Sample size

This subsection describes how sample sizes were determined at randomisation, at the time the study plan was written, and at analysis stage.

Sample size determination at the randomisation stage

As noted in the Randomisation section above, the sample size at randomisation was 824 settings from AEA Categories 5 and 6. This sample size was determined by the number of settings that were eligible for BMP. Note that this sample size corresponds to the maximum available treatment group only since this trial did not include a randomised non-BMP group.

Sample size determination at the time of the October 2021 study plan

Following randomisation, all eligible institutions were contacted by the DfE and offered the opportunity to participate in the pilot. At this stage settings were not informed of their treatment allocation status (this was only revealed to them after deciding whether to participate). Out of those invited, 469 institutions (53%) agreed to participate; these were then assessed against the remaining eligibility criterion (having a minimum of one eligible student enrolled by November 2018) in January 2019 and 434 were retained in the trial.[18] A similar number of comparison group institutions were expected to be selected via the quasi-experimental methodology.

The sample size calculations for the study-plan stage resulted in a Minimum Detectable Effect Size (MDES) of 0.21 standard deviations for estimating the impact of each BMP treatment arm separately and an MDES of 0.10 standard deviations for the impact of the pooled BMP treatment group.[19] These MDES estimates are reported as standardised mean differences. We also carried out sample size calculations for the subgroup of students who received free school meals during the previous six years. For this subgroup, a total sample of 23,473 students was expected, yielding an MDES of 0.15.

Sample size determination at the analysis stage

For the BMP group, the analysis sample size was primarily determined by how many of the 434 settings retained in the trial were successfully linked to the NPD data extract and did not have missing values for the outcome variable and covariates.

For the comparison group, the analysis sample size was determined by how many units were selected by the quasi-experimental methodology for each estimation model carried out. The process of implementing the methodology is discussed in detail in the Impact Evaluation Results section below and involved weighting and matching approaches. For the primary analysis (based on a statistical weighting approach), all available comparison units were retained in the impact estimation. For the robustness checks based on matching approaches, only comparison group units that were ‘matched’ to the intervention group (that is, found to be sufficiently similar) were retained in the final analysis sample.

In addition, the sample retained for each analysis depended on any trimming of observations with extreme propensity scores that was applied.

For the analysis of Research Question 5, the sample size was further constrained by the number of BMP settings that completed each wave of the survey used to examine whether and how BMP funding received was spent by settings.

The final sample size in our main analysis sample is summarised in the Participant Flow Including Losses and Exclusions section below. The sample size associated with the impact analysis for each research question is reported alongside the main findings in the Impact evaluation results section below.

[18] At the time that sample size calculations were performed for the study plan there were 436 settings remaining in the trial not 434. We have now updated the ‘study plan’ sample size calculations (presented in the Participant Flow Including Losses and Exclusions section below) to include 434 schools rather than 436.

[19] The MDES here is expressed in standard deviation units (that is, it is measured as a standardised mean difference). The Sample Size Calculations section in the study plan discusses in detail how power calculations were carried out and how the MDES can be alternatively expressed as a difference in proportions. In the Participant Flow Including Losses and Exclusions section below we present sample size calculations again to show MDES sizes in the final analysis dataset. We also convert the MDES into units of a risk ratio in order to correspond with the units that the majority of our impact analysis results are reported in.

Statistical analysis

Primary analysis

The primary analysis used a doubly-robust estimation (Funk et al., 2011) approach to address Research Question 1. This method combines propensity score estimation with regression analysis to estimate causal impact. It is equivalent to an inverse-probability of treatment weighting estimator augmented by a regression model.

Step 1: Selection of covariates

Suitable covariates for the statistical analysis were identified in the preparation for the study plan. Covariates were selected if they were judged on substantive grounds to satisfy the following conditions:

- covariates are associated with the a priori likelihood of being exposed to the BMP funding;
- covariates are associated with the likelihood that a student achieves the required pass in maths level 2 resit exams; and
- covariates are not themselves able to be influenced by the receipt of BMP funding; that is, they are pre-determined or 'fixed' characteristics.

Covariates meeting these conditions are characteristics that could be a source of bias in impact estimation if they were ignored. Without adjusting for these covariates we would not be able to determine whether any observed differences in outcomes between the BMP and non-BMP group reflected a causal impact of the intervention or were instead simply a reflection of underlying differences in these confounders. Including covariates in the analysis that meet these criteria allows us to adjust for bias that would otherwise affect the impact estimation.

A set of covariates meeting these conditions were identified in the study plan. These range from individual student-level characteristics, setting-level characteristics, and area-level characteristics used to designate AEA status to different areas. These 'AEA covariates' are the list of factors incorporated in the calculation used to determine AEA status. We do not include the AEA category itself as a covariate, because there is no overlap at all between BMP and non-BMP settings in terms of AEA category. This is because BMP settings are located in AEA Category 5 and 6 areas, whereas the comparison group is drawn from AEA Category 4 only (or 3 and 4). This lack of any overlap in AEA category between the BMP and non-BMP group makes this unsuitable as a covariate. Instead, we used the composite AEA variables that feed into the definition of the AEA category as covariates, since the individual variables are not themselves directly predictive of treatment status (unlike AEA category).

The covariates are summarised in Table 9 below.

Table 9: Description of covariates used in the analysis

Covariate		Source	Description
Achieving Excellence Area covariates	Access to a good secondary school index	Publicly available DfE data	Percentage of students able to access a 'good' or 'outstanding' secondary within 5 km of their home.
	AEA achievement indicator	Publicly available DfE data	Attainment 8 Score for the local authority. The Attainment 8 score is a measure of student attainment across 8 qualifications.
	AEA progress indicator	Publicly available DfE data	Progress 8 Score for the local authority. The Progress 8 score is a measure of student progress in the same 8 qualifications as the Attainment 8 score. It captures progress made from the end of primary school until the end of KS4.
	System leader coverage indicator	Publicly available DfE data	Number of secondary students per secondary phase Teaching School or National Leaders in Education. The purpose of this indicator is to capture disparities

			in the capacity of schools to improve through school-to-school support.
	Initial teacher training provider coverage index	Publicly available DfE data	Trainees per 10,000 students.
	Quality of leadership indicator	Publicly available DfE data	Percentage of secondary students in a school with 'good' or 'outstanding' leadership.
	Academy sponsor coverage	Publicly available DfE data	Students per lead or outstanding sponsor academy.
Setting-level covariates	Institution size	Derived using ILR and school census data	Mean number students enrolled at the setting for the academic years 2016/2017, 2017/2018, and 2018/2019.
	Number of Year 12s meeting the eligibility criteria for BMP in 2018/2019	Derived from NPD, ILR, and school census data	This covariate is a proxy for the original covariate described in the study plan. The reasons for the departure from this original plan are discussed below.
	Proportion of Year 12s meeting the eligibility criteria for BMP in 2018/2019	Derived from NPD, ILR and school census data	This covariate is a proxy for the original covariate described in the study plan. The reasons for the departure from this original plan are discussed below.
	Institution type	NPD data (KS5_NFTYPE)	We generated dummy variables for the most commonly observed among our sample: academy converter, further education sector, academy—sponsor led, other institution type, and institution type missing. These are entered as separate covariates in the statistical analysis, with 'institution type missing' acting as the reference category (and omitted).
	Region	YPMAD data (YPMAD_main_inst_ region)	Geographic region of the setting. We grouped regions into three categories: the North (consisting of the North East, North West, and Yorkshire and Humber), the South (consisting of the South East, South West and London), and the East and Midlands (consisting of East Midlands, West Midlands and the East). These are entered as separate covariates in the statistical analysis, with the East and Midlands category omitted and treated as the reference category.
	Proportion of students who were eligible for FSM between age 10 and 15	Derived from YPMAD data (YPMAD_ EverFSMage10to15)	This is a proxy for the original covariate described in the study plan. The reasons for the departure from this original plan are discussed below.
	Proportion of students achieving level 2 or more in maths and English before enrolment	Derived from YPMAD data (YPMAD_L2plusFuncAc_ Eng_Maths)	The proportion of Year 12s who achieved maths and English GCSE or equivalent qualifications.
	Age	Derived from NPD data	Student's age as of the first Monday of September 2018.

Student-level covariates	Gender	NPD data	Student’s gender.
	Ethnicity	NPD data (KS5_KS4_ETHNIC)	Student ethnicity. We derived the following categories based on ONS definitions: White, Mixed/multiple, Asian, Black, Other. These are entered as separate covariates in the statistical analysis, with ‘other’ ethnicity omitted and treated as the reference category.
	Eligibility for free school meals	NPD data (YPMAD_EverFSMage10to15)	Whether the student was recorded as FSM eligible at any time between the ages of 10 and 15.
	Special Education Need status (SEN)	NPD data (KS5_KS4_SENF)	Whether student has any recorded special educational needs.
	Prior attainment at KS2 in English	NPD data (KS4_KS2ENG24P_PTQ_EE)	Prior attainment level at English KS2.
	Prior attainment at KS2 in maths	NPD data (KS4_KS2MAT24P_PTQ_EE)	Prior attainment level at maths KS2.

Step 2: Propensity score estimation

Having identified covariates, the next step was to derive propensity scores for all observations in the sample. Formally, propensity scores represent the conditional probability of being exposed to the BMP given a set of observable covariates. Another way to interpret propensity scores is as a ‘balancing score’. The central idea is that conditional on having similar propensity scores there should be no systematic difference in the covariates between individuals in the intervention and comparison groups that would otherwise lead to bias in impact estimation. The assumptions underlying propensity score theory are set out by Rosenbaum and Rubin (1983).

We used a logit model to estimate propensity scores at the individual level, with a binary outcome taking the value 1 for individuals in the BMP intervention group and 0 otherwise. We used the setting-level indicator of BMP eligibility for the main analysis (as discussed in Participant Selection above, we also estimated impact using an individual-level BMP indicator as an additional analysis). The logit model was estimated using the student- and institution-level covariates described above.

Step 3: Estimating propensity score weights

Our estimation procedure defines a sample Average Treatment Effect on the Treated (ATET) estimand, defined as follows:

$$E(Y_i^1 - Y_i^0 \middle| \mathrm{BMP_i} = 1)$$

Where Y_i^1 is the potential outcome (achieving the required pass in the maths level 2 resit) for student *i* if they were exposed to the BMP, and Y_i^0 is the potential outcome for the same student if they had not been exposed to BMP. BMP$_{\mathrm{i}}$ denotes that the ATET is defined in respect of students who were exposed to BMP in practice.

The ATET formulation means that our findings represent the impact of the intervention for students belonging to BMP schools that were included in the analysis sample. This is different from an alternative estimand called the sample Average Treatment Effect (ATE), which represents the average impact of BMP across the entire student sample (including both BMP students and non-BMP students). The sample ATET was calculated as the difference in outcomes for students in our sample that were exposed to BMP compared to what their outcomes would have been had they not been exposed. This type of analysis involves constructing a comparison group to resemble treatment group students as closely as possible. The derived comparison group represents our best estimate of what outcomes for treatment group students would have been, had they not been exposed to BMP.[20]

To implement the ATET, the next step was to use the estimated propensity scores to define weights that are consistent with a sample ATET estimator. The weights take the value 1 for all observations in the intervention group and 1/(1-π_{ij}) for institutions in the comparison group, where $\pi_{ij}$is the propensity score for individual i in setting j.[21] We also stabilised our weights before weighting the data by dividing by the mean weight.

Our impact estimates carry an Intention to Treat (ITT) interpretation. This means that estimates represent the impact for students of belonging to a setting that promised any funding from the BMP, regardless of whether or to what extent this funding was actually spent during the pilot timeframe.

Step 4: Diagnostic checks on the weighted data

We carried out three diagnostic checks to assess how well the model has worked.

First, we assessed covariate balance before and after weighting by comparing standardised mean differences in each covariate between the intervention and comparison group. If the weighting model were successful, we would expect minimal differences in the distribution of covariates between the intervention group and the comparison group after weighting. We considered an absolute standardised mean difference of 0.1 or below to represent good balance in the distribution of the covariate between the treatment and control group after weighting. Overall, we hoped to find an overall reduction in standardised mean differences after weighting, with fewer than half of the covariates presenting absolute differences over 0.1.

Second, we carried out a visual inspection of common support to assess the extent of ‘overlap’ in the propensity score distributions between the two groups to motivate an effective weighting model. We created histograms showing the distribution in the intervention group and comparison group. If these two distributions were well aligned, this would suggest that there was some a priori similarity between the two groups that could be leveraged to provide a well-balanced sample in the weighted data. If, instead, the distributions had limited overlap, this would be suggestive of strong underlying differences between the intervention and comparison group samples that could be difficult to overcome with quasi-experimental techniques.

Third, we compared the pseudo-R^2 from the logit model of BMP intervention group status before and after weighting. 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. If the weighting model had successfully improved covariate balance we would have expected the pseudo-R^2 to drop after weighting, and ideally be close to 0. A low 0 pseudo-R^2 in the weighted sample is a good indication of good balance between the treatment and derived comparison group.

If any of these diagnostic checks suggested issues with the propensity model we adjusted the model. As per the study plan, we proceeded with this adjustment by firstly expanding the comparison group to include AEA Category 3 areas, secondly re-running the matching model using trimming options. The study plan also allowed for the possibility of revisiting the set of covariates if required.

Step 5: Outcomes regression

The last step was to run a logistic regression modelling the relationship between the outcome and intervention exposure, conditional on the covariates. The doubly-robust estimator incorporates predicted values from this regression together with the propensity score weights to produce an impact estimate. This is equivalent to running a regression on data that is weighted by the inverse propensity score weights and includes additional adjustment for covariates to improve precision.

[20] We define our estimand as a sample ATET rather than a sample ATE because the BMP policy being assessed for this pilot is delivered in AEA Category 5 and 6 areas only. The key causal question of interest, therefore, concerns the impact of the policy for students receiving education in these areas. The comparison group for this evaluation is drawn exclusively from outside AEA Category 5 and 6 areas. Therefore, the impact estimate that is of greatest interest concerns the impact of the intervention for treatment group participants, not the impact of the intervention for participants in other AEA category areas beyond the reach of the pilot intervention.

[21] If we were estimating a sample ATE, we would use a weight of 1/π_{ij} for intervention group observations.

The model was estimated using the `teffects ipwra` command in Stata.[22] We reported our estimate in the form of a relative risk.[23]

For the main primary analysis we carry out this estimation for a pooled BMP intervention group (combining all funding models) compared to a comparison group. This provides a headline estimate for whether the provision of this BMP funding has had an impact on student level 2 maths resit attainment or not.

In line with the study plan, we also estimated separate doubly-robust models for each BMP treatment arm, compared to a comparison group. Since this involves estimating three separate models, we apply a Bonferroni correction, setting the confidence level to $\alpha = 0.05/3 = 0.017$.

Robustness check—two-stage matching

We used a two-stage matching procedure as a robustness check for the doubly-robust analysis. This involved, first, matching BMP institutions with suitably similar comparison group institutions, and then matching individuals within these matched institutions using individual-level characteristics. We used a pooled treatment group for this analysis, comparing students eligible for BMP across all three intervention arms with a matched comparison group.

In the first step, we matched BMP settings to comparison group institutions using a radius matching with replacement approach. The study plan stated that this should be done with a caliper of 0.1 standard deviations of the propensity score, and with a comparison group drawn from settings in AEA Category 4 areas in the first instance. We used the same institution-level and area-level covariates at this stage as for the primary analysis (these covariates are described in Table 9).

We used the same diagnostic checks as for the primary analysis to assess the success of the matching model, namely, (1) comparing covariate balance before and after matching, (2) checking the extent of common support, and (3) comparing the pseudo-R^2 before and after matching. If the model did not satisfy these robustness checks we adjusted it. The study plan anticipated doing this in stages, starting by expanding the available comparison group to include settings in AEA Category 3 and then making other adjustments as needed.

After finalising the institution-level matching model, we retained only the settings that were ‘matched’ for the next step of the estimation and discarded the remainder. The rationale for this stage is to adjust for potential sources of bias that might exist at the setting level by focusing only on comparison settings that are as similar as possible to BMP settings.

We then carried out matching at the student level within these matched settings. This involved estimating a new propensity score model at the student level. We used a logit model for this, containing the same student-level covariates used in the primary analysis (and described in Table 9). We again examined covariate balance for the student-level matching to examine whether matching reduced standardised mean differences between student-level covariates, and assessed the degree of common support and to what extent the pseudo-R^2 was reduced by matching. We used the `teffects psmatch` command in Stata to estimate standard errors using the nonparametric method proposed by Abadie and Imbens (2012).

As with the primary analysis, these impact estimates have an ITT interpretation, reflecting the impact of receiving any BMP funding on student outcomes (regardless of whether and to what extent money was spent in practice). We again report results as a relative risk. Similar to the primary analysis, we also repeated this robustness check using the student-level indicator of BMP exposure.

Relative effectiveness

Research Question 2 examines whether there are differences in the relative impact of the three BMP funding models. We address this research question in two ways.

[22] The functional form of the estimator is in the Stata manual. See StataCorp. (2023) accessible here: https://www.stata.com/manuals/causal.pdf (page 393 begins description of effect parameters estimated by the packages in Stata’s `teffects` suite, with the formula underlying `teffects ipwra` given on page 403).

[23] The coefficient reported by the estimation procedure provides the ATET; that is, the percentage point difference between the BMP and comparison groups in the likelihood of a student receiving the required pass in maths level 2 resit by November 2019. We convert this into a risk ratio by predicting the average likelihoods in both the BMP and comparison groups and taking the ratio of these. Formally, we use the `nlcom` command in Stata to perform this conversion, to enable the construction of a confidence interval around the risk ratio. Following a procedure outlined in Cummings (2011), we first ask `nlcom` to estimate the log of the risk ratio, and then exponentiate the end-points.

First, we performed this analysis using a multilevel logit model that included a dummy variable for each treatment arm. This logit model was weighted using the same propensity score weights as the primary analysis (and so did not involve re-estimating propensity scores for each treatment arm sub-sample).

$$(1) \qquad logit(\pi)_{ij} = \beta_0 + \beta_1 Model\_A_j + \beta_2 Model\_B_j + \beta_3 Model\_C_j + \beta_4 X_{ij} + \beta_5 S_j + u_j + e_{ij}$$

π_{ij} represents the probability that an individual i in setting j received required pass in maths level 2 resits up to November 2019. Model_A$_{j,}$ Model_B$_j$ and Model_C$_j$ captures whether setting j is in the BMP intervention group and part of each funding model respectively. X_{ij} are a range of individual level covariates, and S_j are institution-level covariates. u_j represents a setting random effect and e_{ij} is the individual error term. This specification was weighted by the same weights that produced in the primary analysis and incorporated the same covariates. A Bonferroni correction was applied (setting the confidence level at α = 0.017) to account for the fact that this analysis involves testing three coefficients.

Results from the logit model are expressed as odds ratios. The interpretation of the odds ratio represents the odds of achieving the required pass for students in each of the three treatment arms, relative to the comparison group. An odds ratio of more than 1 indicates a positive impact, and an odds ratio of less than 1 indicates a negative impact. An odds ratio of 1 would indicate that there is no difference in the relative odds of achieving the required pass for a student belonging to one of the three BMP funding models, compared to the non-BMP comparison group.

Secondly, we estimated a single regression carried out over the sample of BMP settings only, to directly compare the relative impact of belonging to each of the three funding models. This analysis directly leverages the randomised assignment of the BMP intervention group and did not require inverse propensity scores weights to mitigate endogeneity of the independent variables. We nonetheless incorporated the same covariates into the regression specification to improve precision. In this case, since the specification involved estimating two coefficients, we applied a Bonferroni correction to the confidence level used to construct confidence intervals of α = 0.025 (which is 0.05/2). This implies the following model (in which funding Model A serves as the reference category):

$$(2) \qquad logit(\pi)_{ij} = \beta_0 + \beta_1 Model\_B_j + \beta_2 Model\_C_j + \beta_3 X_{ij} + \beta_4 S_j + u_j + e_{ij}$$

In the reporting of these results, we qualitatively discuss differences and similarities between the findings of each approach used to explore relative effectiveness.

Subgroup analyses

Research Question 3 asks whether there is a differential impact of BMP according to three subgroups of interest: (1) students who were eligible for free school meals at any time in the five years before entering post-16 education, (2) the number of students per setting eligible to receive funding through BMP, and (3) prior attainment in maths level 2.

The main analysis combined all BMP funding arms into a single overall treatment group. For this pooled analysis, we used the same weights derived from the primary analysis to estimate a multilevel logistic regression with an interaction term between the BMP intervention group and relevant subgroup indicator and a random intercept for each institution. We used the same covariates as the primary analysis. BMP exposure was defined at the setting level and our comparison group was drawn from AEA Category 4 areas (as per the primary analysis). This implies the following model:

$$(3) \qquad logit(\pi)_{ij} = \beta_0 + \beta_1 BMP_j + \beta_2 FSM_{ij} + \beta_3 BMP_j * FSM_{ij} + \beta_4 X_{ij} + \beta_5 S_j + u_j + e_{ij}$$

π_{ij} represents the probability that an individual i in setting j received required pass in maths level 2 resits up to November 2019. BMP_j captures whether setting j is part of the BMP intervention group, FSM_{ij} captures whether individual I in setting j was ever eligible for free school meals in the five years before entering post-16 education. X_{ij} are a range of individual level covariates, and S_j are institution-level covariates. u_j represents a setting random effect and e_{ij} is the individual error term. The specification was defined analogously for the other subgroups of interest. The coefficient of interest in each case is $\beta_{3.}$ This represents any additional impact of BMP for those who are part of the given subgroup, compared to those who are not.

Our main results for this analysis are expressed as odds ratios. The interpretation of the odds ratio associated with the FSM subgroup is the odds of achieving the required pass in maths for FSM-eligible BMP students compared to BMP students who were not. This tells us whether BMP has any differential impact on the odds of passing maths level 2 resits for FSM-eligible students. An odds ratio of more than 1 would mean that members of the subgroup were more likely to experience a positive impact due to BMP than those not in the subgroup. A similar interpretation applies to the subgroup analysis according to prior attainment in maths level 2. In this case, the odds ratio represents any additional differences in BMP impact for students who received a prior attainment of grade 3 in their previous GCSE maths attempt, compared to those who received a grade of 2, 1, or U. The interpretation is slightly different for the institution size subgroup since this is defined continuously rather than in a binary way. Here, the coefficient represents any differences in the impact of BMP on the odds of passing maths level 2 resits for students in settings with x+1 eligible students, compared to students in settings with x eligible students. We also present the findings for this analysis in terms of risk ratios.

As a robustness check, we also evaluated impact for each treatment group separately. This involved re-estimating propensity scores in respect of each individual intervention arm and re-estimating the regression model specified above with updated weights. Note that the subgroup analysis compromises statistical power since the estimation procedure involves partitioning the sample into subgroups. Statistical power is further reduced for the robustness check where subgroup analysis was carried out for each intervention arm separately.

Secondary analysis—dosage analysis

Research Question 4 seeks to understand whether students were more likely to achieve the required pass if they attended institutions that spent a greater amount of BMP funding. The idea behind this research question is that the effectiveness of BMP funding could depend on the total amount of additional resources available for institutions to spend (as opposed to the amount provided per eligible student). For example, institutions with more BMP funding to spend in total may be able to afford measures that are not available to other institutions—such as hiring an additional maths teaching assistant. To address this question, we assessed whether the amount of funding spent by BMP institutions in 2018/2019 had an impact on attainment by November 2019. This involved two analyses: a regression analysis and a marginal effects analysis.

Regression analysis was performed using a multilevel logit model with a binary term for different bands of BMP spending in 2018/2019 by each institution. The lowest band was omitted to avoid multicollinearity and is therefore the reference category against which the relative impact of belonging to other spending bands can be compared. The model is as follows:

$$(4) logit(\pi)_{ij} = \beta_0 + \beta_1 Spendband2_j + \beta_2 Spendband3_j + \beta_3 Spendband4_j + \beta_4 X_{ij} + \beta_5 S_j + u_j + e_{ij}$$

π_{ij} represents the probability that an individual i in setting j received required pass in maths level 2 resits up to November 2019. $Spendband2_j$, $Spendband3_{j,}$ and $Spendband4_j$ are binary variables reflecting whether the amount of BMP funding spent by setting j in 2018/2019 was in the second, third, or fourth band of spending relative to the rest of BMP settings. X_{ij} are a range of individual level covariates and S_j are institution-level covariates. u_j represents a setting random effect and e_{ij} is the individual error term. Since we estimated three coefficients of interest in the specification above ($\beta_{1,}$ $\beta_{2,}$ and β_3), we used a Bonferroni adjustment of α = 0.05/3 to calculate confidence intervals.

The regression was carried out using the sample of all BMP settings that participated in the wave two survey that asked about total BMP funding spent in 2018/2019 and could be successfully linked to NPD data. There was no non-BMP comparison group for this analysis. This means that the coefficients of interest capture any additional impact on the likelihood of passing maths level 2 resits for students belonging to settings that spent more BMP money relative to those in the lowest band of spending.

The marginal effects analysis was used to understand the impact of the 'marginal pound spent' from BMP—that is, the added effect on outcomes of spending one more pound. To carry this out, we first ran the following multilevel logit model over total spend:

$$(5) \qquad logit(\pi)_{ij} = \beta_0 + \beta_1 TotalSpend1819_j + \beta_3 X_{ij} + \beta_4 S_j + u_j + e_{ij}$$

Where $TotalSpent1819_j$ is the total amount of BMP funding spent in the 2018/2019 academic year, and other notation is as equation (4) above. We then used the `margins` command with the option `dydx` in Stata. This used predictions from the previously estimated logit model to produce average marginal effects.

Information on the amount of BMP funding spent in 2018/2019 came from the second wave of the survey that was carried out with BMP settings. Only settings that completed this survey were included in this analysis.

Secondary analysis—retake analysis

Research Question 5 seeks to understand the impact of BMP on whether or not students attempted a GCSE maths retake as opposed to being entered for Functional Skills Level 2 or not attempting any relevant maths resit examination.

This analysis was carried out in the same way as the main doubly robust analysis, but with a different outcome. The outcome of interest for this question was whether a student attempted a GCSE maths resit.

Estimation of the Intra-Cluster Correlation

The BMP trial data has a multilevel structure with students clustered within settings. The Intra-Cluster Correlation (ICC) is defined at the setting level, representing the degree of similarity between students belonging to the same institution. We estimated this using a two-level logit model:

$$(6) \qquad logit(p)_{ij} = \beta_0 + u_j + e_{ij}$$

p_{ij} represents the probability that an individual *i* in setting *j* received required pass in maths level 2 resits up to November 2019. u_j represents a setting random effect and e_{ij} is the individual error term.

The ICC has the following form:

$$(7) \qquad \rho = \frac{\sigma_2^2}{\sigma_2^2 + \frac{\pi^2}{3}}$$

Where $\sigma_2^2$is the variance of the setting-level error term and $\frac{\pi^2}{3}$ is the (constant) variance of the student-level error term (that is, the variance of the logistic distribution).

To fit these models, we first use the `melogit` command in Stata 17 followed by `estat icc` to calculate the ICC. We estimated this ICC in respect of the entire pooled BMP intervention and constructed comparison group using the same final analysis sample as the main primary analysis model.

Changes to trial design since the study plan

There have been several changes to the trial design during the course of this evaluation. For a description of earlier agreed changes, please refer to the study plan.[24] The majority of earlier design changes were a result of the onset of the Covid-19 pandemic during BMP implementation. In this subsection we only discuss the more recent changes that have been made since the most recent study plan was finalised.

Amendment of covariates

The first change relates to minor amendments to the set of covariates included in the statistical analysis. Due to an oversight by the evaluation team, the NPD extract requested did not contain the correct data to fully derive all covariates as per the study plan. In cases where we were unable to fully define a covariate in the manner described in the study plan we have instead used a proxy that is as similar as possible to what was originally intended, and in the same spirit. The extent of these deviations is slight and is not expected to make a meaningful difference to the results of our analysis. These differences are described in detail in Appendix H (Additional Technical Appendices document).

Additional analysis—estimating the impact of BMP as a whole

The study plan indicated that we would carry out primary analysis using separate estimation models in respect of each BMP funding model. In addition to this analysis, we have also estimated the impact of BMP as a whole, combining all funding models into one. This provides a headline estimate of whether the overall provision of BMP funding has had an impact on student level 2 maths resit attainment of not.

Additional analysis—estimating impact according to student-level BMP eligibility

As discussed in the Participant Selection section above, our main impact estimation uses the setting-level indicator of BMP exposure. The team also produced additional analysis, as a robustness check, to see how findings vary using an alternative student-level indicator of BMP exposure.

[24] The version history at the top of the document summarises key changes made across versions of the study plan and the relevant sections of the plan that these changes pertain to.

Additional analysis—exploiting experimental assignment for relative effectiveness analysis

We also chose to add a further piece of simple additional analysis to compare the relative effectiveness of each treatment arm. The planned analysis for Research Question 2, as documented in the study plan, involved comparing each treatment arm with the derived comparison group. We extended this to include an additional analysis that directly invokes the randomised assignment of the BMP funding models. This involves simply comparing the relative effectiveness of each funding arm in relation to the others, with no non-BMP comparison group.

To implement this analysis, we focused on the BMP group only and ran a single multilevel regression model that included a term for belonging to funding Model B or C. The model included the same covariates as the primary analysis in order to improve precision but did not require any quasi-experimental adjustment to mitigate bias. Given the randomised assignment to BMP funding arms, we did not need to adjust for any underlying systematic differences between the groups that might have otherwise led to bias in comparing them.[25] We used a Bonferroni correction to compute confidence intervals and p-values for this analysis. Since this approach involved estimating two coefficients, we applied a significance level of α = 0.025 (which is 0.05/2).

This is informative of the relative impact of Models B and C over and above Model A. This is a valuable addition since there is no other aspect of the methodology that directly invokes the randomised structure of treatment assignment for impact estimation. This means that this analysis provides the most robust means possible from this data of drawing unbiased causal inference, albeit for a measure of relative effectiveness only.

Content of data requests

When the study plan was written we anticipated requesting access to DfE data in two distinct batches: a first request covering covariate data and a second covering outcome data. The reason for this was to provide an extra safeguard against the risk of 'data mining' in the specification of impact estimation models. The idea was that we would specify all impact estimation models using the first batch of data only, which contained no outcomes information. In practice, although we did receive data in two separate batches, the content of each batch was not as initially intended. The first did not contain all information needed to derive the intended covariates, meaning that additional data to derive covariates needed to be added into the second data request (alongside outcomes information). The practical impact of this change on the impact estimation is negligible. We have conducted analysis against the prespecified study plan and documented and justified any changes from the original design to minimise any risk of bias.

Stata command used for primary analysis

In the study plan we intended to use the command `teffects aipw` for the doubly robust estimation. In practice we have used `teffects ipwra` because it contains an option for calculating ATET estimates, which `teffects aipw` does not. Other differences between these two commands are minimal: both implement doubly robust estimation as per the study plan.

The `teffects ipwra` command, as with other packages in Stata's `teffects` suite, estimates standard errors that correct for the three-step process of estimating inverse probability weights, fitting a weighted regression model for each treatment level, and constructing the estimate by comparing means of the predicted outcomes for each treatment level. For this reason, we do not additionally bootstrap the standard errors produced by our analyses (as indicated in the study plan). Bootstrapping is computationally intensive and we estimated that it would take several days to complete within the ONS SRS environment with any reasonable number of bootstrap repetitions (for example, 1,000).

Implementation of the primary analysis

The study plan indicated that the primary analysis would involve:

- at the propensity score estimation stage, combing all BMP funding arms into one pooled intervention group; and
- then, using these 'pooled' propensity scores, estimate separate doubly robust models for each funding model.

[25] Given withdrawals from the trial since settings were initially randomised, there is a theoretical risk that the random allocation across treatment arms was not preserved. This is discussed further in the Attrition section below.

In practice, we have implemented this analysis by fitting separate doubly robust models that use propensity scores estimated in respect of each funding model in turn (rather than 'pooled' propensity scores). This is because the Stata command used for the primary analysis is 'all in one' and does not permit propensity scores to be estimated over a different sample than the estimation sample.

Implementation and process evaluation

Research methods

The IPE synthesised the data collected from initial and follow-up interviews with heads of maths at treatment institutions and three surveys of treatment institutions. All survey and interview materials were reviewed and agreed with the DfE and data was collected by experienced NatCen researchers. The IPE research methods are described below and summarised in Table 10.

Initial head of maths interviews

Twenty-four heads of maths (eight from each treatment arm) were interviewed via telephone. Participants were sampled according to the number of eligible students at the institution (four high and four low) and the percentage of maximum possible funding spent as reported in the wave one survey (four high and four low). The interviews aimed to explore usual practice to support GCSE maths re-sitters, reasons for interest in the pilot, how BMP funding was spent and expected to be spent, decision-making processes on how to spend the funds, and expected outcomes. The interview guide was created in-house by NatCen researchers based on research questions and the theory of change (ToC) (see Technical Appendices: Appendix C). Interviews were conducted in May and July 2019 and lasted 60 minutes. No incentives were used.

Follow-up head of maths interviews

Further follow-up interviews with 12 heads of maths (four from each treatment arm) were conducted post-intervention via telephone. The aim of these was to gather information on the provision in 2019/2020 for students resitting GCSE maths, activities introduced using BMP funding, views on, and suggested improvements, for the BMP, and perceived outcomes of the intervention. The interview guide was created in-house by NatCen researchers (see Technical Appendices: Appendix D). Interviews were conducted in October 2020 and lasted 45 minutes. No incentives were used.

Wave 1 institutions survey

We invited all the institutions involved in the pilot to take part in short surveys at three timepoints to monitor the planning and spending of the funding. All three surveys were administered online and sent to the main contact for the pilot at the institution. We also sent a PDF version to enable the main contact to consult other members of staff if needed. The first survey was sent to 439 institutions and achieved 296 responses (67% response rate). It aimed to find out the resources in place to support GCSE maths resit students aside from the BMP, how clear the aims of the BMP were, conditions for participation, and how institutions intended to support students. The survey was created in-house by NatCen (see Technical Appendices: Appendix E) and administered in February and March 2019. The expected completion time was 15 minutes, with some additional time if discussions were needed with colleagues. No incentives were used.

Wave two institutions survey

The second survey was sent to 434 institutions and achieved 246 responses (57% response rate). It aimed to explore institutions' responsiveness to the intervention including how much BMP funding they spent, how they allocated the additional funding, the sources they used to inform their decision-making, and the challenges faced when allocating funding. The survey was created in-house by NatCen (see Technical Appendices: Appendix F) and administered in November 2019. The expected completion time was ten minutes. No incentives were used.

Wave 3 institutions survey

The third survey was sent to 434 institutions and achieved 118 responses (27% response rate). It aimed to find out the resources in place to support GCSE maths resit students aside from the BMP, how much BMP funding institutions had spent, how they used the funds, and how they would allocate any funding that had not been spent yet. It also explored the costs of delivering the intervention in terms of additional staff hours to allocate the BMP. Finally, it investigated perceived outcomes, any unintended consequences, views on the funding models, and overall experiences of the BMP.

The survey was created in-house by NatCen (see Technical Appendices: Appendix G) and administered in October 2020. The expected completion time was 15 minutes. No incentives were used.

Analysis

Interviews were digitally audio-recorded and professionally transcribed. Framework in NVivo, a systematic approach to qualitative data management developed by NatCen, was used to summarise transcribed data by theme and case (Ritchie et al., 2013). The summarised data was then analysed thematically using an inductive approach. This involved a phase of 'detection', which included identifying what participants said about a given topic and categorising this thematically. Once we had identified different themes in the data, we created higher-level classifications that worked as meaningful conceptual groupings for participants' views and experiences.

Quantitative data was analysed in SPSS by means of frequencies and cross-tabulations (to assess the relations between respondent characteristics and survey responses and test their statistical significance). We analysed all the data available for the three waves of institution surveys. Syntax files were used to ensure a record was kept of how analysis was conducted.

Quantitative and qualitative findings were triangulated and synthesised thematically by the main implementation dimensions to provide a comprehensive assessment of implementation and help to explain the impact evaluation findings. To minimise bias and ensure rigour, multiple researchers cross-analysed and cross-interpreted data.

Table 10: IPE methods overview

Research methods	Data collection methods	Participants/ data sources	Data analysis methods	IPE Research questions addressed	Implementation/ logic model relevance
Qualitative	Interviews (initial)	Heads of maths at treatment institutions (24)	Thematic analysis	RQ1, RQ2, RQ4, RQ5, RQ8, RQ9, RQ10, RQ12, RQ13, RQ14, RQ15, RQ17, RQ18, RQ19	Differentiation, responsiveness, evidence of promise
Qualitative	Interviews (follow-up)	Heads of maths at treatment institutions (12)	Thematic analysis	RQ1, RQ3, RQ4, RQ11, RQ12, RQ13, RQ14, RQ17, RQ18, RQ19	Differentiation, responsiveness, evidence of promise
Quantitative	Survey (wave one)	Treatment institutions (296)	Descriptive statistics	RQ1, RQ8, RQ9, RQ10, RQ11, RQ12, RQ15	Differentiation, responsiveness
Quantitative	Survey (wave two)	Treatment institutions (246)	Descriptive statistics	RQ5, RQ8	Differentiation, responsiveness
Quantitative	Survey (wave three)	Treatment institutions (118)	Descriptive statistics	RQ1, RQ6, RQ7, RQ11, RQ16, RQ17, RQ18	Differentiation, dosage, responsiveness, evidence of promise

Costs

Costs incurred by institutions to deliver BMP

A cost evaluation was conducted to compare the average cost of the intervention per funding model from the perspective of participating institutions. Costs were evaluated in accordance with EEF guidelines.[26] Cost data was collected for providers in the three different treatment groups through the institution surveys.

We used this data to construct a per-student, per-year cost. The average cost per year of the intervention to participating settings was repeated over three years to highlight the relative proportions of start-up versus running costs.

Costs of the pilot for the DfE

In addition to the costs of BMP for institutions presented above, we also explored the costs of the programme for the DfE in terms of the amount of funding provided per student.

We do not have access to DfE data on the amounts of BMP funding actually provided over the course of the pilot. However, we can present a simple estimate of the average per-student cost of delivery by calculating total spend according to the intended funding allocations under each model, and the number of eligible students.

For Model A this is simply £500 per eligible student:

$$\frac{\#\ model\ A\ pupils * £500}{\#\ model\ A\ pupils}$$

For Model B this is:

$$\frac{\#\ model\ B\ pupils * £250 +\ (\#\ model\ B\ pupils\ that\ achieved\ a\ pass) * £250}{\#\ model\ B\ pupils}$$

For Model C this is:

$$\frac{(\#\ model\ C\ pupils\ that\ achieved\ a\ pass) * £500}{\#\ model\ C\ pupils}$$

Timeline

Table 11: Timeline

Dates	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–Sep 2019	Develop theory of change, protocol, and study plan	NatCen, DfE
Nov 2018–Jan 2019	Student enumeration and student and setting data collection, informing students about data processing	NatCen, DfE

[26] https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/Cost_Evaluation_Guidance_2019.12.11.pdf

Feb - Mar 2019	First online survey of treatment settings	NatCen
Mar–Jun 2019	Intervention—additional funding for post-16 settings disbursed in pilot areas	DfE
May–Jul 2019	Initial interviews with heads of maths	NatCen
Oct 2019	Baseline data collection from NPD and school and college performance tables	NatCen, DfE
Nov 2019	Second online survey of treatment settings	NatCen
Oct 2020	Third online survey of treatment settings	NatCen
Oct 2020	Follow-up interviews with heads of maths	NatCen
Feb 2021	Revision of study plan and NPD data request (outcomes data)	NatCen
Mar 2021	Access baseline data for matching	NatCen
Apr 2022	Receipt of outcomes data for impact analysis	DfE
Feb 2022–Apr 2023	Impact analysis and Draft report	NatCen
Jul 2023	Peer review, final report	NatCen, DfE, EEF

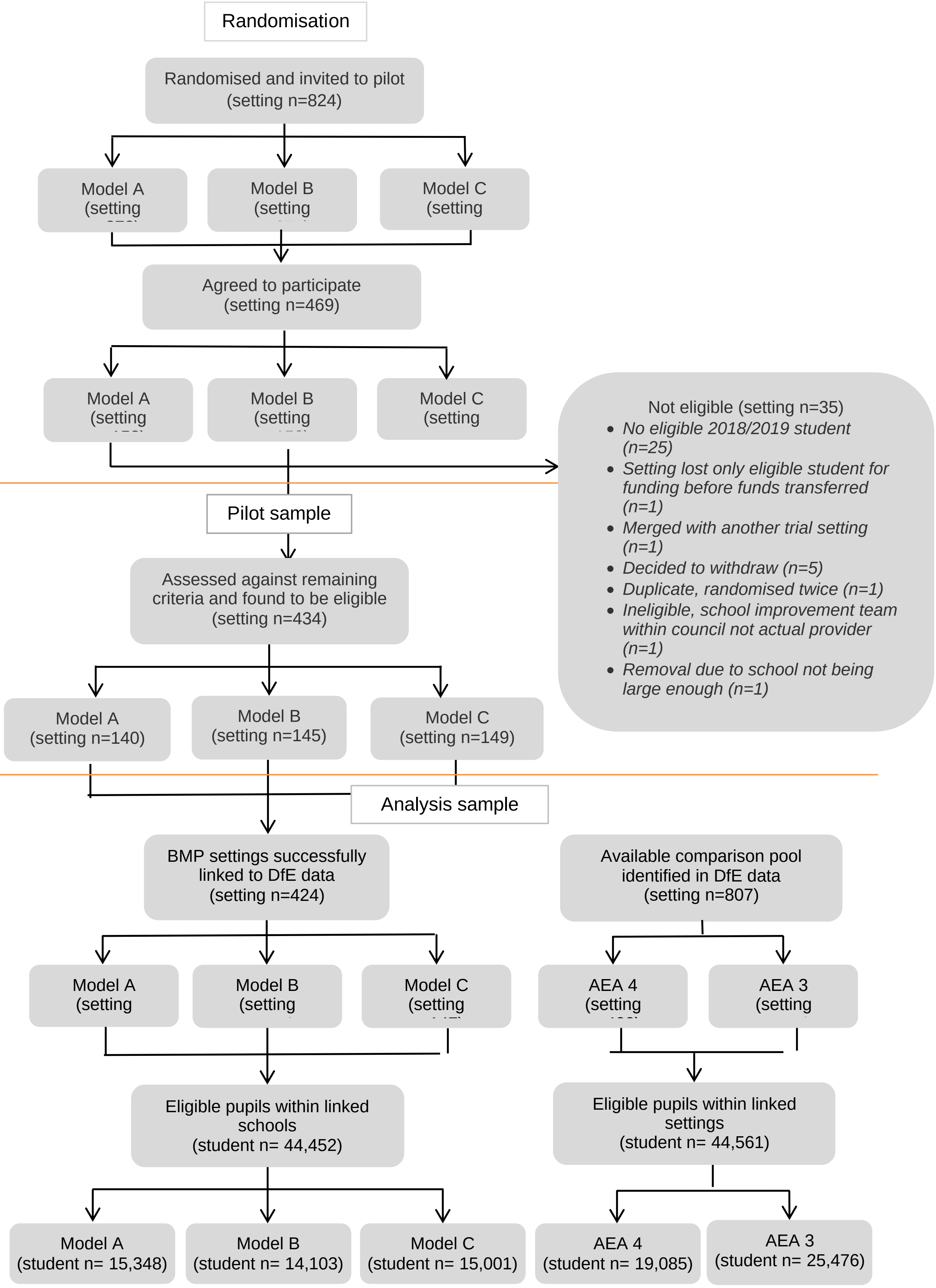
Impact evaluation results

Participant flow including losses and exclusions

Participant flow

Figure 2 illustrates the flow of participants through each stage of the evaluation, from setting randomisation to BMP to the final analysis sample successfully linked to outcomes data in the NPD and ILR datasets. The reasons for attrition at different stages of the trial are discussed in the following section.

Figure 2: Participant flow


Randomisation
Randomised and invited to pilot (setting n=824)
Model A (setting
Model B (setting
Model C (setting
Agreed to participate (setting n=469)
Model A (setting
Model B (setting
Model C (setting
Not eligible (setting n=35)
No eligible 2018/2019 student (n=25)
Setting lost only eligible student for funding before funds transferred (n=1)
Merged with another trial setting (n=1)
Decided to withdraw (n=5)
Duplicate, randomised twice (n=1)
Ineligible, school improvement team within council not actual provider (n=1)
Removal due to school not being large enough (n=1)
Pilot sample
Assessed against remaining criteria and found to be eligible (setting n=434)
Model A (setting n=140)
Model B (setting n=145)
Model C (setting n=149)
Analysis sample
BMP settings successfully linked to DfE data (setting n=424)
Available comparison pool identified in DfE data (setting n=807)
Model A (setting
Model B (setting
Model C (setting
AEA 4 (setting
AEA 3 (setting
Eligible pupils within linked schools (student n= 44,452)
Eligible pupils within linked settings (student n= 44,561)
Model A (student n= 15,348)
Model B (student n= 14,103)
Model C (student n= 15,001)
AEA 4 (student n= 19,085)
AEA 3 (student n= 25,476)

Updated sample size calculations in final analysis data

Table 12 reports MDES calculations for this evaluation at the study plan stage and updated calculations based on the final evaluation sample data. The MDES captures the smallest change in the primary outcome (expressed in standard deviation units) that could be detected by the trial with a given level of power. In this case, the primary outcome is whether or not a student achieved the required pass in maths level 2 resits by November 2019.

At the time of the study plan, sample size calculations reflected our expectations about the likely numbers of eligible students per setting that would be found in the final NPD datasets and the level of the ICC. The ICC shows the degree of similarity between students belonging to the same cluster (in this case, setting). The higher the ICC, the greater the resemblance between students belonging to the same cluster. This means that each additional student contributes less unique information to the sample and power to detect small outcome changes is weakened. We can now update the calculations carried out at the time of the study plan by replacing the assumptions made at the time with the actual sample sizes and ICC values obtained in our final data.

Power calculations are carried out using the PowerUp! software (Maynard and Dong, 2013), based on statistical power of 0.8. For the calculations of MDES sizes in the main (pooled) analysis, and subgroup analysis, we use a significance level of α = 0.05. For calculations of the impact of each treatment arm separately, we apply a Bonferroni correction to adjust for the fact that this analysis involves estimating three coefficients (one for each funding model). This implies a significance level of α = 0.017 (which is 0.05/3).

At the time of the study plan, we did not have access to the final analysis data and we did not adjust our MDES calculations to account for any assumptions about the pre-test/post-test correlation. This means that the original calculations were conservative. In practice, the impact analysis procedure does involve adjusting for a range of covariates that are predictive of final outcomes. Covariate adjustment enables more precise estimation, which leads to gains to statistical power. We can now update the sample size calculations to account for the additional precision gained through the inclusion of covariates. To make this revision, we focus solely on the association between a measure of pre-BMP attainment and the outcome. The measure of pre-BMP attainment used is attainment at KS2 maths (variable KS2_KS2MAT24P_PTQ_EE in the NPD). This is only one of the covariates that we adjust for in our impact estimation procedure in practice (see Table 9 for the full list). The rationale for incorporating correlations for this covariate only is that most EEF trials typically include prior attainment as the sole analysis covariate, along with group status and any design characteristics, in line with EEF Statistical Analysis Guidance (EEF, 2022). Reporting sample size calculations that account for adjustment using this covariate only may therefore be of value to future researchers on EEF trials. Given that we do not show MDES sizes accounting for the inclusion of all our covariates, the reported MDES sizes in Table 12 may still be considered conservative.

Table 12 shows that actual MDES sizes are larger than anticipated in the study plan. This is largely because the final analysis does not include as many BMP settings as anticipated. For reasons that will be expanded on in the main findings section below, the primary analysis has proceeded with the exclusion of some settings in the final analysis dataset—those found to be too distinctive to be considered comparable with settings in the weighted comparison group. The ICCs calculated in our final dataset are also slightly higher than initially expected in the study plan, which again causes the MDES to increase.

Despite these differences, the MDES estimates we find in our final analysis samples remain small.

Converting the MDES expressed as a standardised mean difference into risk ratio units

Table 12 also translates the MDES that is measured as a standardised mean difference (in standard deviation units) into an estimated effect size expressed as a risk ratio. The purpose of this is to show how the minimum effect sizes correspond with the units that the majority of our impact analysis is reported in. We carry out this conversion as follows:

The original MDES in standard deviation units is given by the following formula (where $\bar{Y}_t$ and $\bar{Y}_c$ represent the mean outcome in the treatment and comparison group respectively):

$$MDES = \frac{\bar{Y}_t - \bar{Y}_c}{\sigma}$$

Since our primary outcome is binary, the formula for the variance is:

$$\sigma^2 = Y_c(1 - Y_c)$$

Where Y_c is the proportion of students achieving the required pass in the comparison group. We can consequently convert the standardised mean difference MDES that is expressed in standard deviation units into a risk ratio as follows:

$$Risk\ ratio = \frac{\left(MDES\sqrt{Y_c(1 - Y_c)}\right) + \bar{Y}_c}{\bar{Y}_c}$$

In order to convert the original study plan MDES estimates into a risk ratio, we assume that the proportion of comparison group students achieving the required pass in their maths level 2 resit exam is 24% for the main sample and 17% for the FSM subgroup. These were the assumptions made at the time the study plan was written based on GCSE attainment data for prior cohorts set out in ‘The Attainment Gap’ (EEF, 2018). For the updated analysis stage calculations, we use a comparison group proportion of 13.3% in the main sample and 9.5% for the FSM subgroup. This corresponds to the actual proportions observed in the final dataset retained for the doubly robust impact analysis (see the Primary Analysis section below for details on how the final sample retained for this analysis was determined and see Appendix Table 12 in the Additional Technical Appendices document for the comparison group proportions used in this conversion).

Table 12: Minimum detectable effect sizes at different stages

		Study Plan			**Analysis**		
		Per treatment arm	Overall	FSM	Per treatment arm	Overall	FSM
MDES expressed as a standardised mean difference		0.21	0.10	0.15	0.27	0.14	0.18
MDES expressed as a risk ratio		1.37	1.18	1.33	1.69	1.36	1.56
Pre-test/post-test correlations	Level 1 (student)				0.19	0.19	0.18
	Level 2 (class)						
	Level 3 (school)						
Intracluster correlations (ICCs)	Level 2 (class)	N/A	N/A	N/A	N/A	N/A	N/A
	Level 3 (school)	0.20	0.20	0.20	0.28	0.28	0.29
Alpha		0.0167	0.05	0.05	0.0167	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2	2	2
Average cluster size1		7.96	7.96	1.91	4.07	4.63	2.66
Number of schools	Intervention	145	434	434	80	280	243
	Control	145	434	434	402	399	310
	Total:	290	868	868	482	679	553

Number of students	Intervention	16,301	48,903	11,736	8,166	24,234	9,281
	Control	16,301	48,903	11,736	23,080	23,076	8,827
	Total:	32,602	97,806	23,473	31,246	47,310	18,108

1. MDES calculations are performed using PowerUp! (Dong and Maynard, 2013). We assume a two-level multilevel model, with students clustered within settings.
2. Average cluster size is calculated using the harmonic mean. Harmonic means are recommended for use in power calculations in situations where cluster sizes vary because they are less sensitive to extreme outliers than the arithmetic mean. If cluster sizes are identical then the harmonic mean is equivalent to the arithmetic mean.
3. Updated sample sizes and ICC for the per-treatment arm analysis are based on the final sample used in the doubly robust model performed for Model A. Updated sample sizes and ICC for the pooled intervention group analysis are based on the final sample used for the main primary analysis model. Updated sample sizes and ICC for the FSM subgroup analysis are based on the final sample used for the main primary analysis model, restricted to students who were ever eligible for free school means at any time in the five years before entering post-16 education (variable YPMAD_EverFSMage10to15 in the NPD).
4. The study plan assumed equal sizes of the intervention and comparison groups. This turns out to be a reasonable assumption for the analysis of the overall pooled intervention group and subgroup analysis. For the per treatment group analysis, the intervention group sample is smaller than the comparison group since it has been divided into three.
5. We have lightly updated the MDES calculations originally presented in the study plan. At the time when power calculations were done for the study plan, 436 settings were remaining in the trial. After this, a further two dropped out. We have now amended the study plan calculations to include the 434 settings that ultimately took part in BMP and also correct for inconsistencies in the application of Bonferroni adjustment for the initial study plan calculations.

Attrition

Level of attrition and reasons for attrition

As shown in the preceding section, participants were lost from the trial at various stages. First, some institutions left the trial between randomisation and the signing of MOUs. Among the 824 institutions initially randomised, 57% (469) agreed to participate.

From these 469, a further 35 institutions withdrew for the following reasons:

- institution had no eligible students (25 institutions);
- the institution lost its only eligible student before funds were transferred (one);
- the institution was merged with another institution in the trial (one);
- institution chose to withdraw (five);
- a duplicate institution had been randomised twice (one);
- ineligible, school improvement team within council not actual provider (one); and
- removal due to school not being large enough (one).

This left 434 institutions (53%) remaining in the trial, which were expected to form the key analysis sample.

The next key stage where some of the sample was lost was in linking records to our data extract in the NPD. From these 434 institutions, we successfully linked 424 of them to any student data from our NPD extract. There were ten settings where we could not identify any eligible Year 12 student in our NPD data extract.

For this trial there is no experimentally chosen control group. The comparison group simply consists of students in AEA Category 4 (or 3) areas who were estimated to meet the eligibility criteria for the programme. We do not therefore encounter 'attrition' from the comparison group per se as no non-BMP settings were ever approached for inclusion in the quasi-experimental analysis.

A final reason for loss of sample in our analysis is related to the process of cleaning and preparing the NPD data. As will be shown below, there were few missing values in this data for any of the chosen covariates or the key outcome. Observations that had any missing values were not included in the final estimation models.

The attrition experienced on the trial influences statistical power and the interpretation of impact estimates. Below, we discuss each of these implications in turn.

Implications of attrition for statistical power

Firstly, attrition leads to a reduction in statistical power to detect changes in outcomes. Table 12 above displays updated MDES calculations for this trial that account for the actual rate of attrition that arose in practice. As discussed in the above subsection, the attrition observed in this trial leads to an MDES that is larger than that initially anticipated in the study plan. However, MDES estimates remain small given the large sample size for outcomes data that we are able to access through the NPD. This means the attrition does not result in a trial that is under-powered.

Implications of attrition for quasi-experimental analysis

The majority of the impact analysis contained in this report is produced through a quasi-experimental analysis. For all this QED analysis, attrition is relevant for interpreting the resulting characteristics of the intervention group population that the findings relate to. Attrition does not, however, run any risk of introducing bias in this case.

To understand how attrition affects the interpretation of QED findings, recall that the QED approach involves developing a non-BMP comparison group that is designed to resemble the intervention group as closely as possible in all respects apart from exposure to the BMP funding. This method of analysis is consistent with an ATET estimand in which the results capture the average impact of BMP for settings that received any funding and were included in the analysis sample.

Bias could, in theory, be affected if attrition had an influence on the extent of systematic differences between the intervention group and non-BMP comparison group. However, for the QED analyses, there is no reason to believe that attrition will make a difference to the risk that the selected comparison group is not readily comparable to the BMP intervention group. This is because under the QED analyses, the non-BMP comparison group is not ‘fixed’ in advance but is instead specifically refined to resemble as closely as possible the BMP intervention group at hand. In other words, the QED approach is specifically designed to overcome bias for whatever intervention group remains in the sample after attrition. The derived comparison group can vary to adjust for any changes in the characteristics of the treatment group that are brought about by attrition.

The main implication of attrition for the QED analyses is to influence the characteristics of the sample that the impact findings relate to. The ATET interpretation of our impact estimates means that they capture the average effects of BMP for exactly those settings and students that are retained in the trial after attrition. These attributes are described in Table 13 and Table 14 of the next section. The QED analysis therefore provides a good estimate of the effect of the intervention on settings that agreed to take part and received funding.

Implications of attrition for experimental analysis

An additional analysis performed as a robustness check for RQ3 is the only analysis in this report that directly exploits the randomised allocation of BMP. This experimental analysis (described in the Methods section) involves comparing the three randomly assigned intervention funding models with one another to assess the relative effectiveness of each. For this analysis, attrition does present a potential risk of bias.

By design, random allocation into different experimental treatment arms should prevent bias by ensuring that, on average, there are no differences in the characteristics of schools and students assigned to each group. This property of randomisation may, however, be jeopardised if settings with certain characteristics are more likely to withdraw from the trial if they are assigned to particular treatment arms. For example, it is plausible that settings where heads of maths did not anticipate that many students resitting maths would receive the required pass would be more likely to withdraw from the trial if they were assigned to a payment-by-results model instead of the guaranteed payments model. In other words, if the propensity to withdraw is systematically related both to treatment allocation and to the likelihood of achieving the required outcome, then attrition could introduce bias to this experimental analysis.

We cannot rule out the possibility that the level of attrition in this trial leads to this kind of bias for this robustness check analysis. However, Figure 2 does indicate that the extent of attrition was similar across the three funding models. Among the 824 settings originally randomised, a similar proportion of the settings in each funding model were retained in the analysis sample (50% of Model A settings, and 52% of Model B and C settings). This provides some assurance that there has not been significant non-differential attrition from across the treatment arms.

Student and school characteristics

Covariate descriptives

Table 13 presents setting-level descriptive information for the BMP and non-BMP sample and Table 14 shows student-level descriptive information. These tables capture the characteristics of the sample before any statistical weighting or matching is carried out.[27] The treatment group consists of BMP settings (or students within those settings) in AEA Category 5 and 6 areas, while the comparison group consists of non-BMP settings (and students within those settings) and AEA Category 4. For these tables we pool the three BMP funding models together to form the treatment group.

We draw several reflections from this table. First, we note that the two groups are remarkably well balanced in terms of individual student-level characteristics with the majority of standardised mean differences being close to zero. There are minimal differences between BMP and non-BMP students in terms of their average age in months, gender, special educational needs status, FSM status, and prior attainment in English and maths at KS2. We do observe some differences in ethnicity—with students in the BMP intervention group being somewhat more likely to be white—but, overall, this table indicates that there is substantial alignment in student-level characteristics between our treatment and comparison groups.

The picture is different at the institution level. We find that BMP and non-BMP settings differ markedly in terms of the characteristics used to define AEA group. This is no surprise since BMP settings are located in areas with a different AEA classification to all areas in the comparison group. For example, we find that BMP settings have lower average Attainment 8 scores than non-BMP settings, a lower percentage of students in schools with 'good' or 'outstanding' leadership, and a lower proportion of students are recorded as having access to a good secondary school within 5km of their home. There is also a lower density of lead or outstanding sponsor academies and trainees. These findings are consistent with BMP being purposively targeted towards areas with higher need. In terms of other institution-level characteristics, we also observe some differences between the BMP and non-BMP group. For example, BMP settings are bigger and are distributed differently across regions; the two groups are, however, similar in terms of the proportion of students who have ever been eligible for free school meals and in the settings' proportion of Year 12s who achieve at least level 2 in maths and English.

A second observation from these findings is that while the intervention and comparison groups contain a similar number of settings, there are many more BMP students than comparison group students. This reflects the fact that the BMP schools are typically larger and contain more eligible students than non-BMP schools.

Finally, we note that there are very few missing values in any of these covariates. Given the low incidence of missingness in this data, we do not undertake further exploration of the patterns of missingness, nor attempt inverse probability weighting techniques to adjust for missingness. Note that these tables present a slightly more broken-down version of the covariates that we used in our analysis. Specifically, in the statistical analysis we use slightly broader categories for institution type, region, and ethnicity than those shown here (see Table 9 for more details).

In Appendix I of the Additional Technical Appendices, we present these descriptive results for each treatment arm separately. We find some differences. Differences in setting- and institution-level covariates appear to be most pronounced for institutions in Model A, and least pronounced for those in Model B.

We also report descriptives for the pooled intervention group against a comparison group that includes non-BMP settings (and students) from AEA Category 3 as well as AEA Category 4. From these additional results, we can observe that standardised mean differences between the BMP and comparison group are larger when the comparison group includes settings in AEA Category 3 areas. This is as expected given how AEA Categories are defined: AEA Category 3 is a more distant category from AEA Categories 5 and 6 (which the BMP settings are drawn from).

[27] The impact estimation findings reported later in this section will show weighted results.

Table 13: Institution-level covariate descriptives

Covariate		Unit	Treatment group				Comparison group				Standardised mean difference
			n	Mean	Standard deviation	# missing	n	Mean	Standard deviation	# missing	
Achieving Excellence Area covariates	Access to a good secondary school index	Percentage	424	56.92	17.42	0	422	66.54	13.78	0	-0.59
	AEA achievement indicator	Number	424	46.69	2.34	0	422	48.41	1.96	0	-0.74
	AEA progress indicator	Number	424	-0.14	0.16	0	422	-0.03	0.10	0	-0.75
	System leader coverage indicator	Number	424	2842.30	1427.57	0	422	1707.78	768.54	0	0.89
	Initial teacher training provider coverage index	Number	424	39.79	19.23	0	422	43.42	18.37	0	-0.19
	Quality of leadership indicator	Percentage	424	71.78	14.90	0	422	81.40	9.52	0	-0.72
	Academy sponsor coverage	Number	424	1616.40	1545.13	0	422	700.00	462.30	0	0.75
Other setting-level characteristics	Institution size	Number	424	1999.82	2366.39	0	414	1620.29	2263.69	<10	0.16
	Number of Year 12s in 2018/2019 meeting eligibility criteria for the programme	Number	424	102.26	200.89	0	422	60.37	164.44	0	0.23
	Percentage of Year 12s in 2018/2019 meeting eligibility criteria for the programme	Percentage	424	27.42	20.51	0	422	26.28	24.60	0	0.05
	Institution type: Academy converter	Percentage	424	28.77	45.32	0	422	32.46	46.88	0	-0.08
	Institution type: Further education sector	Percentage	424	25.71	43.75	0	422	15.17	35.91	0	0.26
	Institution type: Academy – sponsor-led	Percentage	424	20.28	40.26	0	422	13.74	34.47	0	0.17
	Institution type: Other	Percentage	424	22.41	41.75	0	422	33.18	47.14	0	-0.24
	Institution type: Missing	Percentage	424	2.83	16.60	0	422	5.45	22.73	0	-0.13
	Region: North East	Percentage	423	3.55	18.52	<10	416	1.44	11.94	<10	0.13
	Region: North West	Percentage	423	17.73	38.24	<10	416	8.41	27.79	<10	0.28
	Region: Yorkshire and the Humber	Percentage	423	13.95	34.69	<10	416	18.51	38.88	<10	-0.12
	Region: East Midlands	Percentage	423	16.31	36.99	<10	416	4.81	21.42	<10	0.37
	Region: West Midlands	Percentage	423	20.09	40.12	<10	416	15.87	36.58	<10	0.11
	Region: East	Percentage	423	12.06	32.60	<10	416	12.50	33.11	<10	-0.01
	Region: South East	Percentage	423	14.42	35.17	<10	416	17.79	38.29	<10	-0.09
	Region: South West	Percentage	423	1.65	12.77	<10	416	16.83	37.46	<10	-0.53
	Region: London	Percentage	423	0.24	4.86	<10	416	3.85	19.25	<10	-0.26
	Percentage of students who were eligible for FSM between age 10 and 15	Percentage	424	25.71	15.22	0	422	24.61	20.09	0	0.06
	Percentage of Year 12s in 2018/2019 with Level 2+ in maths and English at KS4 prior to enrolment	Percentage	424	59.60	26.62	0	422	57.93	33.01	0	0.06

The treatment group sample is all BMP settings in AEA Category 5 and 6 areas. The comparison group sample is settings in AEA Category 4 areas that had at least one student meeting the eligibility criteria for BMP. Sources are NPD, ILR, and BMP sample file information. Cells with 1–9 observations are suppressed and replaced with '<10' as a statistical disclosure control practice.

Table 14: Student-level covariate descriptives

Covariate		Unit	Treatment group				Comparison group				Standardised mean difference
			n	Mean	Standard deviation	# missing	n	Mean	Standard deviation	# missing	
Student-level covariates	Age	Number	44,250	16.01	0.07	202	25,476	16.01	0.07	0	0.00
	Gender: Male	Percentage	44,452	50.82	49.99	0	25,476	51.30	49.98	0	-0.01
	Ethnicity: White	Percentage	43,349	79.76	40.18	1,103	24,712	74.33	43.68	764	0.13
	Ethnicity: Mixed/ multiple	Percentage	43,349	4.51	20.74	1,103	24,712	5.08	21.96	764	-0.03
	Ethnicity: Asian	Percentage	43,349	10.77	31.00	1,103	24,712	12.71	33.31	764	-0.06
	Ethnicity: Black	Percentage	43,349	3.97	19.51	1,103	24,712	6.30	24.31	764	-0.11
	Ethnicity: Other	Percentage	43,349	1.00	9.94	1,103	24,712	1.57	12.43	764	-0.05
	Eligibility for free school meals	Percentage	44,452	38.29	48.61	0	25,476	37.71	48.47	0	0.01
	Special Education Need status (SEN)	Percentage	43,667	27.10	44.45	785	24,998	27.42	44.61	478	-0.01
	Prior attainment at KS2 in English*	Number	41,559	3.66	0.87	2893	23,580	3.68	0.87	1896	-0.02
	Prior attainment at KS2 in maths*	Number	41,557	3.55	0.73	2895	23,576	3.56	0.72	1900	-0.01

The treatment group sample consists of all settings and students belonging to BMP institutions in AEA Category 5 and 6 areas. At the student level, this sample corresponds to students who enrolled in a post-16 study programme in 2018/2019 and did not have prior achievement of GCSE maths in 2017/2018. The comparison group sample consists of settings and students belonging to institutions in AEA Category 4 areas that had at least one student meeting the eligibility criteria for BMP. Sources are NPD, ILR and BMP sample file information.

* The variables measuring prior attainment at KS2 in English and maths are coded from 0–6, where each score represents a level of attainment (that is, 'level 0', 'level 1', 'level 2', etc).

Resit pass rates and exams entered for

Table 15 and Table 16 report student-level descriptive statistics for variables that are related to the key outcomes for this trial. These are:

- the percentage of students achieving the required pass by November 2019—our primary outcome;
- the percentage of students attaining grade 3 in their previous maths level 2 attempt (compared to grades 1, 2, or U); this is informative of prior attainment of students in our sample, before BMP funding was allocated;
- the percentage of students who entered for a GCSE maths resit in 2018/2019 or November 2019; this includes all students who entered for a GCSE exam within that timeframe (regardless of whether they also entered for a Functional Skills Level 2 exam)—our secondary outcome;
- the percentage of students who entered for a Functional Skills Level 2 exam in 2018/2019 or November 2019; this includes all students who entered for a Functional Skills Level 2 exam within that timeframe (regardless of whether they also entered for a GCSE exam): this provides some additional contextual information in relation to our secondary outcome;
- the percentage of students who only entered for a Functional Skills Level 2 exam in 2018/2019 or November 2019 (that is, those who did not also enter for a GCSE); and
- the percentage of students who did not enter for either GCSE maths or Functional Skills Level 2 in 2018/2019 or November 2019.

The tables show average differences in each variable between the BMP group and comparison group, before any weighting or matching adjustment is done. Table 15 reports these differences where the BMP group is defined according to the setting-level indicator of BMP exposure (which is our main definition of treatment status). Table 16 reports the same differences for an alternative definition of BMP exposure, defined at the student level.

Note that this trial reports on student level 2 maths resit attainment in the 2018/2019 academic year and up to November 2019 as this corresponds to the time horizon of interest for the BMP trial. We did not request NPD data covering more recent exam seasons for our evaluation sample; however, we note that assessment scores have changed since then.[28]

The tables show that the percentage of students achieving the required pass is lower in the BMP group than in the comparison group. We also find that average attainment in the previous maths level 2 attempt (up to the end of 2017/2018) is slightly lower in the BMP group than comparison group. This is likely to reflect the fact that BMP was targeted towards settings in areas with higher levels of education disadvantage. If BMP has had a positive impact, these descriptives suggest that, at best, the effect of the programme can only have been to attenuate the disparity in attainment between students in these areas compared to their peers in other AEA category areas. The overall percentage of students in our sample who are achieving the required pass in their maths level 2 resit is low across the sample: at between 12–13% for each group. This figure is lower than reported national average pass rates for students who re-sat maths in 2018/2019.[29]

Across our sample, a substantial majority of students resit a maths GCSE at some point in 2018/2019 or November 2019 (at around 94–95% across groups). This is not surprising since our sample consists of students estimated to be eligible for BMP, meaning that they were expected to continue studying maths as a condition of funding. There does not appear to be any strong difference in the proportion of students re-sitting maths GCSE between the intervention and comparison groups (although, when BMP eligibility is defined at the student level we do observe a slightly higher percentage of students re-sitting the GCSE in the intervention group compared to the comparison group).

A very low proportion of students (around 2% across groups) took a maths Functional Skills Level 2 exam in 2018/2019 or November 2019. This is again fairly even across groups, although when BMP eligibility is defined at the student level we do find a slightly lower proportion of students re-sitting this exam in the BMP group compared to the non-BMP group.

[28] See https://explore-education-statistics.service.gov.uk/find-statistics/level-2-and-3-attainment-by-young-people-aged-19

[29] We have not come across publicly reported attainment figures that relate exactly to our sample and time horizon, but a range of similar data is published each year. For example, DfE figures show that among students aged 16, 17, or 18 at the start of the 2018/2019 academic year there were 198,727 GCSE entries, of which 18.8% received a pass (regardless of prior achievement in maths at level 2). For this same population, there were 8,801 Functional Skills Level 2 entries, of which 90.3% received a pass. See DfE A-level and other 16-18 results (revised): 2018/19 (Table 6b).

It turns out that in our sample of interest, all students who entered for a Functional Skills Level 2 exam also entered for a GCSE within the time horizon. That is, there are no students that entered for Functional Skills Level 2 only. This means that the students in our sample who did not take a GCSE exam are precisely those students that took no maths exam at all (around 4–6% of students, across groups).

This finding affects the interpretation of the results for Research Question 5, which assesses whether there is any impact of BMP on whether students entered for a GCSE exam. This research question is focused on students with a prior maths level 2 attainment of 2 or below, who were eligible to resit either a GCSE or Functional Skills Level 2 exam to meet the condition of funding. One rationale for this question was to understand whether, for this subsample, BMP induces students to be re-entered for the more ambitious GCSE exam rather than the Functional Skills Level 2 exam that they might otherwise have sat. These descriptive statistics already indicate that this question is not relevant for this cohort as there are no students who entered for Functional Skills Level 2 exams only across our time horizon. This means that Research Question 5 will, in effect, test for the impact of BMP on whether students re-entered a GCSE exam as opposed to no maths exam at all.

Across all variables presented in these tables we find that the two definitions of BMP eligibility give broadly similar results. The results shown in Table 15 and Table 16 correspond to students from AEA Category 4, 5, and 6 areas only. In Appendix I of the Additional Technical Appendices document we show these tables again, this time including AEA Category 3 areas. We find that the differences are larger in this case (that is, the comparison group outperforms the BMP group to a slightly greater extent). This is expected given that AEA Category 3 settings will, on average, be less disadvantaged than those in Categories 4, 5, and 6.

Table 15: Descriptive statistics relating to maths level 2 resit attainment and exams entered for (intervention group definition = setting-level BMP indicator)

Variable	Unit	Treatment group				Comparison group				Standardised mean difference
		n	Mean	Standard deviation	# missing	n	Mean	Standard deviation	# missing	
Percentage of students achieving the required pass by November 2019	Percentage	44,452	11.98	32.47	0	25,476	13.29	33.95	0	-0.040
Percentage of students attaining grade 3 in previous maths level 2 attempt (compared to grades 1, 2 or U)	Percentage	44,452	43.03	49.51	0	25,476	45.84	49.83	1353	-0.056
Percentage of students who entered for a GCSE maths resit in 2018/2019 or November 2019	Percentage	44,452	94.83	22.14	0	25,476	94.69	22.43	0	0.006
Percentage of students who entered for a Functional Skills Level 2 exam in 2018/2019 or November 2019	Percentage	44,452	2.04	14.12	0	25,476	1.90	13.64	0	0.010
Percentage of students who only entered for a Functional Skills Level 2 exam in 2018/2019 or November 2019, and did not also enter for a GCSE	Percentage	44,452	0.00	0.00	0	25,476	0.00	0.00	0	.
Percentage of students who did not enter for either a GCSE maths resit or a Functional Skills Level 2 exam in 2018/2019 or November 2019.	Percentage	44,452	5.17	22.14	0	25,476	5.31	22.42	0	-0.006

1. The ‘setting-level’ BMP indicator is drawn from BMP sample file information and denotes settings that were part of the BMP pilot. We consider all eligible students in these settings to be part of the BMP treatment group. The comparison group in this case corresponds to all students in AEA Category 4 areas who met the eligibility criteria for BMP.
2. Sources are NPD, ILR, and BMP sample file information.

Table 16: Descriptive statistics relating to maths level 2 resit attainment and exams entered for (intervention group definition = student-level BMP indicator)

BMP group definition	Unit	Treatment group				Comparison group				Standardised mean difference
		n	Mean	Standard deviation	# missing	n	Mean	Standard deviation	# missing	
Percentage of students achieving the required pass by November 2019	Percentage	41,916	12.13	32.65	0	28,012	12.95	33.57	0	-0.025
Percentage of students attaining Grade 3 in previous maths level 2 attempt (compared to Grades 1, 2, or U)	Percentage	41,916	43.37	49.56	0	28,012	45.08	49.76	0	-0.035

BMP group definition	Unit	Treatment group				Comparison group				Standardised mean difference
		n	Mean	Standard deviation	# missing	n	Mean	Standard deviation	# missing	
Percentage of students who entered for a GCSE maths resit in 2018/2019 or November 2019	Percentage	41,916	95.60	20.50	0	28,012	93.55	24.56	0	0.092
Percentage of students who entered for a Functional Skills Level 2 exam in 2018/2019 or November 2019	Percentage	41,916	1.31	11.39	0	28,012	2.99	17.03	0	-0.120
Percentage of students who entered for a Functional Skills Level 2 exam in 2018/2019 or November 2019, and did not also enter for a GCSE	Percentage	41,916	0.00	0.00	0	28,012	0.00	0	0	.
Percentage of students who did not enter for either a GCSE maths resit or a Functional Skills Level 2 exam in 2018/2019 or November 2019.	Percentage	41,916	4.40	20.50	0	28,012	6.45	24.56	0	-0.092

1. The ‘student-level’ BMP indicator considers students to be part of the treatment group if they are individually flagged as being a BMP beneficiary. The comparison group in this case corresponds to all students in AEA Category 4, 5, and 6 areas who met the eligibility criteria for BMP but are not individually-flagged as being a BMP beneficiary.
2. Sources are NPD, ILR and BMP sample file information.

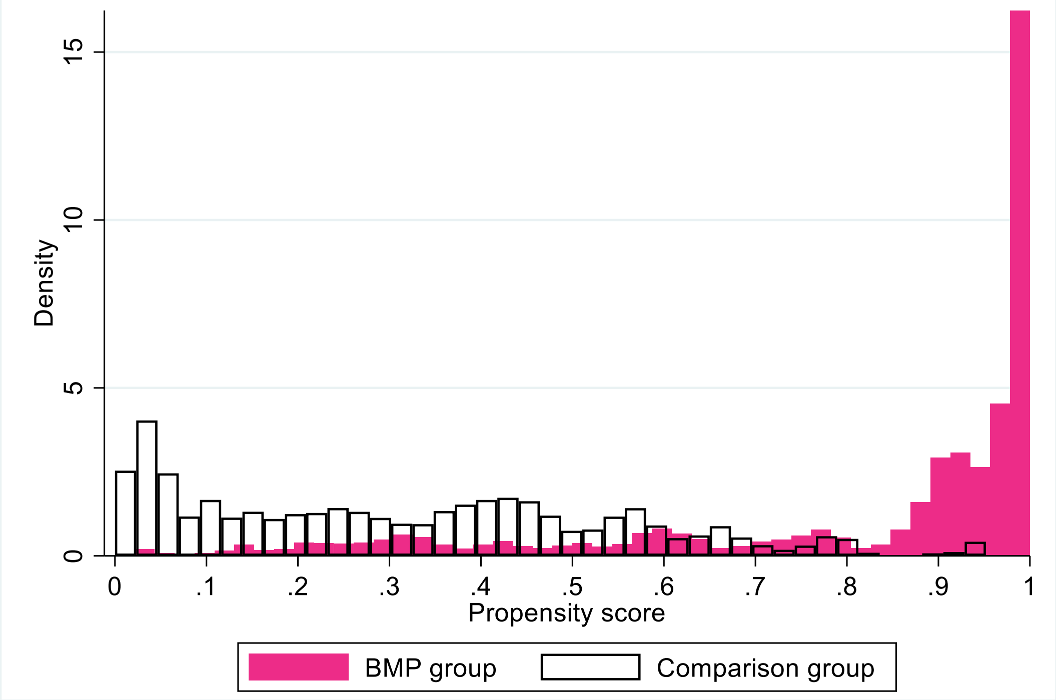
Primary analysis (RQ1)

Model specification—initial model (as per study plan)

Our primary analysis uses a doubly robust approach to estimate the impact of BMP on whether or not students achieve the required pass in maths level 2 resits. The first step in this analysis is to fit a propensity score model for BMP exposure using a range of covariates (described in Table 9). As a starting point, we fit this model for an intervention group consisting of all eligible students in BMP settings and a comparison group consisting of all eligible students in AEA Category 4 areas. To begin with, we estimate a propensity score model in respect of all BMP funding models (the pooled intervention group) and a non-BMP comparison group. Later, we will repeat this analysis for each BMP funding model separately.

Figure 3 plots the distribution of propensity scores from this initial model. We find that there is limited overlap in the propensity score distributions between the BMP and comparison groups: the former are skewed at the upper end of the distribution with few comparison group students receiving propensity scores in this range. This figure suggests that significant adjustment will be required to generate a balanced dataset from this starting point of limited overlap. It is likely that the differences found in the mean values of the institution-level and AEA covariates (Table 13) are what is driving these differences in the respective distributions.

Figure 3: Primary analysis Model 1—propensity score distributions



The figure shows the distribution of propensity scores in the raw student data. The sample covers 40,622 eligible students from 423 BMP settings and 23,083 students from 405 non-BMP settings in AEA Category 4 areas who would have been eligible for the funding if it were offered in their school. Propensity scores are based on a logit model of BMP exposure. Sources are NPD data, ILR data, and BMP sample file information.

Before proceeding to the outcomes analysis, we performed some further diagnostic checks to assess whether weighting the data according to inverse propensity score weights generated from these propensity scores would result in a sufficiently well-balanced dataset. In addition to the visual inspection of propensity score distributions, the study plan also proposes two further diagnostic checks. The first is a comparison of standardised mean differences in each covariate between the BMP and comparison groups, before and after weighting. Table 17 and below illustrate standardised mean differences before and after weighting. In the table, the cells shaded in red are those with absolute standardised mean differences greater than 0.1 and those in green, less than 0.1. This shows that, for this model, weighting does not lead to improvements in standardised mean differences across covariates. More than 50% of the covariates remain with an absolute difference of more than 0.1 after weighting, which violates the first diagnostic check set out in the study plan.

Figure 4 and Table 17 again illustrate that the greatest differences between the covariates are for the AEA indicators. Overall, these results indicate that the weighting model has not been effective: weighting does not reduce the number of covariates for which absolute standardised mean differences are below 0.1—in fact, the situation is slightly worse after weighting.

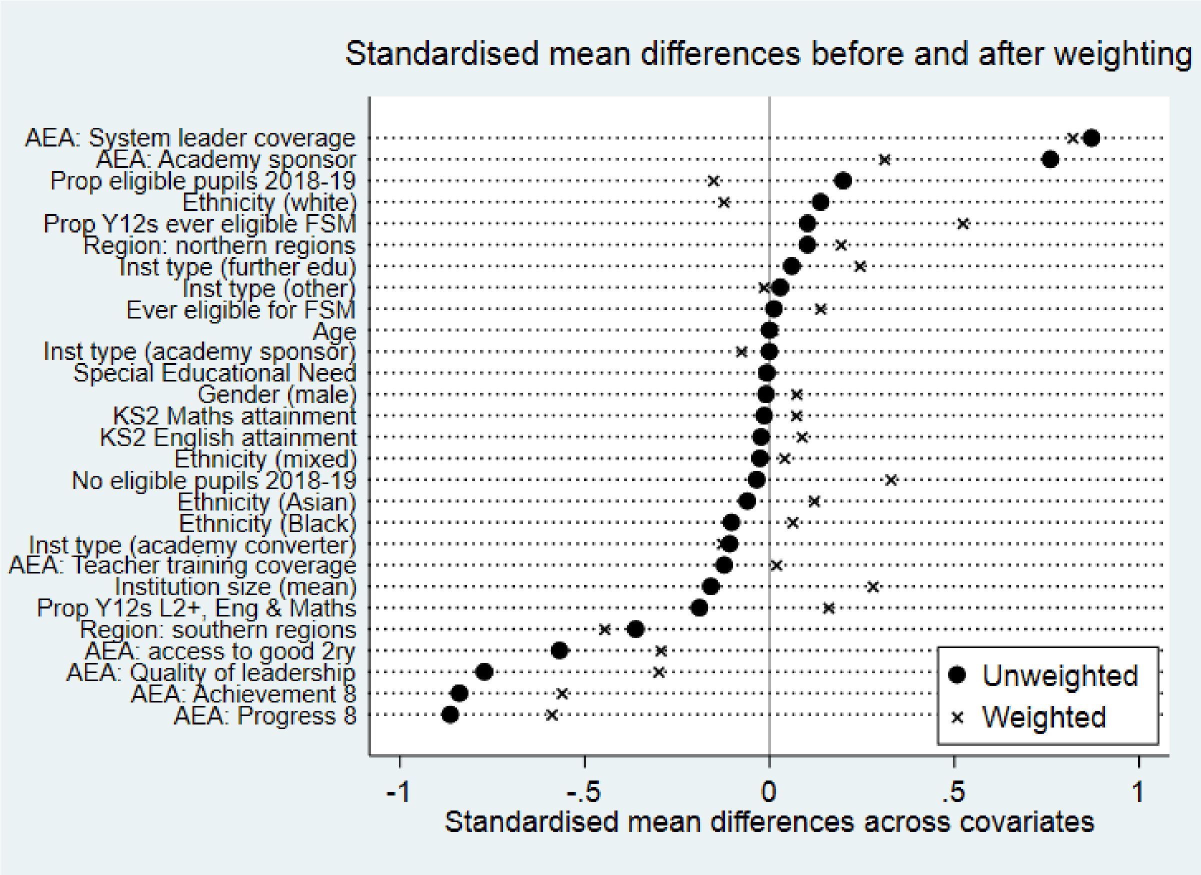
Table 17: Primary analysis Model 1—standardised mean differences before and after weighting

	Unweighted raw sample			**Weighted sample**		
Covariate	**Treated mean**	**Control mean**	**Standardised mean difference**	**Treated mean**	**Control mean**	**Standardised mean difference**
Institution size mean	6171.25	6828.48	-0.16	6150.17	5083.87	0.27
Number of eligible Year 12s in 2018/2019	495.40	507.21	-0.03	493.78	391.17	0.32
Proportion of eligible Year 12s in 2018/2019	45.28	42.40	0.20	45.18	47.64	-0.14
Institution type: proportion of settings that are academy converters	3.42	5.67	-0.11	3.52	6.16	-0.12
Institution type: proportion of settings belonging to the further education sector	85.84	83.69	0.06	85.66	76.50	0.24
Institution type: proportion of settings that are academy sponsors	2.96	2.97	0.00	3.00	4.43	-0.08
Institution type: proportion of settings with an ‘other’ institution type	7.33	6.59	0.03	7.40	7.74	-0.01
Region: Southern regions	17.46	32.91	-0.36	17.63	37.28	-0.45
Region: Northern regions	37.62	32.73	0.10	37.92	28.85	0.19
Proportion of Year 12s ever eligible for FSM	31.61	30.51	0.10	31.55	25.79	0.52
Proportion of Year 12s who achieved a Level 2 or higher in maths and English at KS4 prior to enrolment	37.71	41.14	-0.19	37.81	34.98	0.15
AEA Access to a good secondary school index	57.88	66.58	-0.57	57.76	61.99	-0.29
AEA Achievement 8 indicator	46.48	48.25	-0.84	46.52	47.69	-0.57
AEA Progress 8 indicator	-0.15	-0.03	-0.86	-0.15	-0.07	-0.59
AEA System leader coverage indicator	2566.87	1658.01	0.87	2566.96	1685.58	0.82
AEA Initial teacher training provider coverage index	40.81	42.97	-0.12	40.93	40.60	0.02
AEA Quality of leadership indicator	72.46	81.21	-0.77	72.38	76.34	-0.30
AEA Academy sponsor coverage	1545.23	737.80	0.76	1553.87	1187.76	0.31
Age	16.01	16.01	0.00	16.01	16.00	0.01
Gender	50.82	51.30	-0.01	50.57	47.01	0.07
Ethnicity: White	79.76	74.33	0.13	81.82	86.45	-0.13
Ethnicity: Mixed/ multiple	4.51	5.08	-0.03	4.42	3.63	0.04
Ethnicity: Asian	10.77	12.71	-0.06	9.90	6.57	0.12

	Unweighted raw sample			Weighted sample		
Covariate	**Treated mean**	**Control mean**	**Standardised mean difference**	**Treated mean**	**Control mean**	**Standardised mean difference**
Ethnicity: Black	3.97	6.30	-0.11	3.24	2.23	0.06
Eligibility for free school meals	38.29	37.71	0.01	39.04	32.58	0.14
Special Education Need status (SEN)	27.10	27.42	-0.01	27.96	28.68	-0.02
Prior attainment at KS2 in English	3.66	3.68	-0.02	3.66	3.59	0.08
Prior attainment at KS2 in maths	3.55	3.56	-0.01	3.55	3.50	0.07

1. This table reports standardised mean differences before and after weighting using student-level, institution-level, and AEA covariates. Propensity scores are derived using a logit model. Cells shaded in red are those with absolute standardised mean differences > 0.1 and those in green are absolute standardised mean difference differences < 0.1.
2. The treatment group contains students from BMP schools in AEA Categories 5 and 6; the comparison group comprises students who would have been eligible for BMP had it been offered in their school, from AEA Category 4. The sample sizes are 40,622 students from 423 BMP settings and 23,083 students from 405 non-BMP settings.
3. Three covariates are omitted as they are perfect linear combinations of other included covariates (and the model includes a constant term). These are Region: East and Midlands; Institution type: 'missing'; and Ethnicity: 'other'.
4. The sources are NPD data, ILR data, and BMP sample file information.

Figure 4: Primary analysis Model 1—standardised mean differences before and after weighting.



1. This figure reports standardised mean differences in covariates before and after weighting. Propensity scores are derived using a logit model.
2. The treatment group contains students from BMP schools in AEA Categories 5 and 6; the comparison group comprises students who would have been eligible for BMP if it was offered in their school from AEA Category 4. The sample sizes are 40,622 students from 423 BMP settings and 23,083 students from 405 non-BMP settings.
3. Three covariates are omitted as they are perfect linear combinations of other included covariates (and the model includes a constant term). These are Region: East and Midlands; Institution type: missing; and Ethnicity: other.
4. The sources are NPD data, ILR data, and BMP sample file information.

The second diagnostic check is to check the pseudo-R^2 of the propensity score model in the raw data, compared to the weighted data. 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. We adapt a procedure recommended in Sianesi (2004), which discusses diagnostic checks for propensity score matching. Sianesi recommends that for studies that use matching methods, propensity scores should be re-estimated in the matched sample and used to recalculate the pseudo-R^2. The expectation is that if matching has worked well there should be few differences in the covariate distribution between the treatment and comparison groups, and the pseudo-R^2 will be low. We can apply this same logic to the present case of propensity score weighting and compare the pseudo-R^2 before and after weighting. In doing so, we observe a modest decrease from 0.46 before weighting to 0.30 afterward. This indicates that weighting has helped reduce the overall extent of differences between the intervention and comparison groups in the weighted data, but that covariates are still to some extent predictive of intervention allocation (rather than being entirely orthogonal, as we would hope for a very well-balanced dataset). We conclude that this initial model is ineffective and not suitable for our impact estimation.

Model specification—adjusting the model

We adjusted the model to improve its performance according to the procedures outlined in the study plan. We first expanded the comparison group to include settings from AEA Category 3. We then applied different trimming options to improve balance in the sample. Our objective in selecting the final model was to find a specification that would satisfy the diagnostic tests outlined in the study plan while minimising any trimming to the treatment group. In total, we examined diagnostics for the following models:

- Model 2: expanding the comparison group to include AEA Category 3;
- Model 3: expanding to AEA Category 3, plus trimming observations with propensity scores below 0.1 or above 0.9;[30]
- Model 4: expanding to AEA Category 3, plus trimming observations with propensity scores above 0.97;
- Model 5: trimming observations with propensity scores above 0.97; and
- Model 6: trimming observations with propensity scores below 0.03 or above 0.97.

A summary of the key diagnostic checks from each of the models we tested is shown in Table 18. Among these alternatives, the model that performs best according to our diagnostic checks is *Model 5.* This is what we select for the final estimation model. Our rationale is as follows:

Ideally, model specifications that avoid trimming from the intervention group are preferred. We therefore start by examining whether simply expanding to include AEA Category 3 in the comparison group will be sufficient to improve the model, without needing to carry out any trimming. The rationale for attempting this is that increasing the size of the comparison group may increase the number of available observations that are similar to the intervention group, thereby improving overall model balance. However, this turns out not to be the case and we find that Model 2 does not perform much better than the first model (in fact, the pseudo-R^2 after weighting is less favourable). We also find that Model 4 performs worse than Model 5, where the only difference is the inclusion of AEA Category 3 in Model 4. We therefore conclude that the final model should in fact include a comparison group from AEA Category 4 only. We speculate that the reason why expanding to AEA Category 3 is unsuccessful is that there are meaningful differences in setting and area-level characteristics between the AEA categories, and those in Category 3 are indeed more different to those in Category 4 from the intervention group. This means that including these observations causes propensity scores in the intervention group to be shifted further to the right (that is, more markedly different from the comparison group). This leads to relatively more difficulty in obtaining covariate balance. We therefore rule out models that include AEA Category 3.

Secondly, we observe that dropping observations from the lower end of the propensity score distribution (which are more likely to be those in the comparison group) is much less effective at improving model balance than dropping those at the top end of the distribution (which are more likely to be intervention group observations). We can see this from comparing Models 5 and 6. The two models are almost indistinguishable in terms of covariate balance after weighting, so the loss of additional observations at the bottom of the distribution does not lead to any gains.

[30] This choice of trimming is informed by Crump et al. (2009). This paper recommends trimming observations with propensity scores above 0.9 and below 0.1 as an effective 'rule of thumb' for employing propensity score estimation methods.

We are therefore left with only the option of dropping observations from the top end of the propensity score distribution to achieve improvements in the diagnostic results. Comparing Models 3 and 5, we find that Model 3 performs much better in terms of covariate balance, however, this model carries a serious trade-off in that it requires trimming a large proportion of BMP group schools from the sample. We consider this to be an unacceptably large loss of the intervention group sample. We instead select Model 5 as a compromise between dropping enough observations with extremely high propensity scores to improve diagnostic results while preserving as much of the intervention group as possible. This model meets the diagnostic tests laid out in the study plan. While it does not sacrifice as much of the intervention group as Model 3, this model still drops around 34% of settings in the intervention group.

We discounted the other option mentioned in the study plan—of updating the selection of covariates to improve balance. Changing the choice of covariates only makes sense if there is evidence to suggest that the original selection was inappropriate. The presence of poor model diagnostics does not in itself constitute evidence that the covariates chosen were the wrong ones, according to the criteria set out in the Statistical Analysis section. We instead believe that the difficulty in developing a well-balanced comparison group is due to the fact that the BMP was targeted in specific AEA areas that are meaningfully different from the non-BMP areas from which the comparison group is drawn. In other words, pronounced differences in area-level characteristics between the BMP and non-BMP group are to be expected for this intervention design and are not simply a result of an inappropriate selection of covariates.

Table 18: Primary analysis—diagnostic results for alternative model specifications

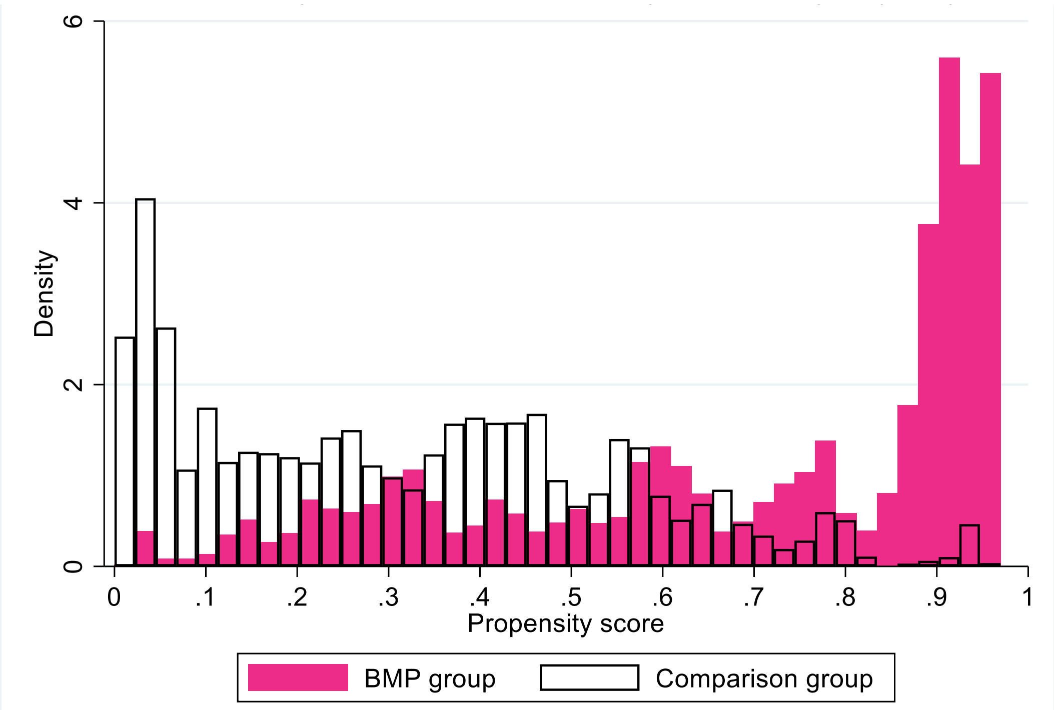
Model			# dropped observations		Remaining observations in estimation sample				Prop. covariates with absolute standardised mean differences < 0.1 after weighting	Pseudo-R^2
#	Trimming	Comparison group	Students	Settings	Intervention group students	Intervention group settings	Comparison group students	Comparison group settings		
1	No trimming	AEA Category 4 only	6,223	18	40,622	423	23,083	405	35.71	Before weighting: 0.461 After weighting: 0.298
2	No trimming	AEA Categories 3 and 4	7,800	29	40,622	423	40,591	779	32.14	Before weighting: 0.533 After weighting: 0.308
3	Observations trimmed with propensity scores outside [0.1, 0.9]	AEA Categories 3 and 4	51,683	681	15,745	196	21,585	354	75.00	Before weighting: 0.533 After weighting: 0.046
4	Observations trimmed with propensity scores > 0.97	AEA Categories 3 and 4	24,475	197	23,953	261	40,585	773	60.71	Before weighting: 0.533 After weighting: 0.136
5	Observations trimmed with propensity scores > 0.97	AEA Category 4 only	22,618	167	24,234	280	23,076	399	64.29	Before weighting: 0.461 After weighting: 0.133
6	Observations trimmed with propensity scores outside [0.03, 0.97]	AEA Category 4 only	24,277	216	24,216	278	21,435	352	64.29	Before weighting: 0.461 After weighting: 0.134

1. This table reports standardised mean differences before and after weighting using student-level, institution-level, and AEA covariates. Propensity scores are derived using a logit model.
2. The treatment group contains students from BMP schools in AEA Categories 5 and 6; the comparison group comprises students who would have been eligible for BMP had it been offered in their school, from AEA Categories 3 and 4 (depending on the model reported).
3. Three covariates are omitted as they are perfect linear combinations of other included covariates (and the model includes a constant term): these are Region: East and Midlands; Institution type: 'missing'; and Ethnicity: 'other'.
4. The sources are NPD data, ILR data, and BMP sample file information.
5. In Models 1 and 2, although there is no explicit trimming of observations with extreme propensity score values, some observations are still dropped from the estimation sample if they have missing covariate values.

Final model selected—diagnostic results

The final distribution of propensity scores from our selected model is shown in Figure 5. We find that there is more alignment in the distributions than in the initial model without trimming—although it is still the case that the distribution is skewed towards 1 for BMP observations and toward 0 for the comparison group.

Figure 5: Primary analysis final model—propensity score distributions



The figure shows the distribution of propensity scores after trimming observations with propensity scores > 0.97. The sample covers 24,234 students from 280 BMP settings and 23,076 students from 399 non-BMP settings who would have been eligible for the funding had it been offered in their school. The propensity score model is based on a logit model of BMP exposure using all student-level and institution-level covariates. Sources are NPD data, ILR data, and BMP sample file information.

Covariate balance before and after weighting is shown in Table 19 and visualised in Figure 6. Compared to the initial model with no trimming, there is more overlap in the propensity score distributions between the intervention and comparison groups, and covariate balance is also improved after weighting. Nevertheless, the weighting approach has not been entirely successful in overcoming differences between the treatment and comparison groups. The pseudo-R^2 in the weighted data is 0.13, which implies that we cannot quite consider BMP exposure to be 'as if' randomly assigned after weighting (as we would hope). However, this is an improvement over the initial model with no trimming.

Table 19: Primary analysis final model—standardised mean differences before and after weighting

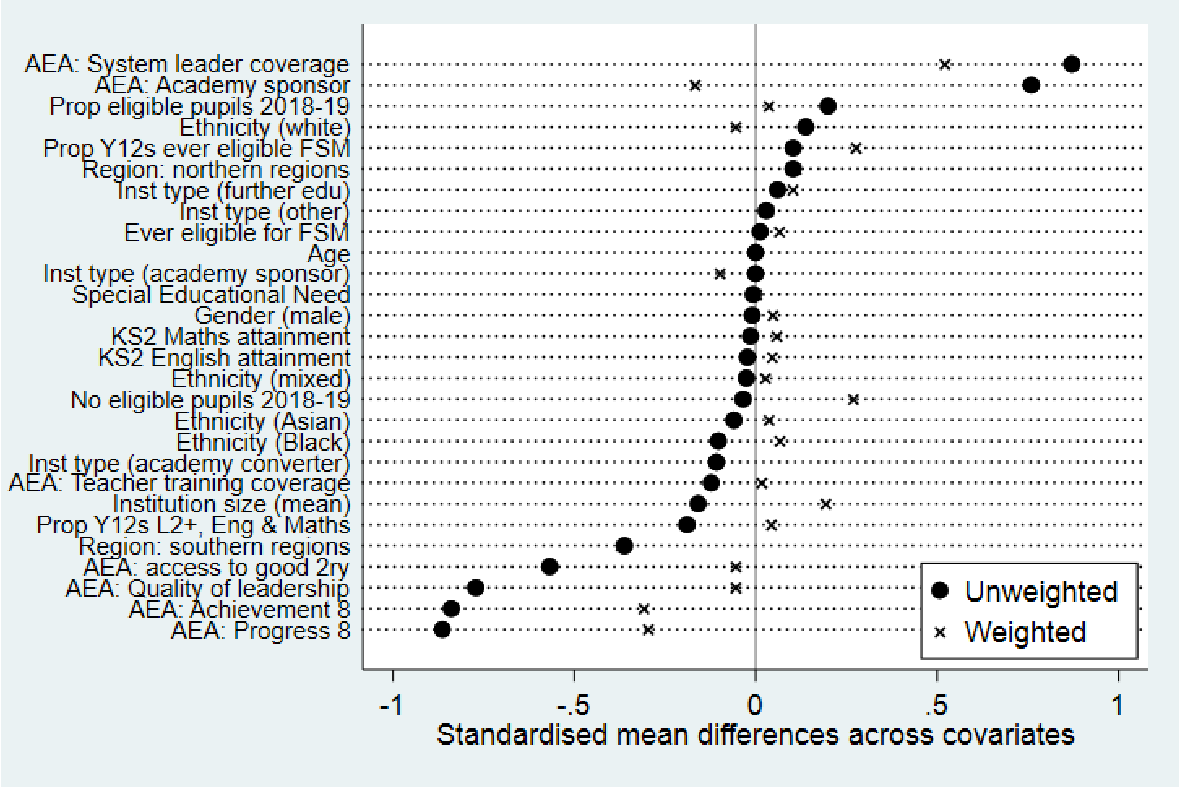
	Unweighted raw sample			Weighted sample		
Covariate	**Treated mean**	**Control mean**	**Standardised mean difference**	**Control mean**	**Treated mean**	**Standardised mean difference**
Institution size mean	6171.25	6828.48	-0.16	6244.96	5534.76	0.19
Number of eligible Year 12s in 2018/2019	495.40	507.21	-0.03	510.55	428.25	0.26
Proportion of eligible Year 12s in 2018/2019	45.28	42.40	0.20	44.58	44.04	0.04
Institution type: proportion of settings that are academy converters	3.42	5.67	-0.11	4.08	6.51	-0.11

	Unweighted raw sample			Weighted sample		
Covariate	**Treated mean**	**Control mean**	**Standardised mean difference**	**Control mean**	**Treated mean**	**Standardised mean difference**
Institution type: proportion of settings belonging to the further education sector	85.84	83.69	0.06	85.53	81.94	0.10
Institution type: proportion of settings that are academy sponsors	2.96	2.97	0.00	2.91	4.80	-0.10
Institution type: proportion of settings with an ‘other’ institution type	7.33	6.59	0.03	7.22	6.33	0.04
Region: Southern regions	17.46	32.91	-0.36	21.71	38.57	-0.37
Region: Northern regions	37.62	32.73	0.10	36.37	31.15	0.11
Proportion of Year 12s ever eligible for FSM	31.61	30.51	0.10	30.64	27.89	0.28
Proportion of Year 12s who achieved a Level 2 or higher in maths and English at KS4 prior to enrolment	37.71	41.14	-0.19	38.33	37.63	0.04
AEA access to a good secondary school index	57.88	66.58	-0.57	61.80	62.58	-0.05
AEA Achievement 8 indicator	46.48	48.25	-0.84	47.33	47.87	-0.31
AEA Progress 8 indicator	-0.15	-0.03	-0.86	-0.09	-0.06	-0.30
AEA System leader coverage indicator	2566.87	1658.01	0.87	2216.15	1687.83	0.52
AEA Initial teacher training provider coverage index	40.81	42.97	-0.12	41.22	40.93	0.02
AEA Quality of leadership indicator	72.46	81.21	-0.77	75.87	76.52	-0.05
AEA Academy sponsor coverage	1545.23	737.80	0.76	1016.81	1160.35	-0.17
Age	16.01	16.01	0.00	16.01	16.01	0.00
Gender	50.82	51.30	-0.01	51.33	49.05	0.05
Ethnicity: White	79.76	74.33	0.13	83.00	85.19	-0.06
Ethnicity: Mixed/ multiple	4.51	5.08	-0.03	4.55	4.01	0.03
Ethnicity: Asian	10.77	12.71	-0.06	8.12	7.13	0.04
Ethnicity: Black	3.97	6.30	-0.11	3.64	2.50	0.07
Eligibility for free school meals	38.29	37.71	0.01	38.30	35.22	0.06
Special Education Need status (SEN)	27.10	27.42	-0.01	28.70	28.69	0.00
Prior attainment at KS2 in English	3.66	3.68	-0.02	3.66	3.63	0.04
Prior attainment at KS2 in maths	3.55	3.56	-0.01	3.55	3.51	0.06

1. This table reports standardised mean differences before and after weighting using student-level, institution-level, and AEA covariates. Propensity scores are derived using a logit model. Cells shaded in red are those with absolute standardised mean differences > 0.1 and those in green differences of < 0.1.
2. The treatment group comprises students from BMP schools and the comparison group consists of students from non-BMP schools in AEA Category 4 who would have been eligible for BMP had it been offered in their school. The unweighted raw sample describes this entire sample; the weighted sample refers to the estimation sample (including trimming of all observations with propensity scores greater than 0.97). Overall, the weighted sample contains 24,234 students from 280 BMP settings and 23,076 students from 399 non-BMP settings.

3. Three covariates are omitted as they are perfect linear combinations of other included covariates (and the model includes a constant term). These are Region: East and Midlands; Institution type: 'missing'; and Ethnicity: 'other'.
4. The sources are NPD data, ILR data, BMP sample file information.

Figure 6: Primary analysis final model—standardised mean differences before and after weighting



1. This figure reports standardised mean differences in covariates before and after weighting. Propensity scores are derived using a logit model.
2. The treatment group comprises students from BMP schools and the comparison group consists of students from non-BMP schools in AEA Category 4 who would have been eligible for BMP had it been offered in their school. The unweighted raw sample describes this entire sample, while the weighted sample refers to the estimation sample (including trimming of all observations with propensity scores greater than 0.97). Overall, the weighted sample contains 24,234 students from 280 BMP settings and 23,076 students from 399 non-BMP settings.
3. Three covariates are omitted as they are perfect linear combinations of other included covariates (and the model includes a constant term): these are Region: East and Midlands; Institution type: 'missing'; and Ethnicity: 'other'.
4. The sources are NPD data, ILR data, and BMP sample file information.

Primary analysis results

Table 20 reports the impact estimates produced by our chosen model. These findings capture the difference in the probability of achieving the required pass in maths by November 2019, for eligible students in BMP schools compared to those in comparison group schools. Table 21 reports the same results, expressed as risk ratios.

The findings reveal no impact of the BMP funding on supporting students to achieve the required pass in maths level 2 resits. We observe that the difference in the probability of passing is close to zero. Similarly, the risk ratio is very close to one, indicating that there is no difference in the relative likelihood of achieving the required maths pass for students in the BMP intervention group.

Table 20: Primary analysis final model—impact estimates

BMP treatment indicator	ATET (difference in probability of passing maths)	Standard error	z	P > |z|	95% confidence interval
Setting level	0.00092	0.012	0.08	0.954	(-0.022, 0.024)

1. This table reports impact estimates from doubly robust estimation, performed using `teffects ipwra` in Stata with a logit specification for the treatment and outcomes model.

2. BMP exposure is defined at the setting level using an indicator drawn from the BMP sample file maintained by the evaluation team to monitor institutions that had signed an MOU and not withdrawn from the trial. This specification includes 24,234 students from 280 BMP settings and 23,076 students from 399 non-BMP settings.
3. Standard errors for the setting-level BMP exposure indicator are clustered at the institution level.
4. The sources are NPD data, ILR data, and BMP sample file information.

Table 21: Primary analysis final model—risk ratios

BMP treatment indicator	Risk ratio	95% confidence interval
Setting level	1.008	(0.834, 1.214)

1. This table reports impact estimates in the form of risk ratios, from doubly robust estimation, performed using `teffects ipwra` in Stata with a logit specification for the treatment and outcomes model.
2. Other notes as per the above table.

Separate doubly robust models

In line with the study plan, we also investigate this research question by estimating separate doubly robust models for each of the three BMP treatment arms. We use the same `teffects ipwra` Stata command as the main primary analysis for the pooled BMP intervention group above. This involves re-estimating propensity scores in respect of each treatment arm subsample and the comparison group. For consistency with the pooled analysis, we again trim observations with propensity scores greater than 0.97 and define the comparison group using settings from AEA Category 4 areas. We also apply a Bonferroni correction to the calculation of confidence intervals, to adjust for the fact that we are estimating three separate coefficients. This implies a significance level of α = 0.017 (which is 0.05/3).

The estimation models, when broken down by treatment arm, suffer from some of the same difficulties in estimation as the main primary analysis and are difficult to fit. The model for group A proceeds as expected, but the models of group B and C cannot be fit (that is, the doubly robust model refuses to run). We can redress this issue for the B and C models through a slight coarsening of the institution size covariate, to group together the ‘institution size – other’ and ‘institution size – missing’ categories into one, rather than entering them as separate covariates as in the main analysis. This slight coarsening of one covariate lessens the demands of fitting the estimation model using this reduced sample and allows coefficients to be estimated. Nonetheless, the fact that we encounter difficulties in achieving model fit over this reduced sample is symptomatic of broader challenges with achieving a good comparison group for the BMP intervention sample.

Table 22 shows the results of this analysis, where the ATET is expressed in units of the difference BMP makes to the probability of passing maths. This table does not show evidence that any of the BMP funding models lead to a detectable impact on the likelihood of eligible students passing maths level 2 resits by November 2019. The coefficients are small and statistically indistinguishable from zero. This is reflected in p-values that are greater than 0.05 and the fact that the Bonferroni-adjusted confidence interval crosses zero.

Table 23 presents the same results as risk ratios. In this case, we find that risk ratios are close to one, which indicates no difference in the relative likelihood of passing maths for students belonging to that treatment arm. The confidence intervals associated with the risk ratios cross over one, which means that we cannot statistically distinguish whether the BMP is associated with a positive or negative change in relative risk.

Looking at the point estimates for this analysis, we do find that students belonging to funding Model B appear to have a slightly lower likelihood of passing maths than those in the comparison group. However, as discussed above, this result is not measured precisely and we therefore treat it with caution.

Table 22: Analysis of relative effectiveness (separate doubly robust models)—impact estimates

BMP funding model	ATET (difference in probability of passing maths)	Standard error	z	P > |z|	95% confidence interval
A	0.003	0.024	0.12	0.906	(-0.054, 0.060)
B	-0.025	0.026	-0.96	0.335	(-0.086, 0.037)

C	0.009	0.017	0.52	0.602	(-0.032, 0.049)

1. This table reports impact estimates from doubly robust estimation, performed separately for each BMP treatment arm using `teffects ipwra` in Stata with a logit specification for the treatment and outcomes model.
2. BMP eligibility is defined at the setting level using an indicator drawn from the BMP sample file maintained by the evaluation team to monitor institutions that had signed an MOU and not withdrawn from the trial.
3. Sample sizes are as follows:
    a. Model A: 8,166 BMP students belonging to 80 settings in AEA Category 5 and 6 areas, 23,080 non-BMP students belonging to 402 settings in AEA Category 4 areas.
    b. Model B: 7,731 BMP students belonging to 106 settings in AEA Category 5 and 6 areas, 23,079 non-BMP students belonging to 401 settings in AEA Category 4 areas.
    c. Model C: 8,751 BMP students belonging to 105 settings in AEA Category 5 and 6 areas, 23,077 non-BMP students belonging to 400 settings in AEA Category 4 areas.
4. Standard errors are clustered at the institution-level.
5. We have applied a Bonferroni correction to account for the fact that this analysis involves estimating three coefficients. The significance level used to define the confidence interval is α = 0.017 (which is 0.05/3).
6. The sources are NPD data, ILR data, and BMP sample file information.

Table 23: Analysis of relative effectiveness (separate doubly robust models)—risk ratios

BMP funding model	Risk ratio	95% confidence interval
A	1.025	(0.615, 1.709)
B	0.833	(0.542, 1.279)
C	1.071	(0.781, 1.468)

This table reports risk ratios from doubly robust estimation, performed separately for each BMP treatment arm using `teffects ipwra` in Stata with a logit specification for the treatment and outcomes model. Other notes as per the table above.

Further analysis in technical appendices

The additional technical appendices document contains further analysis of the impact of BMP on the primary outcome.

In Appendix J we present the results from a robustness check carried out using a two-stage matching approach. Unfortunately, the two-stage matching model does not work well and is unable to produce a satisfactory comparison group. The results from the robustness check suggest a negative and significant impact of BMP. However, given the difficulties faced by this model in establishing a comparison group, we attribute this apparently negative finding to the fact that the BMP funding was targeted to settings with relatively lower levels of educational attainment. In other words, the statistical adjustment performed by the two-stage matching model is not able to fully overcome the systematic differences that exist in reality between BMP settings and non-BMP settings. Please refer to Appendix J for full results and further discussion.

In Appendix K of the Additional Technical Appendices document we present additional exploratory impact analysis of the primary outcome. This uses an alternative definition of BMP exposure that is defined at the student level (discussed above in the Participant Selection sub-section). The results from this additional exploratory analysis are somewhat contrasting and reveal a modest positive impact of the BMP on the probability of achieving the required pass. However, we do not consider this exploratory analysis to constitute strong evidence. First, the fact that there is such close alignment between the two definitions of BMP allocation means that the apparent sensitivity of the results to changing the definition appears to be spurious. Secondly, the BMP exposure variable used for this additional analysis does not conform to the level at which BMP funding was delivered in practice. BMP was fundamentally a setting-level intervention and it is therefore most appropriate to analyse it at this level (as our main analysis does). Finally, this positive result on the primary outcome is at odds with other results across the impact evaluation and IPE components, described below.

Analysis of relative effectiveness (RQ2)

Research Question 2 explores whether there are differences in the relative impact of the three BMP funding models. To address this question, we first carry out a single multilevel logit regression (weighted by the inverse propensity score weights derived from the primary analysis) that contains a term for each of the three treatment arms. Second, we directly exploit the randomised assignment of the different BMP funding models by estimating a further multilevel logit regression over the sample of BMP students only. This model contains a term for belonging to funding Models B and C, which are compared to funding Model A as the reference category.

Multilevel logistic regression

Table 24 reports results from a multilevel logit model that includes a dummy variable for each treatment arm. This model is weighted using the same propensity score weights as the primary analysis (and so does not involve re-estimating propensity scores for each treatment arm subsample). We again apply a Bonferroni correction when calculating confidence intervals to adjust for the fact that this analysis involves testing three separate hypotheses rather than one. This means that the confidence intervals reported below are based on a significance level of α = 0.017 (which is 0.05/3). In Table 24 coefficients are expressed in terms of odds ratios. Table 25 reports the same results expressed as risk ratios. In both cases, we do not find evidence of a detectable impact of any of the BMP funding models. In all cases the p-values associated with each coefficient are greater than 0.05. This is reflected in confidence intervals that cross over one for all the coefficients, which indicates no statistically detectable impact for both odds ratios and risk ratios.

Table 24: Analysis of relative effectiveness (multilevel logistic regression)—impact estimate

BMP funding model	Odds ratio	Standard error	z	P > |z|	95% confidence interval
A	0.82	0.110	-1.46	0.14	[0.598, 1.132]
B	0.94	0.107	-0.58	0.57	[0.712, 1.231]
C	1.12	0.139	0.92	0.36	[0.833, 1.507]

1. This table reports impact estimates from a multilevel logistic regression that includes a separate dummy variable for each BMP treatment arm.
2. BMP eligibility is defined at the setting level using an indicator drawn from the BMP sample file maintained by the evaluation team to monitor institutions that had signed an MOU and not withdrawn from the trial.
3. This includes 24,234 BMP students from 280 settings in AEA Category 5 and 6 areas and 23,076 non-BMP students from 399 settings in AEA Category 4 areas.
4. Standard errors for the setting-level BMP exposure indicator are clustered at the institution-level.
5. We have applied a Bonferroni correction to account for the fact that this analysis involves estimating three coefficients. The significance level used to define the confidence interval is α = 0.017 (which is 0.05/3).
6. The sources are NPD data, ILR data, and BMP sample file information.

Table 25: Analysis of relative effectiveness (multilevel logistic regression): risk ratios

BMP funding model	Risk ratio	95% confidence interval
A	0.86	(0.677, 1.101)
B	0.95	(0.775, 1.169)
C	1.09	(0.874, 1.354)

This table reports impact estimates from a multilevel logistic regression, with students clustered within settings, which includes a separate dummy variable for each BMP treatment arm. Other notes as per the table above.

Experimental analysis

Secondly, we focus on the sample of BMP institutions only and carry out a simple regression analysis to examine the relative effectiveness of belonging to funding Models B and C, compared to A. This analysis directly exploits the fact that allocation to BMP funding arms was randomised and does not use any quasi-experimental adjustment to develop a non-BMP comparison group. Instead, we focus on the BMP group only and run a single multilevel regression model that includes a term for belonging to funding Model B or C. The model also includes the same covariates as the primary analysis in order to improve precision. We again use a Bonferroni correction, this time using a significance level of α = 0.025 (which is 0.05/2), since we are estimating two coefficients.

The results from this analysis are shown in Table 26 as odds-ratio units and in Table 27 as risk ratios. Again, the findings from this analysis show no discernible impacts as coefficients are close to one and associated with wide confidence intervals. Looking only at the point estimates, we see that the odds and risk ratios associated with Models B and C are higher than one. This indicates a slight tendency for students to have a higher likelihood of achieving the required pass if they belonged to one of these two models, rather than Model A. Again though, the lack of precision in these estimates means that we cannot interpret the direction and magnitude of these relationships with confidence.

Table 26: Analysis of relative effectiveness (experimental analysis)—impact estimates (relative to funding Model A)

BMP funding model	Odds ratio	Standard error	z	P > |z|	95% confidence interval
B	1.123	0.117	1.11	0.269	[0.888, 1.419]
C	1.150	0.121	1.32	0.187	[0.907, 1.457]

1. This table reports impact estimates from a multilevel logistic regression carried out over the BMP sample only, that includes a separate dummy variable for each BMP treatment arm. The reference category is funding Model A.
2. BMP eligibility is defined at the setting level using an indicator drawn from the BMP sample file maintained by the evaluation team to monitor institutions that had signed an MOU and not withdrawn from the trial.
3. This includes 41,622 BMP students from 423 settings in AEA Category 5 and 6 areas (comprising 14,119 students from 136 Model A settings, 12,821 students from 141 Model B settings, and 13,682 students from 146 Model C settings.
4. Standard errors for the setting-level BMP exposure indicator are clustered at the institution level.
5. We have applied a Bonferroni correction to account for the fact that this analysis involves estimating two coefficients. The significance level used to define the confidence interval is α = 0.025 (which is 0.05/2).
6. The sources are NPD data, ILR data, and BMP sample file information.

Table 27: Analysis of relative effectiveness (experimental analysis): risk ratios (relative to funding Model A)

BMP funding model	Risk ratio	95% confidence interval
B	1.09	[0.912, 1.288]
C	1.11	[0.931, 1.322]

This table reports impact estimates from a multilevel logistic regression carried out over the BMP sample only, that includes a separate dummy variable for each BMP treatment arm. The reference category is funding Model A. Other notes as per the table above.

Overall, the analyses carried out to explore relative effectiveness do not provide any evidence of an impact associated with any of the funding models. In each analysis, coefficients were not statistically distinguishable from the implicit null hypothesis of no impact.

Subgroup analysis (RQ3)

Research Question 3 examines whether there is a differential impact of BMP according to three subgroups of interest: (1) students who were eligible for free school meals at any time in the five years before entering post-16 education, (2) the number of students per setting eligible to receive funding through BMP, and (3) prior attainment in maths level 2 exams. As outlined in the study plan, we carry out this analysis by first combining all BMP treatment arms into one pooled intervention group, and then for each intervention arm separately.

Pooled treatment group

Table 28 reports the results from multilevel logistic regressions, where all BMP funding arms were pooled into a single intervention group. Each regression was weighted using the weights produced in the primary analysis and includes an interaction term for BMP treatment assignment and the subgroup of interest.

Table 28 reports the findings from this analysis, expressed as odds ratios. These results do not provide evidence of any differential impact of BMP funding according to these subgroups. The odds ratios for the FSM and setting size subgroups are very close to one, indicating no impact. For the subgroup based on prior attainment, the point estimate on the odds ratio is less than one. This would be consistent with students who received a grade 3 in their previous attempt at maths level 2 exams being less likely to experience an impact on resit results through BMP than their peers with prior attainment of 2, 1, or U. This could suggest that BMP funding is particularly used to support students with lower attainment to begin with. However, this is weak evidence since the confidence interval around this estimate is particularly wide and extends below and above one.

Table 28: Subgroup analysis (multilevel logistic regression)—impact estimates

Subgroup	Odds ratio	Standard error	z	P > |z|	95% confidence interval	n
Student ever eligible for free school meals	0.99	0.09	-0.10	0.92	[0.823, 1.192]	47,310
Number of BMP eligible students per setting	1.00	0.00	-0.52	0.61	[0.999, 1.000]	47,310
Highest prior maths level 2 attainment of grade 3	0.75	0.21	-1.02	0.31	[0.435, 1.300]	47,310

1. This table reports impact estimates from separate multilevel logistic regression models with an interaction term for the BMP exposure indicator and each subgroup of interest, and a random intercept for each institution.
2. BMP eligibility is defined at the setting level using an indicator drawn from the BMP sample file maintained by the evaluation team to monitor institutions that had signed an MOU and not withdrawn from the trial.
3. The regression model includes the same covariates as the primary analysis and is additionally weighted using the same weights. Note that in the estimation model for the FSM subgroup, the FSM covariate is omitted from estimation due to collinearity. The same is true for the number of eligible students covariate in the subgroup analysis exploring differential impacts according to number of eligible students.
4. The sources are NPD data, ILR data, and BMP sample file information.

Table 29 presents the same results as Table 28, this time expressed as risk ratios. This reveals a similar pattern of results. Risk ratios are very close to one for the first two subgroups, indicating that belonging to the subgroup does not make a difference for the impact of BMP. There is a lower risk ratio associated with the subgroup for prior attainment at maths level 2 exams, which indicates that in our sample students that achieved a grade 3 in their previous attempt were less likely to be impacted by BMP than those that achieved a grade 2 or below. However, again there is a wide confidence interval associated with this risk ratio, which crosses one, so we cannot be confident in this finding. Overall, this set of results does not provide evidence of any differential impact of the BMP according to the identified subgroups.

Table 29: Subgroup analysis (multilevel logistic regression)—risk ratios

BMP funding model	Risk ratio	95% confidence interval
Student ever eligible for free school meals	0.991	[0.842, 1.167]
Number of BMP eligible students per setting	1.00	[1.000, 1.000]
Highest prior maths level 2 attainment of grade 3	0.83	[0.58, 1.18]

1. This table reports impact estimates from a multilevel logistic regression with an interaction term for the BMP exposure indicator and each subgroup of interest, and a random intercept for each institution.
2. These were calculated using the `margins` command followed by `nlcom` in Stata for the FSM and prior attainment subgroups. This follows a method outlined in Cummings (2009). For the continuously defined subgroup (institution's number of eligible students), we estimated risk ratios using a generalised linear model. Other notes as per the table above.

Separate treatment arms

Next, we repeat the subgroup analysis carried out above but for each treatment arm separately. This time, the estimation procedure involves re-estimating propensity scores and weights in respect of each treatment arm separately. For consistency with the primary analysis, we again trim observation that have propensity scores greater than 0.97. We also apply a Bonferroni correction equivalent to a significance level of α = 0.017 (which is 0.05/3). A caveat with this analysis is that statistical power is to some extent compromised as the estimation procedure requires not only partitioning the sample into subgroups but also dividing it further by treatment arm.

Table 30 presents these results in terms of odds ratios and Table 31 as risk ratios. The pattern of results, when the sample is separated out by BMP funding model, remains similar. Across the board, we observe confidence intervals that sit almost exactly at one or extend above and below one. This indicates either a null result or lack of statistical significance associated with all coefficients.

Odds and risk ratios associated with the FSM and setting size subgroups are always close to one. The picture is more variable for the subgroup for prior attainment, which has an odds and risk ratio below one for the funding Models A and C, but a positive (above one) odds ratio for funding Model B. A coefficient of below one in this case would indicate the BMP funding is more effective at supporting attainment of students whose prior attainment was the lowest to begin with. For Model C, this finding would imply that we do not have evidence of ‘creaming and parking’ associated with the PBR component of this funding model. However, the confidence intervals associated with the prior attainment subgroup are the widest of all the subgroup analyses indicating that this is measured with the least precision. It is, therefore, difficult for us to draw any firm inferences based on this set of results.

Table 30: Subgroup analysis (multilevel logistic regression)—impact estimates by BMP treatment arm

BMP funding arm	Subgroup	Coefficient (odds ratio)	Standard error	z	P > |z|	95% confidence interval	n
Model A	Student ever eligible for free school meals	1.05	0.16	0.28	0.78	[0.722, 1.512]	31,246
	Number of BMP eligible students per setting	1.00	0.00	1.17	0.24	[1.000, 1.001]	31,246
	Highest prior maths level 2 attainment of grade 3	0.75	0.22	-0.99	0.32	[0.375, 1.501]	31,246
Model B	Student ever eligible for free school meals	0.96	0.13	-0.34	0.74	[0.705, 1.301]	30,799
	Number of BMP eligible students per setting	1.00	0.00	-0.09	0.93	[0.999, 1.001]	30,799
	Highest prior maths level 2 attainment of grade 3	1.11	0.24	0.49	0.62	[0.659, 1.882]	31,799
Model C	Student ever eligible for free school meals	0.97	0.15	-0.18	0.86	[0.677, 1.398]	31,788
	Number of BMP eligible students per setting	1.00	0.00	-0.94	0.35	[0.999, 1.001]	32,788
	Highest prior maths level 2 attainment of grade 3	0.44	0.20	-1.78	0.08	[0.148, 1.323]	32,788

1. This table reports impact estimates from separate multilevel logistic regression models with an interaction term for each BMP funding model and each subgroup of interest, and a random intercept for each institution.
2. BMP eligibility is defined at the setting level using an indicator drawn from the BMP sample file maintained by the evaluation team to monitor institutions that had signed an MOU and not withdrawn from the trial.
3. The regression model includes the same covariates as the primary analysis and is additionally weighted using inverse propensity score weights. Note that in the estimation model for the FSM subgroup, the FSM covariate is omitted from estimation due to collinearity. The same is true for the number of eligible students covariate in the subgroup analysis exploring differential impacts according to number of eligible students.
4. We have applied a Bonferroni correction to account for the fact that this analysis involves estimating three treatment arms. The significance level used to define the confidence interval is α = 0.017 (which is 0.05/3).
5. The sources are NPD data, ILR data, and BMP sample file information.

Table 31: Subgroup analysis (multilevel logistic regression)—risk ratios by treatment arm

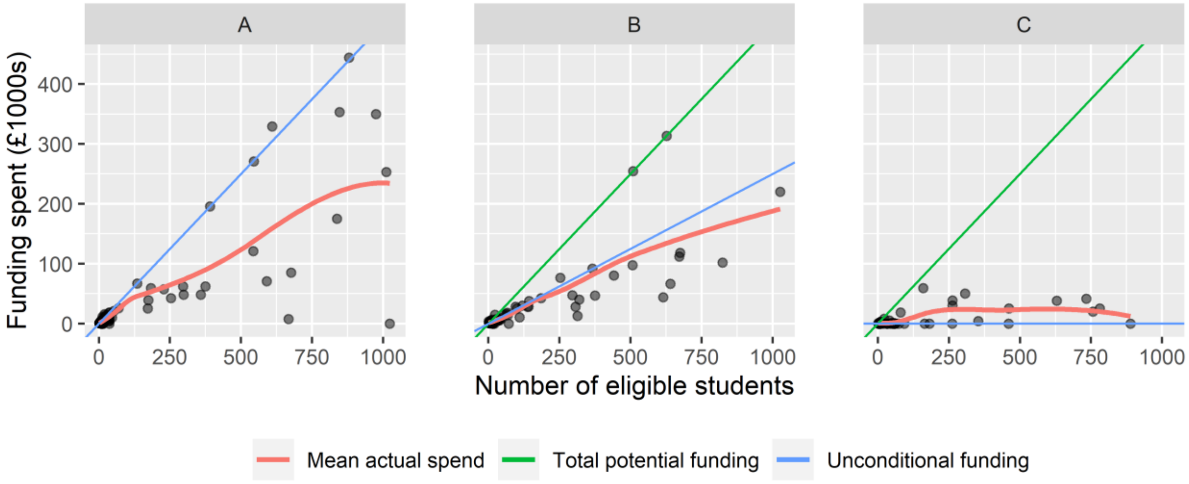
BMP funding arm	BMP funding model	Risk ratio	95% confidence interval
Model A	Student ever eligible for free school meals	1.04	[0.758, 1.423]
	Number of BMP eligible students per setting	1.00	[1.000, 1.001]
	Highest prior maths level 2 attainment of grade 3	0.84	[0.544, 1.289]
Model B	Student ever eligible for free school meals	0.96	[0.744, 1.249]
	Number of BMP eligible students per setting	1.00	[1.000, 1.001]
	Highest prior maths level 2 attainment of grade 3	1.07	[0.771, 1.484]
Model C	Student ever eligible for free school meals	0.98	[0.715, 1.334]
	Number of BMP eligible students per setting	1.00	[0.999, 1.000]
	Highest prior maths level 2 attainment of grade 3	0.61	[0.317, 1.152]

This table reports impact estimates from a multilevel logistic regression with an interaction term for the BMP exposure indicator and each subgroup of interest and a random intercept for each institution. These were calculated using the `margins` command followed by `nlcom` in Stata for the FSM and prior attainment subgroups. This follows a method outlined in Cummings (2009). For the continuously defined subgroup (institution's number of eligible students), we estimated risk ratios using a generalised linear model. Other notes as per the table above.

Dosage analysis (RQ4)

Research Question 4 examines whether the impact of BMP depends on the total amount of funding spent by institutions. To address this question, we used information from wave two of the survey of post-16 settings that participated in BMP (see details in the Methods section). This wave asked about BMP funding spent in 2018/2019. Previous analysis of the survey data (shown in Figure 7) illustrates that the total amount spent by post-16 institutions varies both with the number of students eligible for BMP and the funding model that the institution is allocated to (see the study plan). The hypothesis under test by this research question is whether schools that spent more money were able to make more effective investments, such as recruiting an additional teaching assistant.

Figure 7: Total amount spent (£000s) by post-16 settings, according to treatment group



Regression analysis

First, we carry out a regression analysis. This involves a multilevel logistic regression with a term for different bands of BMP funding spent in 2018/2019, including all student- and institution-level covariates. The specification includes a random intercept for each institution and additionally includes terms for all student and institution-level covariates to increase the precision of the estimate. This analysis is carried out using information from all BMP settings that participated in the wave two survey. This means that there is no non-BMP comparison group for this analysis.

Table 32 reports the findings from this analysis, expressed as odds ratios. For the second and third bands of spending, the odds ratio is very near to one—indicating that there are no discernible differences in the odds of students achieving the required pass in those settings compared to students from settings that spent the lowest amount of BMP funding. There is a positive point estimate for the highest spending category. This would be consistent with the hypothesis that settings that spent the most BMP money were able to obtain more benefits from the programme, potentially through the financing of higher-value and more effective forms of support. However, this evidence is again rather weak as there is a wide confidence interval around this estimate: this crosses one, meaning that we cannot tell with 95% confidence whether the true effect is positive, negative, or zero.

Table 32: Dosage analysis (multilevel logistic regression)—impact estimates

Quartile of BMP spend in 2018/2019	Odds ratio	Standard error	z	P > |z|	95% confidence interval
2	0.95	0.17	-0.28	0.78	[0.627, 1.446]
3	1.07	0.15	0.44	0.66	[0.754, 1.507]
4	1.30	0.20	1.73	0.08	[0.905, 1.869]

1. This table reports impact estimates from a single multilevel logistic regression model with a binary variable for each band of total BMP funding spent in 2018/2019 and a random intercept for each institution.
2. The sample includes BMP institutions that participated in the wave two survey and could be linked to the NPD only. This contains 28,683 students from 237 institutions altogether (of which 4,022 are from 60 institutions in the lowest spending band, 907 are from 60 institutions in the second spending band, 2,802 are from 59 institutions in the third spending band, and 20,952 are from 58 institutions in the highest spending band.
3. The regression model includes the same covariates as the primary analysis.
4. We have applied a Bonferroni correction to account for the fact that this analysis involves estimating three coefficients. The significance level used to define the confidence interval is α = 0.017 (which is 0.05/3).
5. Sources are NPD data, BMP sample file information, and data from wave two of the BMP settings survey.

Marginal effects analysis

We also carried out a marginal effects analysis to explore RQ4. Estimates from this analysis represent the predicted average change in the probability of students achieving the required pass by November 2019 for each additional pound of BMP funding spent in 2018/2019. Table 33 reports the results from this analysis, showing that an additional pound of BMP funding spent does not have an impact.

Table 33: Dosage analysis—marginal effect of each additional £1 of BMP funding spent

Independent variable	Average marginal effect (dy/dx)	Standard error	z	P > |z|	95% confidence interval
Total spend in 2018/2019	0.000	0.000	0.46	0.646	[0.000, 0.000]

1. This table reports the results of a marginal effects analysis carried out following estimation of a multilevel logistic regression of the likelihood of achieving the required pass against the total amount of BMP funding spent in 2018/2019. The marginal effects analysis is carried out using the `margins, dydx` command in Stata.
2. The sample includes BMP institutions that participated in the wave two survey and could be linked to the NPD only. This contains 28,683 students from 237 institutions.
3. Sources are NPD data, BMP sample file information, and data from wave two of the BMP settings survey.

This null finding is by no means surprising considering that it would be remarkable for an additional £1 spent on any intervention to meaningfully influence expected attainment outcomes. To understand whether marginal effects of BMP spend may be evident at higher increments, we have repeated this analysis with a different scaling. Table 34 displays the results of this same estimation approach, this time examining whether each additional £1,000 of BMP funding spent in 2018/2019 led to any predicted average change in the likelihood of achieving the outcome. The results of this rescaled analysis show that an additional £1,000 of BMP funding also does not translate into any tangible change in the average likelihood of achieving the outcome.

Table 34: Dosage analysis—marginal effect of each additional £1,000 of BMP funding spent

Independent variable	Average marginal effect (dy/dx)	Standard error	z	P > |z|	95% confidence interval
Total spend in 2018/2019	0.000	0.000	0.47	0.642	[0.000, 0.000]

1. This table reports the results of a marginal effects analysis carried out following estimation of a multilevel logistic regression of the likelihood of achieving the required pass against the total amount of BMP funding spent in 2018/2019. The marginal effects analysis is carried out using the `margins, dydx` command in Stata.
2. The sample includes BMP institutions that participated in the wave two survey and could be linked to the NPD only. This contains 28,683 students from 237 institutions.
3. Sources are NPD data, BMP sample file information, and data from wave two of the BMP settings survey.

Retake analysis (RQ5)

Research Question 5 assesses the impact of BMP on whether students attempted a GCSE maths examination retake, as opposed to being entered for Functional Skills Level 2 only, or not attempting any relevant maths resit at all. As with the other impact evaluation research questions, the time horizon of interest is the 2018/2019 academic year and the November 2019 resit period. For this question, the sample of interest is students who achieved a grade 2 or below (including U) in their previous maths level 2 attempt. These are the students who can study towards either a GCSE maths or Functional Skills Level 2 maths exam as part of the condition of funding (whereas students who achieved a grade 3 are required to enter a GCSE exam).

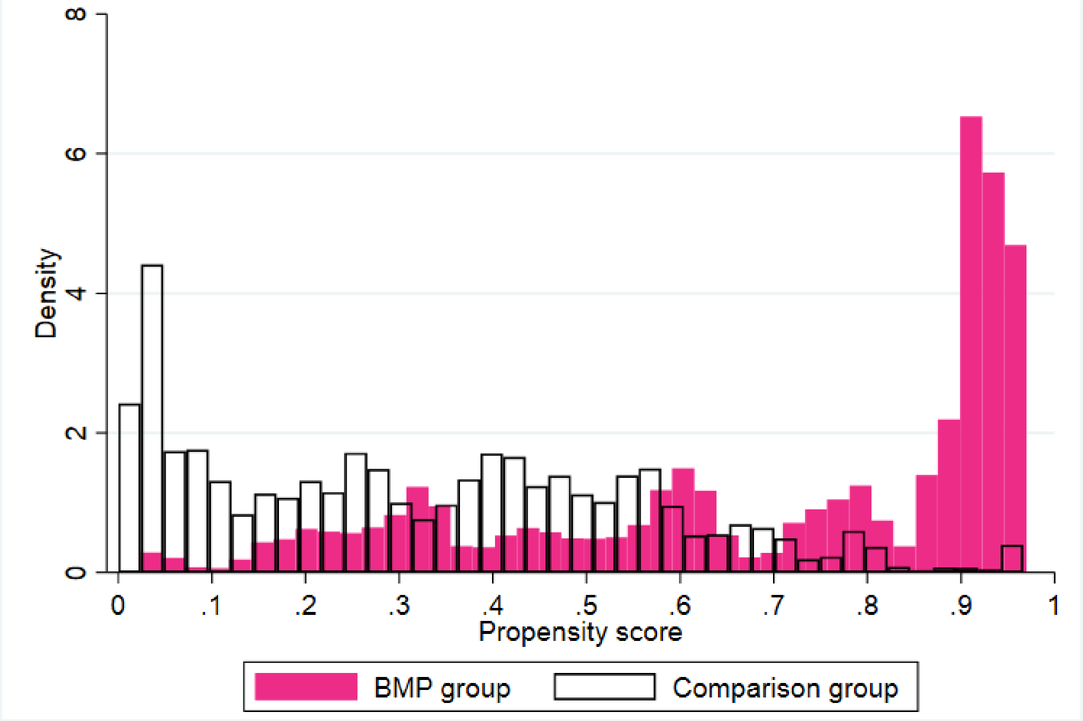
Descriptive analysis presented above in Table 15 and Table 16 already show that the vast majority of eligible students did re-take a GCSE exam within this timeframe. Comparatively few students entered for a Functional Skills Level 2 exam. Among those who did, all of them also entered for a GCSE exam within the same timeframe. This means that there are no students in our main analysis sample that entered for a Functional Skills Level 2 exam only. The main outcome for this analysis takes the value one if a student re-took a GCSE exam, and zero if they did not. Given that there are no students who entered for a Functional Skills Level 2 exam without also entering for a GCSE exam, this analysis will, in effect, capture the impact of BMP on whether students resit a GCSE exam compared to no exam at all. The interpretation is, therefore, slightly different to what was originally intended as this analysis will not shed light on whether BMP has an influence on causing students who would otherwise have entered for a Functional Skills Level 2 exam only to ‘upgrade’ to enter for the more difficult GCSE exam.

This analysis was carried out in the same way as the main doubly robust analysis, but with a different key outcome. For consistency, we applied the same trimming rule as in the primary analysis. This means that we removed observations with propensity scores greater than 0.97. We were also required to make a small change to the covariates used for weighting since it turned out that the two covariates capturing whether institutions were of the academy converters or academy sponsor types were found to perfectly predict the outcome. This meant that all students belonging to institutions of these two types re-took a GCSE exam, which prevents the weighting model from being identified. We proceeded with the analysis by dropping these two covariates, which makes the weighting model slightly less granular than intended. It will, in effect, mean that institutions of academy converter and academy sponsor types are now included in the omitted institution type variable along with those where the type is listed as missing in the NPD data. The root of this issue is the fact that a reasonably small proportion of students belonged to settings of those two types while the vast majority met the outcome under test; this, together, meant that there was no variation in the outcome within these two institution types that could be exploited to identify model coefficients. This change to the intended set of weighting covariates is minor and makes little appreciable difference in practice.

Model diagnostics

The diagnostic results for this analysis are extremely similar to the final model of the primary analysis. Figure 8 reports the distribution of propensity scores in the BMP and non-BMP group. Table 35 shows standardised mean differences before and after weighting. We find that weighting improves the number of covariates for which there is an absolute standardised mean difference of less than 0.1 to 17 of 26 (65.4%). Figure 9 shows the same information in a visual form. The pseudo-R^2 drops from 0.46 before weighting to 0.15 afterwards.

Figure 8: Retake analysis—p: Propensity score distributions



The figure shows the distribution of propensity scores among the sample of students that achieved a grade 2 or below in their initial GCSE attempt after trimming observations with propensity scores > 0.97. The sample covers 13,356 eligible students from 239 BMP settings and 12,320 students from 324 non-BMP settings in AEA Category 4 areas who would have been eligible for the funding if it were offered in their school. Propensity scores are based on a logit model of BMP exposure. Sources are NPD data, ILR data, and BMP sample file information.

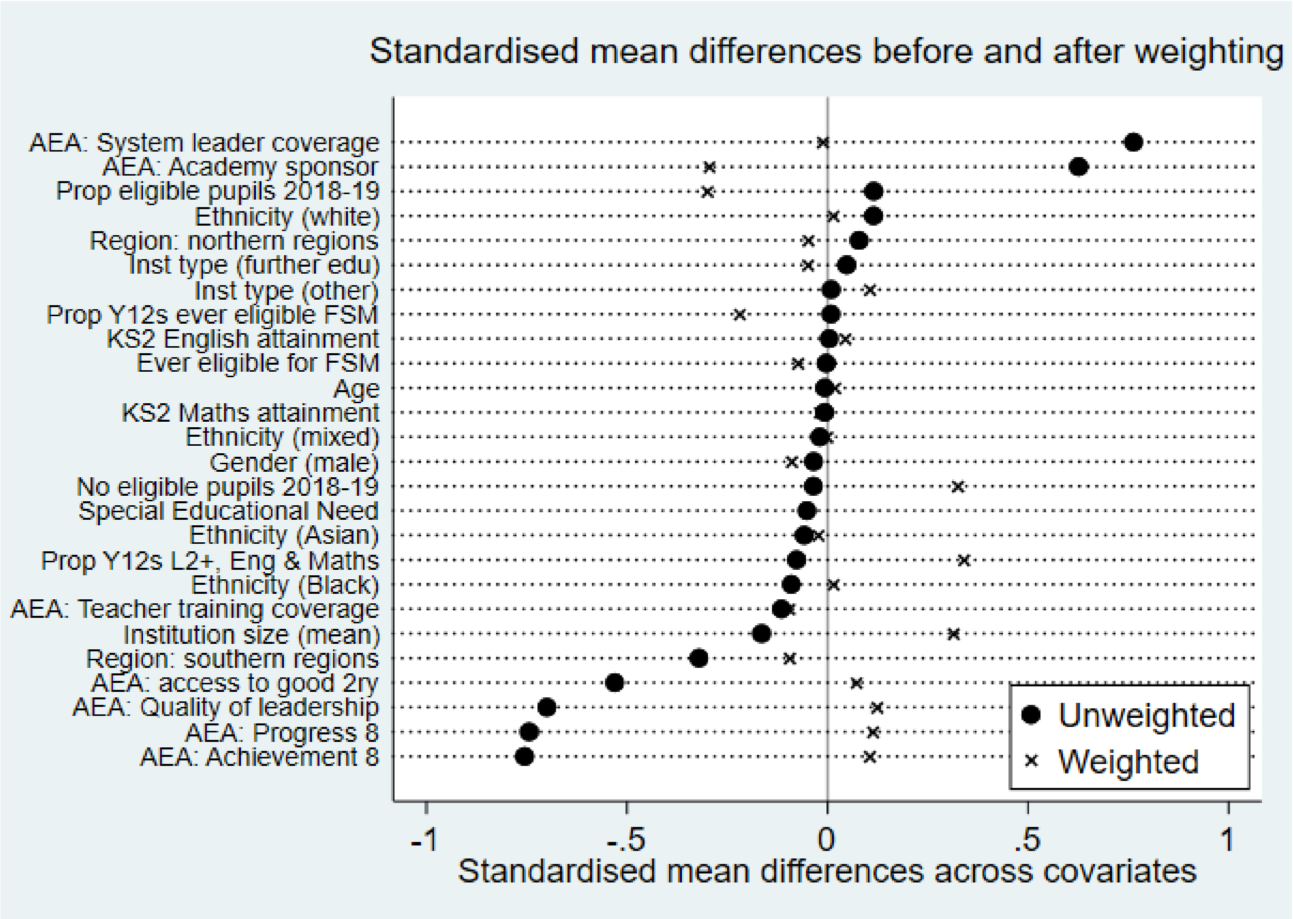
Table 35: Retake analysis—standardised mean differences before and after weighting

	Unweighted raw sample			Weighted sample		
Covariate	**Treated mean**	**Control mean**	**Standardised mean difference**	**Treated mean**	**Control mean**	**Standardised mean difference**
Institution size mean	6395.99	7307.12	-0.22	6456.79	5692.85	0.20
Number of eligible Year 12s in 2018/2019	518.07	542.46	-0.07	534.67	445.98	0.29
Proportion of eligible Year 12s in 2018/2019	47.46	45.33	0.17	46.70	45.83	0.07
Institution type: proportion of settings belonging to the further education sector	88.93	87.25	0.05	88.11	85.97	0.06
Institution type: proportion of settings with an ‘other’ institution type	5.80	5.15	0.03	6.06	5.22	0.04
Region: Southern regions	17.11	32.02	-0.35	21.75	40.89	-0.42
Region: Northern regions	37.99	34.03	0.08	37.41	30.74	0.14
Proportion of Year 12s ever eligible for FSM	32.75	32.08	0.06	31.88	28.37	0.35
Proportion of Year 12s who achieved a level 2 or higher in maths and English at KS4 prior to enrolment	34.93	37.27	-0.15	35.58	35.05	0.03
AEA Access to a good secondary school index	57.89	66.67	-0.58	62.24	62.74	-0.03

	Unweighted raw sample			Weighted sample		
Covariate	**Treated mean**	**Control mean**	**Standardised mean difference**	**Treated mean**	**Control mean**	**Standardised mean difference**
AEA Achievement 8 indicator	46.44	48.19	-0.83	47.32	47.95	-0.37
AEA Progress 8 indicator	-0.15	-0.04	-0.86	-0.10	-0.06	-0.33
AEA System leader coverage indicator	2567.25	1657.72	0.88	2222.82	1650.70	0.56
AEA Initial teacher training provider coverage index	40.93	43.01	-0.12	41.81	40.65	0.07
AEA Quality of leadership indicator	72.48	81.46	-0.81	76.16	75.39	0.06
AEA Academy sponsor coverage	1536.32	723.50	0.77	1012.13	1214.21	-0.23
Age	16.00	16.01	-0.01	16.00	16.00	0.01
Gender	51.89	52.68	-0.02	51.99	48.59	0.07
Ethnicity: White	79.27	73.7	0.13	82.60	85.83	-0.09
Ethnicity: Mixed/ multiple	4.62	5.00	-0.02	4.70	3.69	0.05
Ethnicity: Asian	10.84	12.94	-0.06	8.28	6.71	0.06
Ethnicity: Black	4.16	6.64	-0.11	3.74	2.58	0.07
Eligibility for free school meals	43.15	42.78	0.01	43.79	39.50	0.09
Special Education Need status (SEN)	35.54	36.77	-0.03	37.86	38.22	-0.01
Prior attainment at KS2 in English	3.43	3.45	-0.01	3.43	3.39	0.04
Prior attainment at KS2 in maths	3.32	3.32	-0.01	3.31	3.26	0.06

1. This table reports standardised mean differences before and after weighting using student-level, institution-level, and AEA covariates. Propensity scores are derived using a logit model. Cells shaded in red are those with absolute standardised mean differences > 0.1 and those in green, < 0.1.
2. The sample for this analysis is eligible students who received a grade 2, 1, or U in their previous maths level 2 attempt in 2017. The treatment group contains students from BMP schools in AEA Categories 5 and 6, and the comparison group is comprised of students who would have been eligible for BMP had it been offered in their school, from AEA Category 4.
3. The sample sizes before weighting are 25,323 students from 373 BMP settings and 13,799 students from 341 non-BMP settings. The sample size after weighting (including trimming) are 13,356 students from 239 BMP settings and 12,320 students from 324 non-BMP settings.
4. Three covariates are omitted as they are perfect linear combinations of other included covariates (and the model includes a constant term): these are Region: East and Midlands; Institution type: 'missing'; and Ethnicity: 'other'. Institution types 'academy converter' and 'sponsor covariates' are also omitted from this model since they turn out to perfectly predict attainment of the outcome.
5. The sources are NPD data, ILR data, and BMP sample file information.

Figure 9: Retake analysis—standardised mean differences before and after weighting



1. This figure reports standardised mean differences in covariates before and after weighting. Propensity scores are derived using a logit model.
2. The sample for this analysis is eligible students who received a grade 2, 1, or U in their previous maths level 2 attempt in 2017. The treatment group contains students from BMP schools in AEA Categories 5 and 6, and the comparison group is comprised of students who would have been eligible for BMP had it been offered in their school, from AEA Category 4.
3. The sample sizes before weighting are 25,323 students from 373 BMP settings and 13,799 students from 341 non-BMP settings. The sample size after weighting (including trimming) are 13,356 students from 239 BMP settings and 12,320 students from 324 non-BMP settings.
4. Three covariates are omitted as they are perfect linear combinations of other included covariates (and the model includes a constant term): these are Region: East and midlands; Institution type: 'missing'; and Ethnicity: 'other'. Institution types 'academy converter' and 'sponsor covariates' are also omitted from this model since they turn out to perfectly predict attainment of the outcome.
5. The sources are NPD data, ILR data, and BMP sample file information.

Retake analysis results

Table 36 reports the results from this analysis where the coefficient represents percentage point units of change in the likelihood of students retaking a GCSE. The pattern of results turns out to be similar to the primary analysis. There is no discernible impact on the outcome. The same results are expressed as risk ratios in Table 37. Here, we again find no impact when BMP exposure.

Table 36: Retake analysis—impact estimates

BMP treatment indicator	ATET	Standard error	z	P > |z|	95% confidence interval
Setting-level	0.02	0.03	0.61	0.54	[-0.047, 0.089]

1. This table reports impact estimates from doubly robust estimation, performed using `teffects ipwra` in Stata with a logit specification for the treatment and outcomes model.
2. BMP eligibility is defined at the setting level using an indicator drawn from the BMP sample file maintained by the evaluation team to monitor institutions that had signed an MOU and not withdrawn from the trial. The sample for this analysis is eligible students who received a grade 2, 1, or U in their previous maths level 2 attempt in 2017. This specification includes 13,356 students from 239 BMP settings and 12,320 students from 324 non-BMP settings.

3. Standard errors for the setting-level BMP exposure indicator are clustered at the institution-level.
4. The sources are NPD data, ILR data, and BMP sample file information.

Table 37: Retake analysis: risk ratios

BMP treatment indicator	Risk ratio	95% confidence interval
Setting-level	1.02	[0.951, 1.100]

This table reports impact estimates in the form of risk ratios, from doubly robust estimation, performed using `teffects ipwra` in Stata with a logit specification for the treatment and outcomes model. Other notes as per the above table.

In Appendix K of the Additional Technical Appendices document we repeat this analysis using the second BMP eligibility indicator (defined at the student level). The results are somewhat contrasting, showing a positive and statistically significant impact of the BMP on the likelihood of a student re-sitting GCSE maths. As with the primary analysis, we caution against placing too much emphasis on this additional analysis as it is exploratory.

Estimation of ICCs

Table 38 reports the ICC for our primary outcome, where student outcomes are clustered within settings. This ICC was used as the basis for the updated sample size calculations reported above in Table 12.

Table 38: ICC estimation

BMP treatment indicator	ICC
Setting-level	0.279

This table reports the ICC for our primary outcome: whether or not students achieved the required pass in a maths level 2 resit by November 2019. It is estimated based on a multilevel logit model, with students clustered within settings. We use Statas `melogit` command, followed by `estat icc`. The ICC is estimated across the final analysis sample used for the primary estimation model. This includes 47,310 observations altogether, and 679 settings.

Implementation and process evaluation results

This section outlines key findings from the implementation and process evaluation of the BMP pilot. It brings together findings from the three-wave survey (wave one in February and March 2019, wave two in November 2019, and wave three in October 2020) and longitudinal interviews (initial interviews from May to July 2019 and a follow-up in October 2020) with the institutions involved (see Methods). Findings are organised thematically and discussed sequentially, moving from a brief discussion of compliance followed by implementation fidelity, usual practice, and intervention activities, institutions' responsiveness, perceived outcomes, and costs of taking part. The profile of the institutions that responded to the three waves of the survey and their representativeness of the BMP settings is briefly described below for context.

Survey respondents' profile

The first survey was sent to the 439 institutions that were participating in the pilot at the time and achieved a 67% response rate. Of those that responded, 36% were in funding Model A (versus 32% in the whole sample), 34% were in Model B (versus 33%), and 29% were in Model C (vs 34%).[31] Respondents had 145 eligible students, on average, compared to an average of 112 across the whole final sample of 434 BMP institutions.

The second survey was sent to 434 institutions and achieved a 57% response rate. Of those that responded, 37% were in funding Model A, 36% in Model B, and 28% in Model C. Respondents had 144 eligible students on average.

The third survey was sent to 434 institutions and achieved a 27% response rate. Of those that responded, 32% were in funding Model A, 34% in Model B, and 34% in Model C. Respondents had an average of 123 eligible students.

Compliance

Measures of compliance are generally used in the impact evaluation. This accounts for non-compliance and thus gives an indication of the treatment effects among those who participated.

Given the nature of the intervention, non-compliance was not anticipated to be a concern. All eligible institutions that took part in the evaluation received additional funding from the DfE as per the treatment condition each institution was randomly assigned to. All institutions were given access to a resource page that they could use to guide spending of additional resources. However, there was no restriction on how participating institutions could spend the additional funding except for the requirement to spend it on GCSE maths resit students to improve their outcomes by summer 2020. Findings relating to money received but not spent, or spent on a wider cohort, are discussed in the section on Fidelity below.

Fidelity

This section assesses the extent to which the intervention was delivered as intended so as to benefit GCSE maths resit students, as outlined above. The types of activities and resources that institutions spent BMP funding on are outlined in the section Intervention Activities. Implementation deviated from the intended model in three key ways: funding not being spent fully by the end of the intervention, funding being spent on different groups of students, and funding being spent also for other institutions in multi-academy trusts.

Unspent funding

One key issue was that funding was not used in full by all institutions in the two years of the intervention (2018/2019 and 2019/2020). Survey data from across the three waves showed delays in spending. For example, at wave two, 82% of respondents said they had spent some funding in 2018/2019 while 18% said they had not spent any money that year. Spending was lower at this point for funding Model C, where institutions had spent an average (mean) of 16% of their maximum possible funding compared to institutions in funding Models A and B that had spent 60% and 41% respectively.

[31] Percentages do not sum to 100% due to rounding.

When surveyed again at wave three, institutions reported having spent, on average, less than two-thirds of their maximum possible BMP funding (58%).[32] There may have been underlying reasons for this underspend: first, data on spending was self-reported rather than financial administrative data and therefore may have been somewhat inaccurate or based on estimates. Second, compared with previous survey waves, a higher proportion of respondents in wave three were in funding Model C. These institutions had a lower average spend, which is likely to be reflected in the overall lower average spend.

The average proportion spent differed considerably by funding model. Table 39 shows the total amount of BMP spend at wave three by funding model, calculated as a sum of funding spent on different activities. Table 40 shows the average spending as a percentage of maximum possible funding. Institutions in Model A reported the highest proportions; the average percentage spent was lower for models which followed payment by results: Model B (50%) and Model C (43%) where all funding was conditional on student results.

Table 39: Total amount of BMP funding spent in 2018 to 2020, by funding model

Amount spent (£)	Funding model		
	Guaranteed payment (A)	Half guaranteed, half by results (B)	Payment by result (C)
Mean	£86,251	£38,229	£17,192
SD	£175,352	£67,969	£35,870
Min.	£0	£0	£0
Max.	£835,004	£280,000	£191,000

n: A = 28, B = 30, C = 21. The sample sizes reflect the number of institutions that completed wave three of the survey with outliers removed. Outliers were removed if the reported percentage of available BMP funding that was spent was above 100 or below 0.

Table 40: Total amount of BMP funding spent in 2018 to 2020 as a percentage of maximum possible BMP funding, by funding model

Amount spent (£)	Funding model		
	Guaranteed payment (A)	Half guaranteed, half by results (B)	Payment by result (C)
Mean	78	50	43
St. Dev.	24	23	33
Min.	0	16	2
Max.	100	100	100

n: A = 28, B = 30, C = 21.

Delays and lack of spending suggest that the funding designed to support a specific cohort of students was not used to support their learning as intended. Within initial interviews, heads of maths in funding Models B and C highlighted that the dependency of funding on student achievement made it difficult to plan spending effectively, especially when upfront costs were high.

> *'The fact that we've got this potential funding in the future that will reward us for that, that's great, but we've still got to find the funds now to do what we do' (head of maths, Model C).*

[32] Sixteen institutions were excluded from this part of the analysis as they reported spending more than 100% of the maximum possible BMP funding available to them. Outlying values ranged from 101% to 550%.

In this context, institutions raised that they preferred funding to be provided upfront and found the payment by results models unfair due to the financial risk associated with spending money that they might not receive.

> *'I think it's really unfair because you would not really know how much money you were going to get. I don't think we would have been able to spend any additional money on that basis, so for us it wouldn't have really worked. … I would never have been able to employ two staff on the basis that I might get a certain amount of students through a GCSE. That would have been too much of a financial risk' (head of maths, Model A).*

Some institutions assumed a percentage success rate for students (for example, 20%) and did not spend money over the amount they expected to receive. Payment by results was less of an issue where institutions had additional funding from other sources which could be used to cover any shortfall. For more details about other factors that might have affected spending, including understanding of the intervention and challenges to decision-making, see section on Responsiveness.

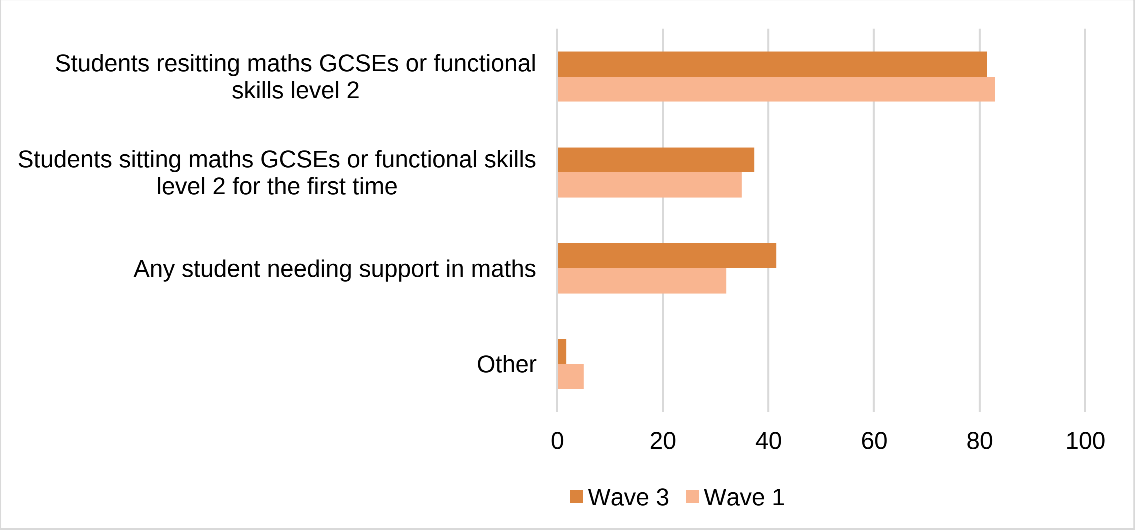
In wave three of the survey, institutions that had not yet spent all the expected funding were asked when the remaining funding would likely be spent: 30% said they were planning to spend it in the 2020/2021 academic year, 52% that they would spend some of it in 2020/2021 and some in 2021/2022, and 19% said they would spend all of it in 2021/2022.

Funding spent on non-eligible students

Implementation also deviated from the intended model in terms of the target groups. The intervention was expected to benefit students who were resitting GCSE and Functional Skills Level 2 maths for the first time, however, survey results show that funding and resources may not have always been used on the intended recipients of the intervention in that other groups benefited from BMP funding (Figure 10). At wave one, just over one-third of institutions (35%) said they had spent or were planning to spend BMP funding on students who were sitting maths GCSE or Functional Skills for the first time, while 83% had spent or were planning to spend it on all students resitting maths GCSE or Functional Skills Level 2. Additionally, just under a third (32%) said that they had spent or would spend the funding on any student who needed support in maths (5% responded 'other').[33]

The percentage of funding that was reported to have been spent on all students resitting maths GCSE or Functional Skills Level 2 (rather than just those who were resitting for the first time) remained high over the intervention. At wave three, 81% of institutions said that BMP funding had been spent on this group. Additionally, the percentage of institutions that reported spending funding on any student needing support in maths rose to 42%.

Figure 10: Groups on which BMP funding was planned for or spent (% institutions)



Sample sizes: wave one, n = 296; wave three, n = 118.

Interviews support the survey results and highlight that in some cases funding was spent on GCSE maths students across the board to help increase overall achievement in maths. Some interviewees explained that the learning of Year

[33] Percentages do not add up to 100% as institutions were asked to select all options that applied.

11 students was supported so that results were not impacted and GCSE students did not need to resit the following year. Similarly, institutions reported that they targeted students sitting Functional Skills Level 1 who would be taking the Level 2 qualification the following year. Within interviews, heads of maths reported difficulties in dedicating resources to resitters. Although more funding and resources were planned to be allocated to this group, this was not always possible in practice for reasons such as limitations on staff time. There was also additional evidence of heads of maths reporting lack of clarity over initial objectives which is discussed further in the section on Responsiveness.

Evidence from interviews also suggests that funding was not always spent only on maths. Interviews showed that in some cases, funding for maths was placed in a combined pot for maths and English. The quote is a respondent's rationale for doing this.

> *'I think it's difficult, and not necessarily wise, to try and separate maths and English. I think they're the same challenges' (head of maths, Model B).*

Statements such as the above suggest that overall spending of BMP funding may have been under-reported by institutions as some of this was spent on areas outside of maths.

These types of divergence from the eligibility conditions of the intervention often stemmed from a lack of clarity around student eligibility. For example, some heads of maths thought that all GCSE maths students were eligible for funding and resources.

> *'We could spend the money how we felt, as long as it supported GCSE maths' (head of maths, Model B).*

Some interviewees stated that they received conflicting information about eligibility. They reported examples where they had initially been led to believe that BMP funding was intended for all GCSE maths resit students but were later informed that it was only for students resitting their maths GCSE for the first time. As a result, there was some uncertainty around the number of eligible students. For further discussion of institutions' views on eligibility, see Responsiveness.

Interviews also showed that heads of maths' understanding of requirements around measures of success was sometimes limited. For example, some were unsure as to whether conditional funding as part of the payment by results models was dependent on students achieving a level 4 or a level 2 in GCSE maths. More generally, there was some initial uncertainty about the objectives of the intervention and a lack of clarity on how funding should be allocated (see Responsiveness).

Funding spent on specific groups or resitters

Interviews and survey results also show instances where funding and resources were targeted at specific groups of maths resitters. Survey results from wave three indicate that although nearly three-quarters (74%) of institutions said that they targeted resources and funding equally at all resitters, just over one quarter (26%) said they did not. Interviews highlight that the payment by results funding models led some institutions to focus time and resources on learners who were very close to getting a pass at the expense of other learners who were less likely to achieve a level 4 in GCSE maths. Heads of maths also raised the concern that staff were sometimes more enthusiastic to work with higher level students as lower level students were perceived to be harder to motivate and more likely to be already disengaged with the learning process.

Funding spent on other institutions within multi-academy trusts

In the second survey wave, 42% of institutions stated they were part of a multi-academy trust: the proportion was lower in funding Model A (38%) than in Models B and C (44% and 43% respectively). There was some evidence that the BMP was used to support GCSE maths resitters in other academies in the same trust, with 13% of respondents overall answering 'yes'. However, there was also a lack of knowledge from survey respondents about whether this happened, particularly among those in funding Model C (28%).

Table 41: Whether BMP had been used in other academies in the same trust, by funding model (% institutions)

Response	Funding model			
	Payment upfront (A)	Half guaranteed, half by results (B)	Payment by result (C)	All
Yes	9%	18%	10%	13%
No	74%	69%	62%	69%
Don’t know	18%	13%	28%	19%
Total	100.0	100.0	100.0	100.0

n: A = 34; B = 39; C = 29.

Usual practice

RQ1 What activities and tools are currently available in pilot institutions to support GCSE maths resitters?

RQ2 Of these activities and tools, which were not available before 2018/2019?

RQ3 Which activities and tools were phased out after 2018/2019?

RQ4 To what extent were the amount of support and the type of activities tailored to the needs of individual students?

RQ5 What funding schemes, other than the intervention, have been used to support GCSE maths resitters?

The following section describes usual practice to support GCSE maths resitters in 2017/2018 to 2019/2020 at intervention institutions. ‘Usual practice’ refers to activities and resources which were not funded through the BMP.

There were difficulties differentiating usual practice from BMP-funded activities in 2018/2019 and 2019/2020. Some heads of maths were not aware whether activities were funded through the BMP or usual practice funding; others reported that activities had been paid for from both sources. In this section we present usual practice to support resitters as far as respondent data made clear.

Usual practice in 2017/2018 to 2019/2020

In the first survey wave, only 7% of institutions said they received additional funding to help improve maths attainment aside from the usual statutory funding per student. Within this 7%, most of the additional funding received was aimed at improving attainment at level 2, including GCSEs. Institutions told us they received an average of £43,000 to support these students (see Table 42). By contrast, institutions received, on average, only £333 to support students working towards level 3 qualifications, such as A-Levels.

Table 42: Additional usual practice funding received per qualification level, 2017/2018 and 2018/2019 to 2019/2020

Year	Qualification level	Mean	Maximum	Minimum	Standard deviation
2017/2018	Level 1	£524	£10,000	£0	£2,182
	Level 2	£42,954	£170,000	£0	£64,761
	Level 3	£333	£4,000	£0	£966
2018/2019 to 2019/2020	Level 1	£16,718	£240,000	£0	£59,634
	Level 2	£116,323	£400,000	£0	£171,503
	Level 3	£4,618	£39,600	£0	£10,719

n: wave one = 21; wave two = 16. ‘Level 1’ = Level 1 Functional Skills; ‘Level 2’ = GCSEs Level 2 Functional Skills; ‘Level 3’ = A Levels.

When asked about the activities and resources in place in the institution to support students who had yet to pass their maths GCSEs, the most common answers were ‘engagement with parents’ (47%), ‘staff training’ (45%), and ‘increased teaching staff' (44%) (see Figure 11). A small proportion of institutions (8%) said they had no activities to support resitters in the 2018/2019 academic year. The most common reason was 'lack of funding', selected by 17 of 24 respondents.

In the initial interviews, heads of maths described a range of activities for resitters funded by non-BMP funding streams. Additional teaching hours for resitters was one popular activity. This could take the form of sessions exclusively for Year 12 resitters or sessions that were shared between resitters and Year 11s working towards their GCSEs. These might be compulsory or optional 'drop-in' sessions. Alternatively, they could involve one to one tutoring sessions, which was seen as an effective way of engaging with students who struggled in large classes.

> *'I think that a lot of the young people that we're working with have come from a school environment where a group is an opportunity to play up and have a bit of fun ... if it's a one to one basis with a tutor ... it's very, very different to what happened at school' (head of maths, Model C).*

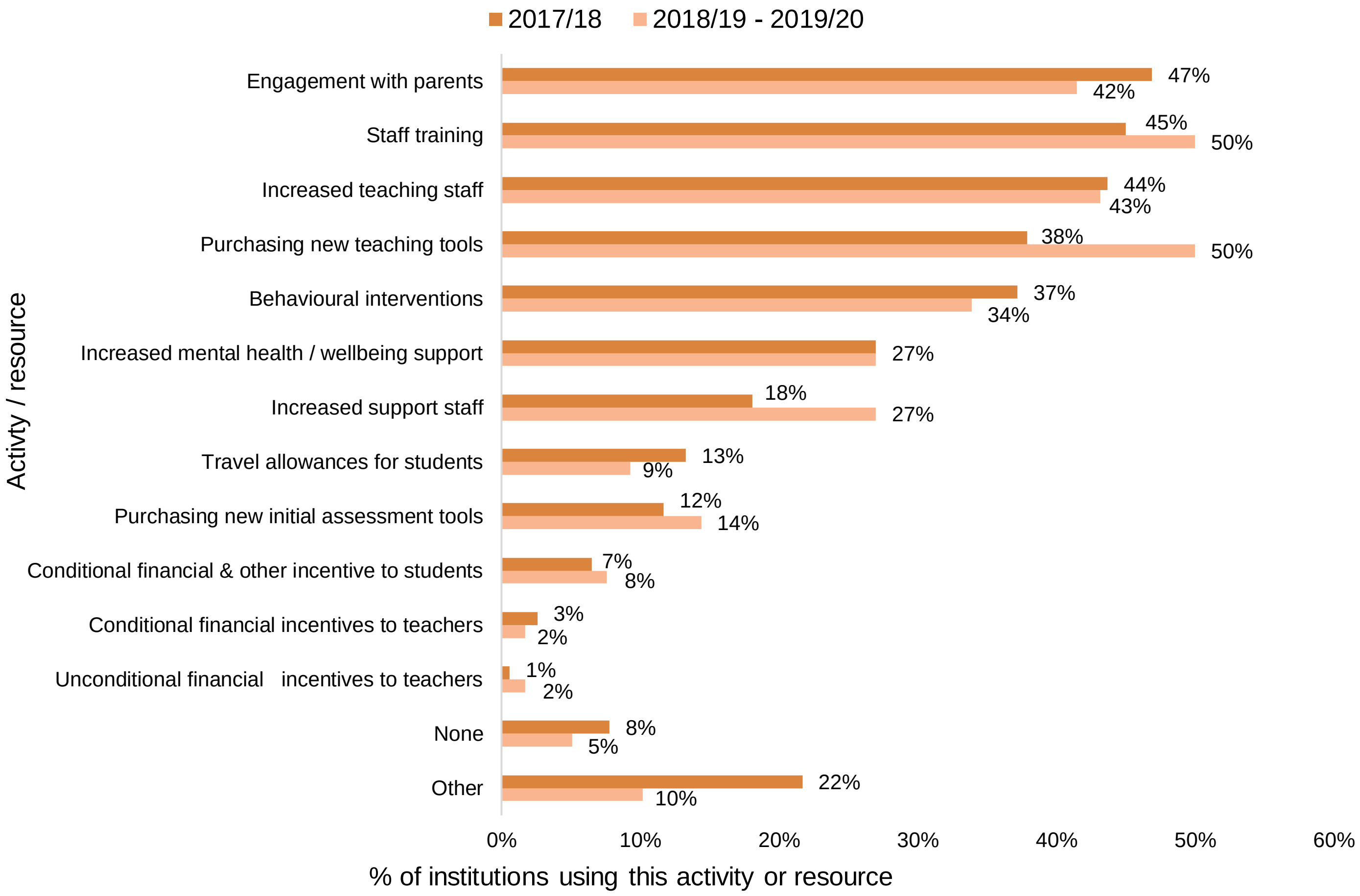
Aside from additional teaching hours, institutions favoured maths apps such as Hegarty Maths or MyMaths. These were valued for their ability to identify a student's maths level and offer appropriate practice activities in response. They were also seen as good value for money, especially compared to activities involving staff.

Usual practice in 2018/2019 to 2019/2020

According to the third survey wave, 18% of institutions said they had received additional funding other than the BMP to support resitters in the time period between 2018 and 2020. Within this 18%, institutions had received an average of £116,322 to support students working towards level 2 qualifications, £16,718 for level 1 qualifications, and £4,618 for level 3 qualifications (see Table 42).

Heads of maths reported that the most common usual practice activities and resources during this period were 'upskilling staff' (50%), 'purchasing new teaching tools' (50%), and 'increased teaching staff' (43%) (see Figure 11). Only 5% had no usual practice activities in place to support resitters. When asked the reason, four cited a lack of funding, two said they did not think these were necessary, and two said activities funded through the BMP were sufficient. Responses on usual practice from the third survey wave are similar to the first wave, however, 'purchasing new teaching tools' and 'increased support staff' were substantially more common in the third wave while 'engagement with parents' was less common.

Figure 11: Usual practice activities and resources in place to support resitters, 2017/2018 and 2018/2019 to 2019/2020



In the follow-up interviews, heads of maths generally described continuity between usual practice activities in place in 2018/2019 and in 2019/2020 and were often unsure what activities were novel to 2019/2020, compared to 2018/2019.

The preference for additional staffing seen in the third survey wave was supported by the follow-up interviews. Heads of maths said they hired revision specialists or attendance officers (who would chase up students—or their parents—when students missed revision sessions) as well as additional staff in the regular maths teaching unit. Funding could also be used to extend the hours of existing staff, which allowed for more sessions or smaller class sizes.

Similar to the 2017/2018 to 2018/2019 period, another use of usual practice funding was new software suites, including 'virtual teaching' apps. Software could also have other uses, such as ProMonitor, a virtual learning platform that logs many types of student data including attendance, behavioural issues, or AQA enhanced results analysis—a tool which feeds back question-by-question results from every exam a resitter has taken. More traditional teaching resources included regular benchmarking of student grades and maths needs through taking past papers, and revision guides.

Heads of maths sometimes made pragmatic choices with students who struggled to achieve a grade 4 at GCSE after multiple attempts. For example, they might hope for a pass in the November resits but keep resits later in the year as a backup. Where a resitter had to do more than one resit, the institution could switch the exam board for each attempt, hoping they might be able to achieve a grade 4 with a different paper. A resitter who repeatedly failed could, as a last resort, sit the International GCSE—thought to be more accessible because of its less complex language. Another pragmatic choice institutions made when GCSE pass rates were persistently low was to move students off GCSE maths and onto the less academically-focused Functional Skills.

Activities of all types—including additional lessons, one to one support or activities set through tutoring software, and either during or after school—could be either mandatory or optional. Attendance was enforced through chasing resitters, following up with parents, and, in some cases, through behavioural contracts. This was described as effective—with attendance of up to 90% at after school revision sessions.

Usual practice activities phased out in 2019/2020

Heads of maths said their institutions most commonly phased activities out for reasons of expense. For example, in the case of maths teaching apps, these might be phased out for being too expensive or because they had been superseded by a newer, superior app.

Specific approaches to resits could also be phased out. For example, one institution had entered all resitters for the November resits in 2018/2019 but changed its approach in 2019/2020. In this year, only those most likely to pass were put forward for the November resits while other resitters were given more time to revise and prepare for resits later in the academic year.

Intervention activities

RQ1 What activities and tools are currently available in pilot and control institutions to support GCSE maths resitters?

RQ2 Of these activities and tools, which were not available before 2018/2019?

RQ3 Which activities and tools were phased out after 2018/2019?

RQ4 To what extent were the amount of support and the type of activities tailored to the needs of individual students?

This section examines what activities and resources were paid for with BMP funding in 2018/2019 and 2019/2020.

Intervention activities in 2018/2019

Institutions had spent some of the BMP funding at the time of the first survey wave in 2019 (see Fidelity). At that time, BMP funding was largely used to increase teaching staff, with institutions spending £7,467 on this, on average (see Figure 12). Funding was also used to increase support staff (£3,319) and purchase new teaching tools (£1,815). For any unused BMP funding, the most popular items institutions intended to spend it on were purchasing new teaching tools (55%), increasing teaching staff (43%), and staff training (38%). Just below half (46%) had yet to decide how they would spend at least some of this money.

In the initial interviews, heads of maths confirmed this preference for increased teaching time; they described additional staff time being used to extend regular maths sessions or reduce class sizes. Alternatively, it could be spent on support staff, such as attendance officers (see above): poor attendance at revision sessions was seen as a major factor preventing resitters' succeeding.

> *'Our number one biggest barrier by a long way, is attendance. Getting the buy-in from the students, convincing them that this is something that they should be doing, we've got engaging and effective members of staff, but they can only work their magic if the students are in the room' (head of maths, Model A).*

Heads of maths also described hiring 'maths coaches'. These coaches would be in the classroom helping resitters during lessons but also had a pastoral role—supporting resitters with their mental health and resilience as they worked towards the exam.

> *'Quite often lots of different hands keep going up because they get stuck at different stages, which is difficult for a maths teacher to circulate around and troubleshoot where every learner is … so these maths learning coaches are to support with that spinning effect and being there as quickly as possible to support learners when they're facing a problem' (head of maths, Model B).*

Staff time was also devoted to one to one support, including in-class, with the help of teaching assistants specifically for resitters, or outside of class, in targeted tutoring sessions.

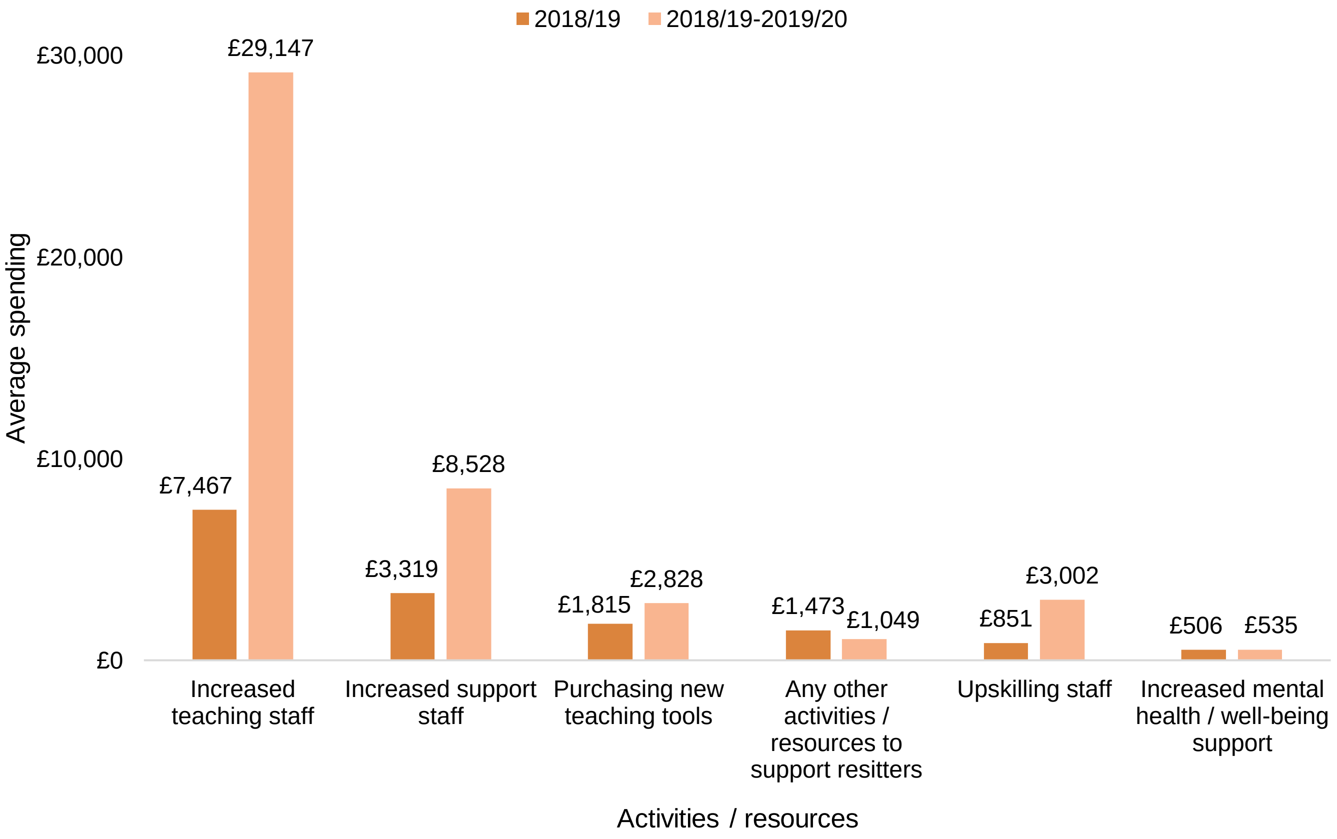
Intervention activities in 2018/2019 to 2019/2020

RQ6 How many contact hours on average did eligible students receive?

Over the two-year period of the programme (2018/2019 to 2019/2020) the activity which attracted by far the most funding was increased teaching staff, with institutions spending on average £29,147 in this area (see Figure 12). The largest sum spent on teaching staff was £600,004. In practice, this meant that in this period 46% of institutions gave their resitters 30 to 90 minutes of additional contact time per week as a result of BMP funding. Another 14% of institutions provided two hours, and a further 14% provided more than two hours.

The next three activities in terms of average money spent were more support staff for maths (£8,528), purchasing new teaching tools (£2,828), and upskilling staff (£3,002).

Figure 12: BMP funding spent, 2018/2019 and 2018/2019 to 2019/2020



Sample sizes: 2018/2019: n = 296; 2018/2019–2019/2020: n = 118.

At the time of the third wave survey in October 2020, 23% of respondents said they still had some BMP funds to be spent (11% said they did not know). Of these 27 respondents, 18 said they would spend it on additional teaching tools, 11 on activities yet to be decided, and eight on increased teaching staff.

Data from the follow-up interviews with heads of maths provides more detail on what institutions were doing with any leftover funding. In terms of staffing, leftover funds were used to reduce class sizes (or maintain their current size) or introduce specialist staff for resitters in either one to one or group settings. Smaller class sizes were considered important to make activities as tailored as possible to the needs of resitters and one to one contact was regarded essential for motivating students who lacked confidence.

> *One to one … relationship is key with Sixth Form … They feel really negative towards the subject, generally, by the time they get to 18 and they're still blooming sitting there for the 947th time … the better the relationship … and the positivity with the teacher can engage, in our experience, those are the areas that we really can have an improvement' (head of maths, Model C).*

At least one institution described using the BMP to help motivate teachers to support resitters. The head of maths at this institution believed that in previous years, when teachers had volunteered their time for resitters, they did so without truly believing it would help resitters pass. Now the institution could pay teachers for these sessions they were more motivated to help resitters towards the grade they needed.

Maths tools tended to focus on providing additional maths practice in a format that was simple and convenient for resitters and the institution. As mentioned above, technology and apps such as MyMaths or iPads for all resitters were a common choice. Conventional tools such as revision guides were also popular, allowing for more tailored and focused preparation for exams. In some cases, paying resitters incentives to attend revision sessions was tried. It was hoped this would improve attendance, similar to the Education Maintenance Allowance payments students received in years past. However, heads of maths pointed out they had not seen evidence that this was effective based on the efforts they had attempted.

Intervention activities phased out in 2019/2020

Institutions had seldom phased out any activities from 2018/2019 to 2019/2020, except in some cases where technology was phased out in favour of alternatives that were more tailored to students' needs. PiXL was one such app that was seen as not customizable enough to the individual student and was replaced by another app called ActiveLearn.

Impact of Covid-19 on usual practice and BMP activities—2019/2020 school year

During the Covid-19 lockdown, some institutions continued teaching online, albeit at a much-reduced level compared to the pre-lockdown timetable. In other institutions, however, heads of maths reported that teaching was largely or entirely suspended over this period. Activities specifically for resitters followed a similar pattern with some institutions suspending them altogether. Where activities did continue, this was in several different ways, for example, institutions might use software suites to supplement their remote teaching or, if they had phased out teaching entirely, these were offered as an alternative. These apps, such as MathsWatch, provided ongoing work for resitters in lockdown, though in some cases this was entirely voluntary. Resitters' participation in activities during the lockdown ranged from almost nothing to as high as 50% of resitters at the institution taking part. Compounding this much reduced (and often non-compulsory) timetable, resitters were described as losing all motivation to continue with their work once it was announced their grades would be centre assessed.

> *'The students, by that point, knew they weren't sitting an exam, and it was a one-year course, so I think they gave up, almost, but yes. It was difficult, but we tried our best' (head of maths, Model B).*

Heads of maths described students experiencing challenges in moving to remote working. Notably, this included lack of access to computers, which in some cases was mitigated by sending home technology like Chromebooks. This was not explicitly covered by BMP.

Responsiveness

This section covers heads of maths' perceptions of the BMP and institutions' engagement with the intervention. It discusses reasons for interest in the BMP, understanding of the intervention, decision-making around the allocation of funding, views on the financial incentive and overall experience of the intervention and suggestions for improvements.

Reasons for interest in the BMP

RQ9 To what extent is the intervention aligned with the objectives and priorities of each school/college?

Institutions were interested in the BMP funding for two main reasons: to help meet their goals and objectives or to solve an existing issue.

Meeting the institution's goals and objectives

Survey findings show that 53% of respondents agreed and 44% strongly agreed that the objectives of the BMP aligned with those of their institution. Some heads of maths, for example, noted that the BMP was aligned with their institution's mission to better support achievement in maths and invest in young people's futures; others noted that the intervention would help to improve the quality of teaching and education—for example, by being able to offer higher salaries—which aligned with the institution's priorities.

Solving existing issues in the institution

Heads of maths were drawn to the intervention as a solution to existing challenges within the institution. Challenges included financial strain from cutbacks, limited staffing capacity, low levels of student achievement, and associated low student retention. Some reported that high proportions of resit students at the institution had highlighted the need to prioritise those students. Others stated that, due to competing pressures at their institution, resit students had not been a top priority, which was an issue they thought should be addressed. The BMP provided an opportunity for institutions to address some of these challenges.

Understanding of the intervention

RQ8 How clear were the objectives of the trial and the funding arrangement?

Understanding of funding arrangements

Understanding of the intervention fell into a broad spectrum. While some were confident in their understanding of the BMP others were unsure about the objectives (see Fidelity) and specific aspects of the funding arrangements.

Allocation of funding model

In the first survey wave, most institutions (97% Model A; 94% Model B; 92% Model C) correctly identified their allocated funding model, with less than 1% unaware of their allocated funding model. In the third survey wave, the proportion of Model B and C respondents who correctly identified their allocated model decreased, while for Model A it remained the same (97% Model A; 65% Model B; 73% Model C). Compared to the first survey wave, a greater proportion of participants (8%) reported that they were unaware of their allocated funding model. Moreover, 20% of Model B and 18% of Model C respondents misidentified themselves as belonging to funding Model A.[34] It is possible that the lower awareness of allocated funding model in the third survey wave is a result of staff turnover.

The interview findings mirrored those from the survey, with more limited awareness of the treatment group for those in funding Models B and C. For instance, one respondent in the payment-by-results model thought they had already received BMP funds. In comparison, Model A had a more complete understanding of the funding arrangements, possibly due to the comparative simplicity of the model.

Amount of funding

The overall amount of funding to be received was clear for 68% of institutions (39% quite clear; 29% very clear). This only varied slightly across funding models.[35] Initial interviews identified some areas of uncertainty around the amount of funding. For example, there was confusion around which resitters were eligible (see Fidelity). Institutions did not know whether the funding applied to apprentices or which cohort of resit students qualified. This meant that some institutions received less funding than expected. Furthermore, some were not clear on the amount of funding that would be received per student. For example, during an initial interview, one participant in Model C thought that the total funds were only £250 per student as opposed to £500. Other participants had no awareness of the amount per student.

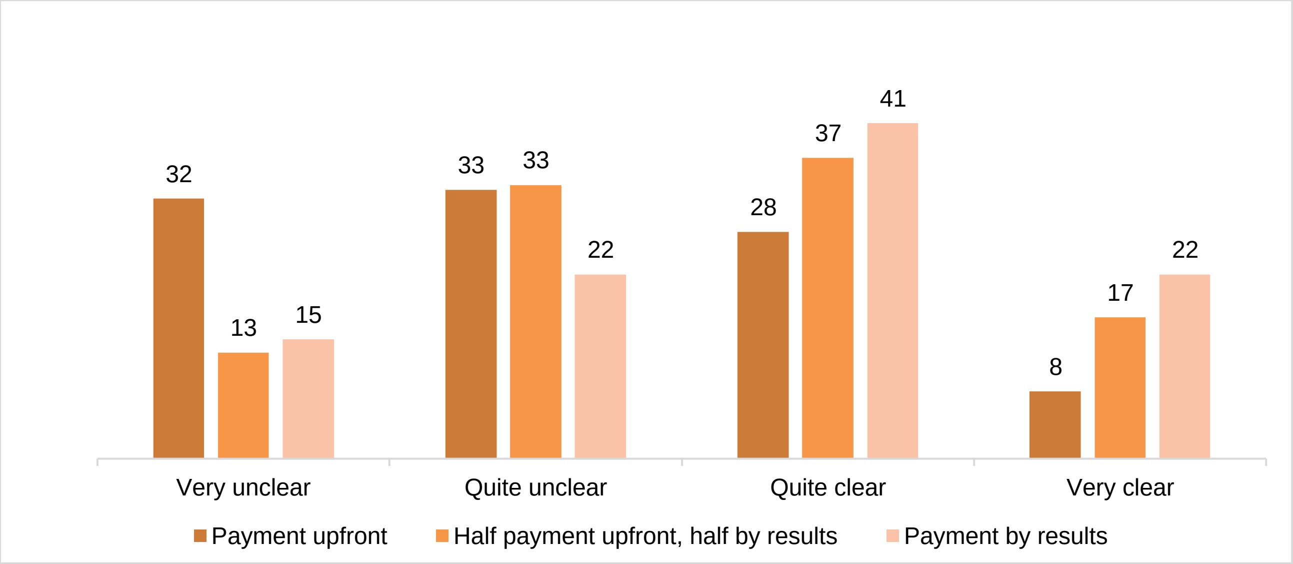
Timing of payments

Less than half of institutions reported that the timing of payments was clear (30% quite clear; 15% very clear). This varied by funding model, with 36% of Model A, 53% of Model B and 63% of Model C respondents reporting that the payment date was quite or very clear (see Figure 13). During interviews, heads of maths in Model A were particularly surprised that the funds were received later in the academic year rather than at the beginning as expected and implied by the term ‘payment upfront’.

[34] In this paragraph, percentages are inclusive of participants who selected ‘don’t know’ options.

[35] Model A: 62% clear (37% quite clear, 25% very clear); Model B: 70% clear (37% quite clear, 33% very clear); Model C: 73% clear (45% quite clear, 28% very clear).

Figure 13: Whether the payment date was clear—by funding model (% institutions)



n: A = 98; B = 93; C = 76. Percentages do not add up to 100% due to rounding.

Clarity of information

While heads of maths generally found the information they received about the BMP to be sufficient, there were some aspects that were perceived to be unclear or limited. Interviewees highlighted that there was little information on whether there were any restrictions on spending. In a small number of cases, guidance was misinterpreted to mean that spending was restricted to certain types of resources. For example, one institution believed that BMP funding could not be used to cover staffing costs.

Of the institutions that found it difficult to decide which resources to spend funding on, 25% reported that it was due to a lack of information about quality resources (see Ease of Decision-Making). Interview findings highlighted a perceived lack of recommendations on how the funding could be allocated. Although the DfE BMP webpage referred to a DfE report on delivering maths to 16- to 18-year-olds and to the EEF toolkit for guidance, participants suggested that more anecdotal evidence of best practice, ideas, and inspiration should be provided.

Cascading of information

RQ12 How well was the information cascaded to teachers/heads of maths?

In the first survey wave, the majority of institutions (71%) reported that information about the BMP had been or would be cascaded to maths teachers by the head of maths; 46% said it was, or was planned to be, cascaded through meetings and 28% via emails. Only 1% of institutions reported that they were not planning to cascade any information about the BMP to maths teachers.

Initial and follow-up interviews revealed that not all heads of maths had received sufficient information about the BMP. Heads of maths noted that information had been received from more senior members of staff or from staff who were no longer at the institution. This meant they had a limited understanding of the intervention. Others expressed the view that information about the BMP should have been directed at business managers as opposed to teaching staff. The information that was cascaded to business managers was often confusing because teaching staff had a more limited understanding of the institution's funding procedures.

Decision-making around the allocation of funding

Decisions around the allocation of funding tended to be made by senior members of staff, most commonly the head of the institution (50%). Spending decisions often involved input from those responsible for the institution's finances (59%). In some cases, a dedicated team was formed to make these decisions; 77% percent said that the head of maths was involved. However, initial and follow-up interviews suggest that heads of maths had varying degrees of input in the decision-making process: while some had complete ownership, others had limited involvement or were not involved at all.

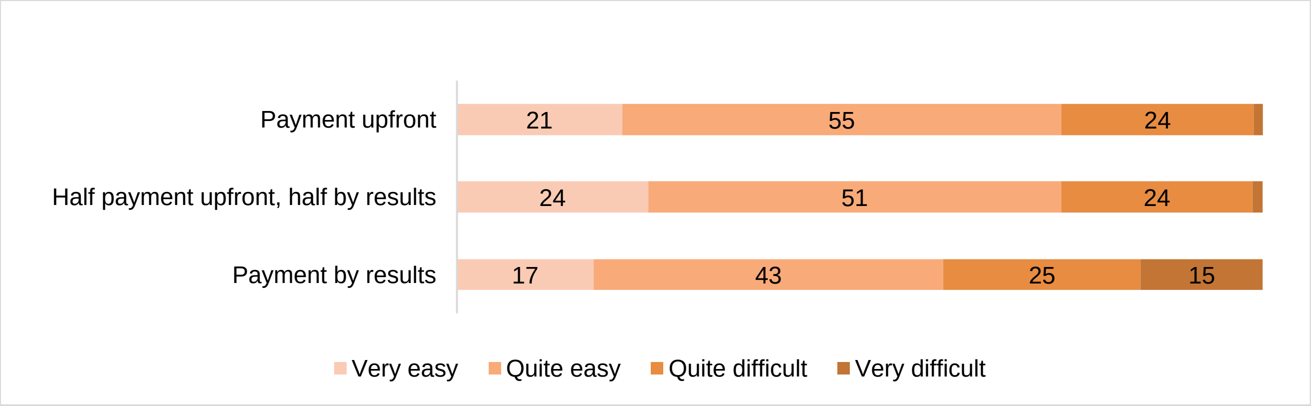
Ease of decision-making

RQ16 Was the time lag between the intervention, the outcome, and the payment acceptable?

Overall, interview participants found making decisions about spending to be relatively straightforward. A key enabler was consensus, which was facilitated by having few decision-makers and reviewing evidence of best practice that those involved in making decisions had sourced (for example, research reports and interventions trialled internally). Having clear areas of weakness that the institution needed to target also facilitated decision-making.

Ease of decision-making varied depending on the funding model. Those in Model C were most likely to experience difficulties with 40% reporting that it was either very or quite difficult to allocate funds. In comparison, only 25% of Model A and 25% of Model B experienced difficulties (see Figure 14).

Figure 14: Whether the allocation of funds was easy or difficult—by funding model (% institutions)



n: A = 88; B = 80; C = 53. One percent of Model B and 1% of Model C reported that allocating funds was 'very difficult'. Percentages do not add up to 100% due to rounding.

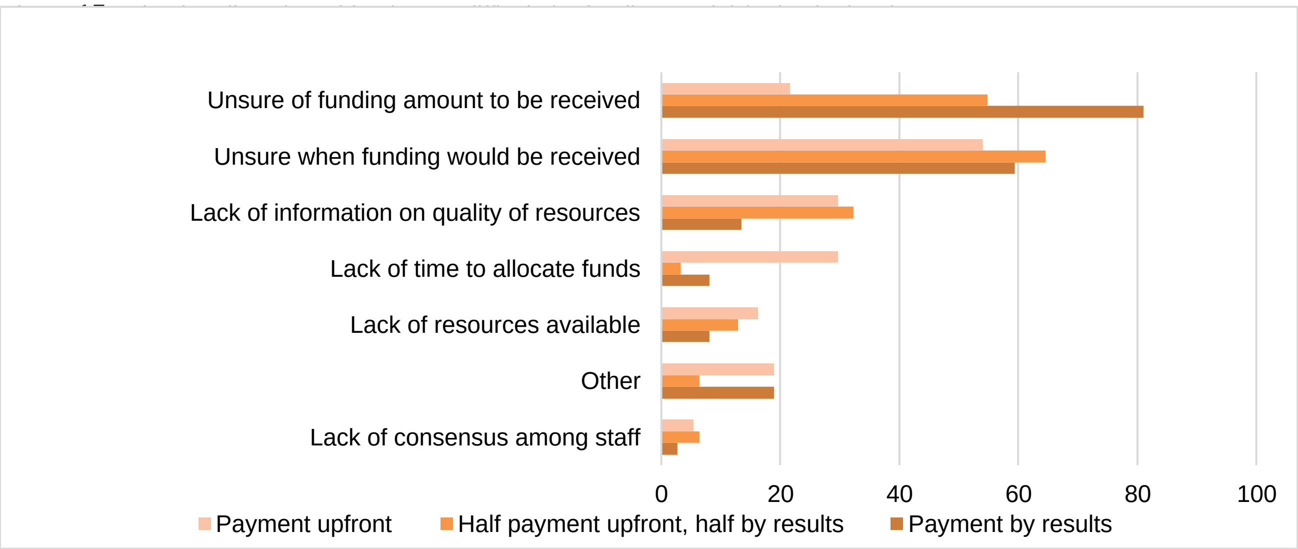
For the institutions that found the process of allocating BMP funding difficult, the two main causes were uncertainty around when the funding would be received (59%) and the amount (52%).

Uncertainty around the amount of funding

In Models B and C, the most prominent cause of difficulty was the uncertainty around the amount of funding that would be received (see Figure 15). This was also reflected in the follow-up interviews, where respondents in Models B and C explained that the uncertainty caused by the payment-by-results component made it difficult to plan, especially when there were high upfront costs involved. While some estimated the amount they would receive and sourced alternative funding to bridge the gap, other institutions held off from spending until funds had been received.

Timing of payments

Uncertainty around the timing of payments was experienced by all funding models (see Figure 15). This uncertainty led to delays in spending (see also Fidelity). This was especially the case for institutions which had a policy that required funds to be received before they could be spent. The timing of the BMP payments caused problems for those in models with a payment-by-results component, with 47% of Model B and 81% of Model C reporting that the timing of payment caused problems compared to 24% in Model A. In the interviews, heads of maths in Models B and C noted that the payment-by-results arrangement was challenging as they needed to source alternative funds to invest upfront, which some found especially difficult.



n = 105; respondents were asked to select up to two reasons.

Other challenges

Heads of maths also discussed a number of additional challenges. It was reported that funds were not always spent on their intended purpose. In some settings, the premium was put into a 'central pot' of money or was placed into a combined pot of funding for the English and maths departments. Heads of maths also experienced delays when spending the funds caused by internal bureaucracy and lengthy sign-off processes. Furthermore, heads of maths explained that care was taken when making decisions to ensure sustainable choices. For example, funds were invested in current GCSE maths students to reduce the number of resit students in the following cohort.

> *'It's about putting things in place, so it lasts three, five, ten years down the line, so we improve every single year, year on year. What is the point of increasing results by 10% one year, then just dropping back the next year?' (head of maths, Model A).*

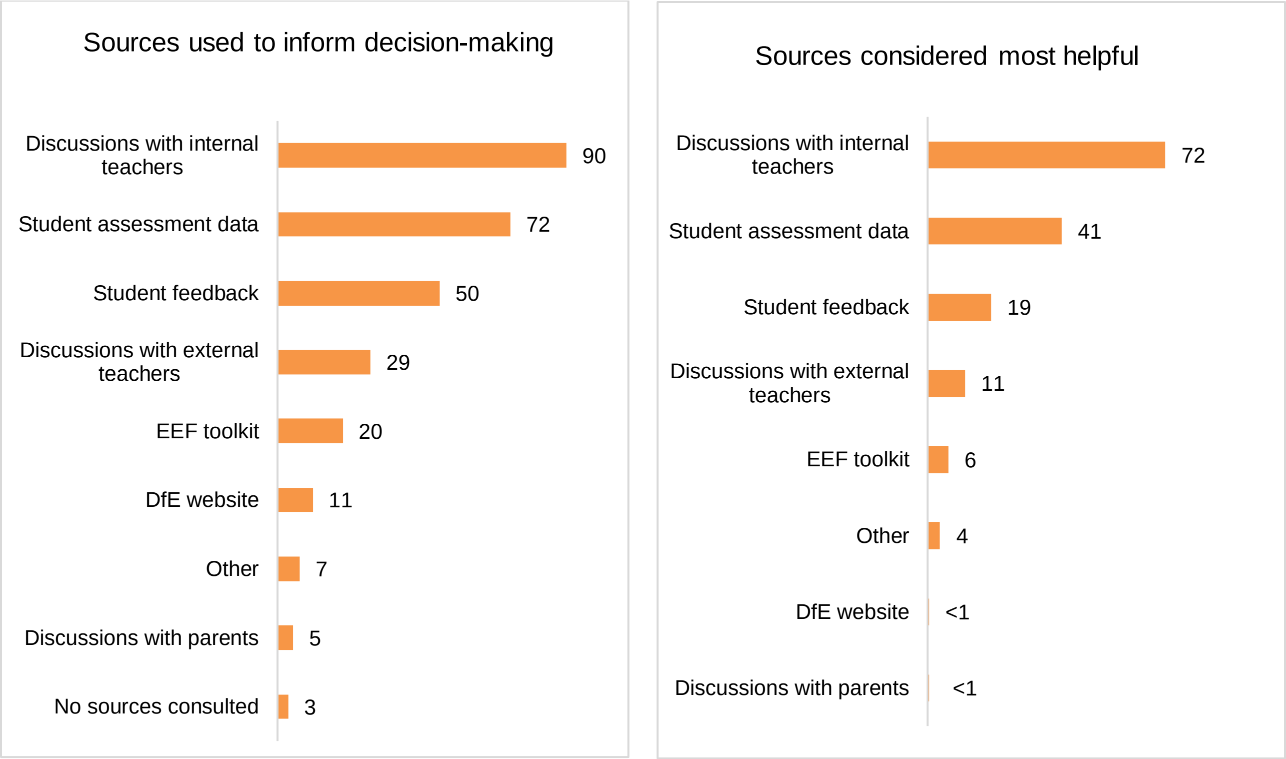
Sources of guidance

RQ5 To what extent did pilot institutions use the guidance on the DfE website?

Institutions consulted both internal and external resources when making spending decisions. The most commonly used were internal. These were discussions with teachers within the institution (90%), student assessment data (72%), and student feedback (50%).

External guidance was used to a lesser degree: institutions reported that they consulted with teachers outside of their institution (29%), the EEF toolkit, which was also suggested as a source of guidance in the DfE BMP webpage (20%), and the DfE website in general (11%). Resources considered most helpful aligned with their popularity except for the DfE website, which was proportionately less likely to be considered helpful, see Figure 16. The sources of guidance considered to be most useful were discussions with teachers within the institution (72%), student assessment data (41%), and student feedback (19%). Findings from interviews were consistent with survey results, with participants highlighting the value of speaking with teachers in both their own and other institutions. The interviews also identified additional sources of guidance such as speaking with experts (for example, maths consultants), consulting social media, and educational websites.

Figure 16: Sources used to inform decision-making and sources considered most helpful (% institutions)



Sources to inform decision-making: n = 246; respondents selected all options that applied. Sources considered most helpful: n = 238; respondents selected up to two options.

There were some institutions that did not consult any guidance: findings from the wave two survey show that 3% did not consult any sources when making decisions. In a small number of cases, interview participants were unaware of any resources that could have been used.

Response to the financial incentive

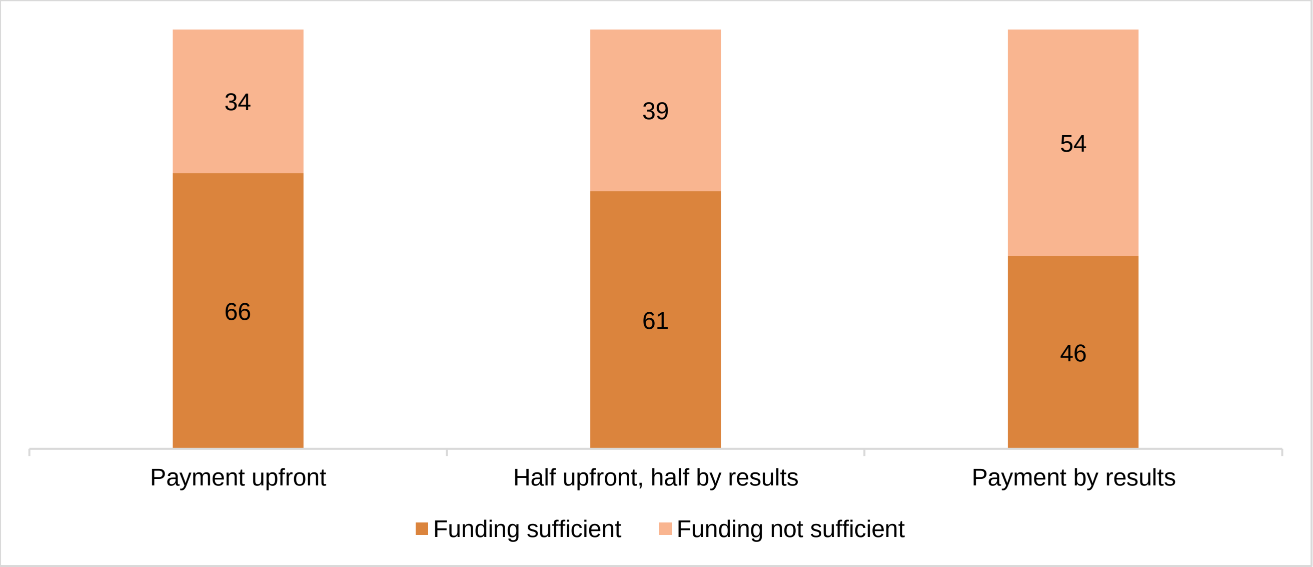
Views on the sufficiency of funding

RQ10 Does the maximum amount (£500/student) seem sufficient to address the problem?

RQ14 How concerned were institutions that the funding received from the DfE would not cover their expenses?

Overall views on the sufficiency of BMP funding were divided, with 58% of respondents reporting that the amount of funding was sufficient to improve maths level 2 resit attainment. Views on the sufficiency of funding varied by model, with institutions in Model A (66%) and Model B (61%) more likely to find the funding sufficient than those in Model C (46%) (Figure 17).

Figure 17: Views on sufficiency of funding—by funding model (% institutions)



n: A = 108; B = 101; C = 87.

One view was that the amount per student should be increased, noting that £500 is too little to make an impact on attainment. This was especially the case for institutions with low numbers of eligible students or where the number of maths resit students varied considerably each year. For example, low numbers of eligible students in the 2018/2019 cohort impacted the following cohorts and might not reflect the institution's needs.

> *'This year, we've got six. That's £3,000. Last time, we had one. That's £500. That's quite a lot of difference … you're not going to employ anyone with £500. You've got a chance of doing something a little bit with £3,000' (head of maths, Model A).*

On the other hand, participants expressed appreciation for any additional funding. This was emphasised by institutions that were under financial strain and viewed any extra money as helpful. In the wave three survey, some respondents in Model C stated they could not comment on the sufficiency of funding as they had yet to receive the payment.

Perception of the funding models

RQ15 *Did implementers understand how 'success' was measured? Did they find it fair?*

RQ16 *Was the time lag between the intervention, the outcome and the payment acceptable?*

RQ13 *To what extent did the intervention help teachers focus on outcomes? How pressured to succeed did they feel?*

Fairness of the model

A large majority of respondents in Model A (92%)[36] perceived their model to be fair, compared to 81% of those in Model B [37] and only 35% in Model C.[38] These views were mirrored by the interview findings. Heads of maths considered Model A the fairest as it did not require sourcing additional funding for upfront investments and provided more certainty around the total amount of funding that could be allocated to additional activities and resources. It was the prominent view that Model C was the least fair—mostly due to the financial difficulties of institutions, which made it more challenging to source funding for upfront costs. On the other hand, one view was that Model C could be effective once schools were able to build up a reserve.

Effectiveness of the model

Across funding models, most institutions (82%) stated that Model A was the most effective in helping to get the best outcomes for maths GCSE or Functional Skills resit students. Respondents in Models B and C stressed that more could

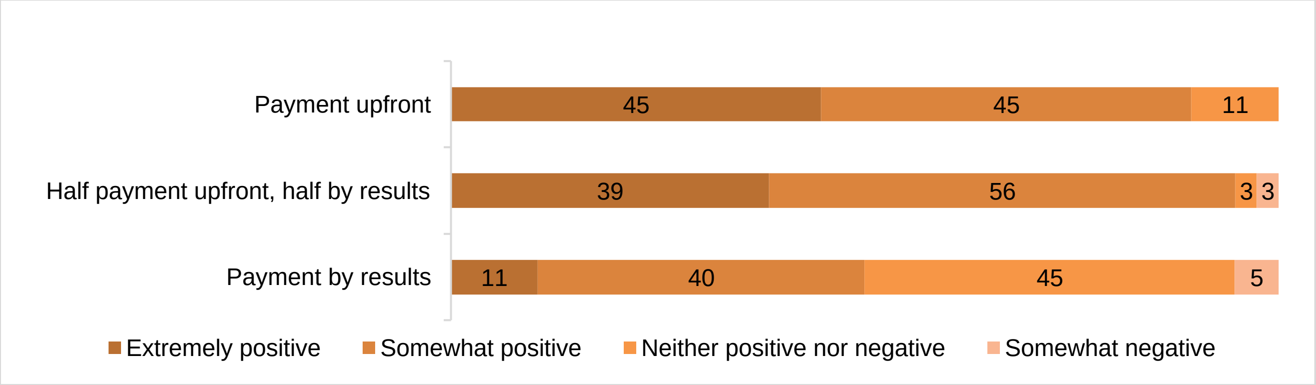
[36] Model A: 30% 'Quite fair' and 62% 'Very fair' (n=98)
[37] Model B: 57% 'Quite fair' and 23% 'Very fair' (n=94)
[38] Model C: 26% 'Quite fair' and 9% 'Very fair' (n=77)

have been implemented with an upfront payment. However, some highlighted that there was also a risk of complacency with this model as none of the funding is dependent on results. Model B was considered to be the second most effective (17%). Those who preferred Model B felt that it struck a good balance between security and incentivisation and that it made them think more carefully about implementing sustainable changes (that is, that benefit both current and subsequent students). Model C was the least preferred with only 1% of respondents perceiving it as the most effective.

Overall experience and suggestions for improvements

Overall, 47% of institutions reported having a ‘somewhat positive’ experience of the BMP and 31% ‘extremely positive’ (see Figure 18). While those in Models A and B reported similar overall experiences, Model C stood out for its high proportion of ‘neither positive nor negative’ responses (45%). Only a small minority of institutions reported having a negative experience, which were concentrated in Models B and C.

Figure 18: Overall experiences of the BMP—by funding model (% institutions)



n: A = 38; B = 39; C = 38. No participants reported an ‘extremely negative’ experience.

Views on eligibility

Some respondents had strong views on who should benefit from the funding. Institutions with large proportions of students resitting lower-level Functional Skills exams believed that it was important to invest in their attainment as well. Another view was that GCSE English resit students should also receive the premium as maths and English were often part of the same department and it was considered unfair to provide extra support for maths but not English.

> *‘I would say, it makes it quite difficult if I have to treat maths and English quite differently because we’re one department, and I think that’s the case with a lot of colleges’ (head of maths, Model B).*

Suggestions and recommendations

Alternative funding arrangements

Institutions that had variable numbers of resit students across cohorts suggested that the allocation of funding to institutions should be made on an application basis. This would allow institutions to specify the amount they need to put the right support in place. Institutions that saw the BMP added to a ‘central pot’ suggested that the funds should be audited by the government in order to avoid other departments spending the funds.

> *‘The more the funding can be ring-fenced, the more likely it is to be spent where it’s supposed to be spent’ (head of maths, Model B).*

Alternative interventions

Some respondents felt the funding would be better spent on adult education rather than young people resitting their maths exams. This would allow adults to achieve maths GCSE or Functional Skills when it was right for them and they were fully motivated. Other respondents expressed the view that increasing funding to FE institutions rather than targeting resit students specifically would be a better investment in terms of attainment.

Evidence of promise

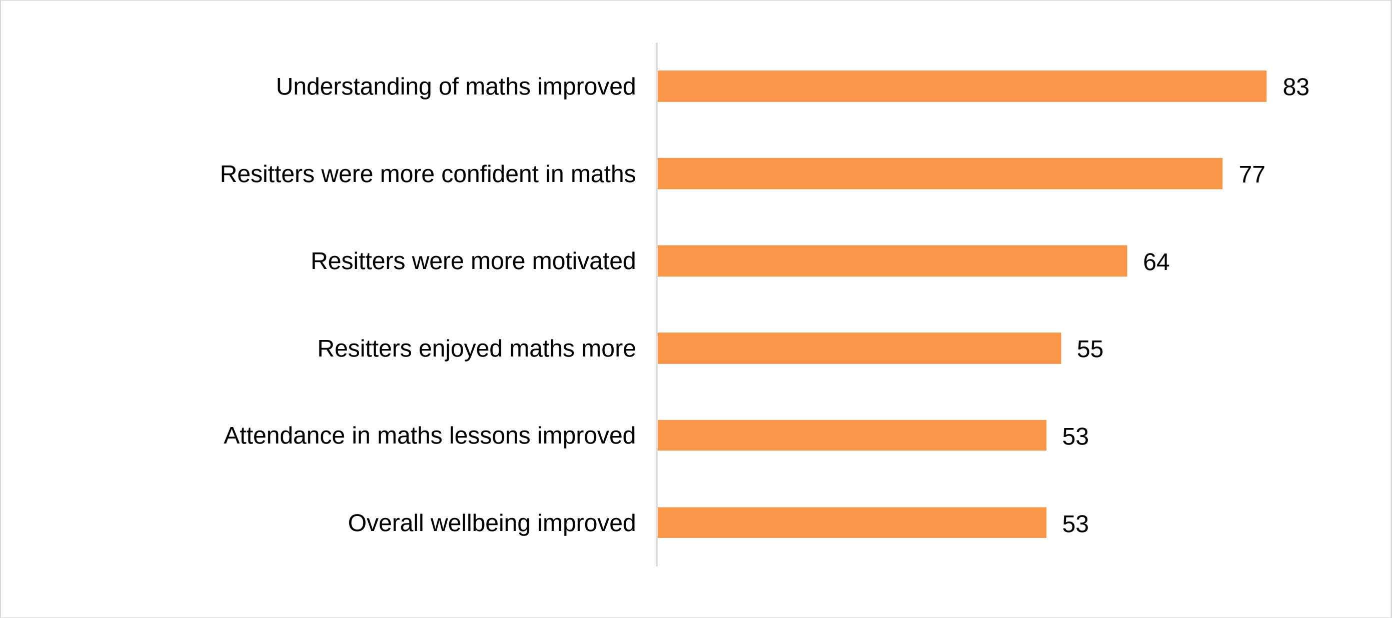
Outcomes for resitters

RQ17 What are the perceived effects on students' soft outcomes?

In the initial interviews, heads of maths expressed an expectation that resitters' grades would improve as a result of BMP funding, although there was more scepticism about this in schools with low spending and low numbers of eligible students. Heads of maths at institutions in Models A and B also observed that attendance had improved and students were more proactive and engaged with maths. In funding Model C, they hoped that BMP would increase students' confidence. This difference in perceived outcomes may be related to differences in mean spending by funding model at the time of these interviews, with institutions in Model C having spent less on average, as shown by the wave one survey (see Fidelity).

In the third survey wave, 86% of respondents reported at least one of the benefits of BMP for resitters included in the survey. Figure 19 shows the percentage of institutions agreeing with each statement. The most commonly reported benefit, selected by the large majority of respondents, was 'resitters' understanding of maths improved as a result of activities funded by the BMP'. A majority of respondents also stated that 'resitters were more confident' and 'more motivated' in maths, and just over half that 'resitters enjoyed maths more' and 'resitters' attendance in maths lessons' and 'overall wellbeing' improved. Institutions in funding Model C were less likely to report any of these benefits (this difference was statistically significant for all perceived benefits except 'resitters' attendance in maths lessons improved').

Figure 19: Benefits of the BMP for resitters (% institutions)



n = 118, of which 101 reported at least one benefit.

These findings are in line with those from follow-up interviews, where a range of 'soft' and 'hard' outcomes of BMP for resitters were mentioned. Heads of maths referred to four 'soft' outcomes: increased confidence, increased motivation, more enjoyment of maths and learning in general, and improved independent study skills. They attributed resitters' increased confidence to realising the progress they made from basic to more advanced maths skills and explained that this was demonstrated by resitters' willingness to attempt questions they would not have approached before.

> *'You can see, in specific groups of students, the confidence is growing. They're applying themselves more, they're not afraid to make a mistake' (head of maths, Model A).*

Heads of maths also thought resitters showed increased motivation as they worked harder and applied themselves more in classes. Resitters' progress from basic to more advanced maths skills and their increased confidence and motivation were perceived to be facilitated by smaller classes and additional teaching and contact time resulting in increased attention to individual learners and support from teachers. Finally, heads of maths pointed to improved independent study skills developed for example by using online learning platforms.

Two main 'hard' outcomes were reported: increased attendance and increased attainment. Heads of maths ascribed increased attendance to having smaller classes and stressed that students were more confident and willing to attend when classes were small. On attainment, one view was that progress and grades (predicated and actual) had improved over the two years of the BMP programme. A contrasting view was that it was difficult to assess the effect of the BMP on attainment as summer 2020 GCSE grades were centre-assessed due to Covid-19.

The BMP was considered to facilitate these outcomes by contributing to better quality teaching. Heads of maths stated that BMP funding enabled resitters to benefit from more support from teaching staff and increased teachers' focus and students' confidence due to smaller class sizes and equipment the institution may not have been able to purchase otherwise. They noted that more support from teaching staff resulted from additional teaching and contact hours, smaller classes, employing more experienced staff, and motivating staff with more opportunities for development and progression.

> *'If they hit it off with that one-to-one or that particular extra day within half-term, it can make all the difference. It can make a fundamental difference between a pass and a fail' (head of maths, Model C).*

Smaller classes and small-group and one to one sessions meant teachers could spend more time on individual learners. This provided an opportunity to know students better and also to focus on their maths foundation skills, which may be overlooked in larger classes including non-resitters.

> *'We were able to take a step back with those students and give them an opportunity to look at their basic foundation skills before moving on' (heads of maths, Model A).*

Mediating factors

Heads of maths' views on whether BMP funding might have benefitted certain resitters more than others varied. Perceived mediating factors included resitters' prior attainment, socio-demographic characteristics, and cohort.

Prior attainment

Views on the influence of prior attainment depended on how this criterion was used to target resitters through BMP-funded activities. Where additional support was given to those with lower attainment, these students were deemed to benefit more. Resitters with lower grades were also considered to benefit more as teachers were better able to identify and target those at risk of failure, and from undertaking Functional Skills examinations. On the other hand, resitters with lower grades and repeat resitters were also thought to be harder to motivate. Where BMP-funded activities targeted resitters close to grade 4 (grade boundary), these were seen instead to benefit more. These students were also considered easier to motivate, and heads of maths noticed that the additional support and smaller classes meant they could progress more quickly as teachers could spend less time on the basics with them.

Socio-demographic characteristics

BMP additional funding was considered especially beneficial for resitters from low-income households; these were thought to have less access to academic support outside of school and be most likely to struggle with confidence. Heads of maths emphasised the importance of the BMP particularly for those who just miss out on the Pupil Premium. There was a view that students with EAL may have benefitted more than others from smaller class sizes and increased attention from teachers. Conversely, it was suggested that resitters with SEND may have benefitted less from the BMP as they already had additional support in place.

Cohort

While some heads of maths did not think cohort affected the extent to which resitters benefitted from the BMP, others reflected that resitters in the second cohort (2019/2020) benefitted more as funding was only spent in that academic year or as everything was already in place by then (although the 2019/2020 cohort was also seen to be affected by Covid-19, as discussed below). Another view was that students who attended in both the 2018/2019 and 2019/2020 academic years benefitted most as they received support over two years. However, it was also noted that it was difficult to compare attainment outcomes between cohorts as summer 2020 GCSE exams were cancelled due to Covid-19 and grades were centre-assessed.

Effect of Covid-19 on outcomes

In the follow-up interviews, heads of maths mentioned a number of perceived negative effects of Covid-19 on 2019/2020 resitters' outcomes. They reported decreased attendance and engagement during school closures as well as a negative effect of school closures on students' mental health. There was also a view that having centre-assessed GCSE grades in 2020 meant students were less motivated to study in order to do well at the exams. Finally, some heads of maths thought GCSE maths exams would have shown more improvement in grades compared to the centre assessments. However, perceived positive effects of the GCSE exams being cancelled were also discussed. These included students being more motivated as grades were based on all the work done rather than a one-off exam and students not having the stress of a high-stakes exam. One view was that exam results were not affected by Covid-19 as grades were based on mock exams which had taken place when students were still in school.

Wider outcomes

RQ19 To what extent did the intervention push teachers to innovate?

Follow-up interviews with heads of maths also highlighted perceived benefits of the BMP for other groups who were not directly targeted by the intervention. These included other students beyond those eligible for the BMP, teaching staff, and the overall institution.

Benefits for other students

As well as resitters eligible for the BMP (that is, those enrolled in 2018/2019), the activities and resources funded through the programme were seen to positively affect future cohorts of maths students and resitters. Heads of maths explained that benefits for all students also resulted from purchasing non-maths-specific technology resources and from smaller class sizes. Smaller class sizes were considered especially important as they facilitated classroom management and meant that teachers could dedicate more time and energy to teaching creatively and effectively.

> *'If you can have 17 people in your class instead of 25, that can make a big difference to how much energy you have to devote to being a creative teacher, rather than just doing marking' (head of maths, Model B).*

Benefits for staff

Perceived benefits for teaching staff included lower workload and pressure, improved opportunities for development and progression (for example, through training and the creation of new roles), and increased confidence.

> *'Relieving the pressure on existing staffing because we could bring more in, which allowed staff to focus on what they do best, meaning the learners get the better experience' (head of maths, Model A).*

Benefits for institution

Institutions were seen to benefit from teaching staff's increased motivation, which translated in improved retention rates.

> *'We've got progression opportunities for staff now, which is a real incentive ... which is also meaning we're retaining staff more' (head of maths, Model A).*

Heads of maths also explained that the BMP contributed to improve best practice by encouraging staff to think about new ways to support resitters, enabling staff to be trained on best practices, and allowing increased financial freedom to bring in new initiatives, teacher incentives, support packages, and department restructurings.

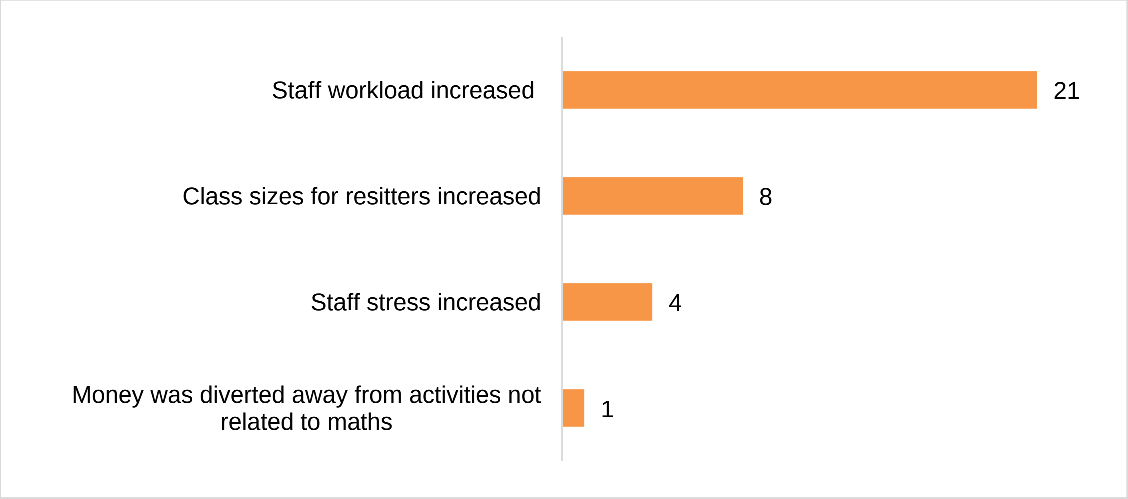
Overall, the BMP was considered important in setting GCSE maths resits as a priority in institutions where there was previously little in place for post-16 students, and useful to contrast low levels of funding in Further Education. Some heads of maths mentioned intending to continue BMP-funded activities and resources once funding ended, for example, by drawing on the regular college budget. However, it was also suggested that while the BMP provided additional funding for institutions, it might not be effective in targeting resit students nor a good investment as this group has especially low pass rates and levels of motivation which are difficult to address.

Unintended consequences

RQ18 What are the unintended effects of the intervention?

In the third survey wave, 31% of respondents stated that they experienced at least one of the challenges of BMP included in the survey. Figure 20 shows the percentage of respondents reporting each challenge. The most common, selected by 20% of institutions, was that 'staff workload increased because of BMP'. A small minority of institutions mentioned 'class sizes increased', 'staff stress increased', and 'money was diverted away from activities not related to maths'. Around 10% of the wave three survey respondents also stated they 'took on additional resitters because of BMP'.

Figure 20: Challenges of the BMP funding (% institutions)



n = 118, of which 36 reported at least one challenge.

During initial interviews, heads of maths expressed concerns that in mixed cohort classes ineligible students might find the intervention unfair. However, this was only reported as a risk rather than something that had been observed. Another concern expressed at that point was that there might be a shift of resources and priorities from other subjects and activities to maths. However, findings from the third survey wave suggest this was not the case. Follow-up interview findings on what heads of maths thought had worked less well with the programme are discussed in the section on Responsiveness.

Cost

Average per student, per year cost to institutions

RQ7 What was the average cost of the intervention to institutions?

As indicated in the evaluation protocol, the cost evaluation compares the average cost of the intervention per funding model, bearing in mind that (1) participating institutions are free to use the funds as they see fit and (2) Models B and C entail a certain amount of financial risk. The intervention consisted in the provision of additional funding for institutions to spend. Therefore, the cost of delivering the programme mainly includes staff time spent on decision-making and administrative tasks around the allocation of BMP funding. The third survey wave asked for an estimate of staff time spent on delivering the BMP and of the cost of delivery (based on staff time and average hourly pay). Across funding models, institutions reported spending an average of 15 additional staff hours over the two years of the programme to deliver the BMP. They stated that this had costed them an average of £2,021. Table 43 to **Error! Reference source not found.** summarise the additional staff time and delivery costs that institutions reported spending on the intervention in the wave three survey.[39]

Table 43: Additional staff time spent delivering the BMP by funding model

Staff time	**Funding model**			
	Guaranteed payment (A)	Half guaranteed, half by results (B)	Payment by result (C)	Overall
Mean	24	14	8	15
SD	33	21	13	24
Min.	0	0	0	0
Max.	150	80	50	150

n = 102 (A = 32; B = 36; C = 34). Outliers were excluded (one respondent reported 1,000 staff hours).

Table 44: Cost of delivering the BMP by funding model

Amount spent (£)	**Funding model**			
	Guaranteed payment (A)	Half guaranteed, half by results (B)	Payment by result (C)	Overall
Mean	£1,728	£2,395	£1,873	£2,021
SD	£2,347	£7,125	£4,674	£5,218
Min.	£0	£0	£0	£0
Max.	£10,000	£40,000	£24,500	£40,000

n = 88 (A = 25; B = 31; C = 31). Outliers were excluded (one respondent reported spending £175,000).

As shown in Table 44, the cost to institutions of delivering BMP consists of a single running cost: staff time. There were no further running costs or any start-up costs. Table 45 translates this into a cost per student, per year over three years.[40] This is calculated as follows.

[39] Please note that these findings should be treated with caution as based on self-reported staff time and costs spent rather than on administrative data.

[40] Although the actual duration of the BMP pilot was two years, we present cost information over a 3-year time horizon in accordance with the EEF cost evaluation guidance, to permit comparability with other evaluations.

Step 1: calculate the total cost of delivering BMP during Year 1.

- In the first year of the pilot, BMP covered a single cohort of eligible students—those who entered post-16 education in 2018/2019 and were eligible for BMP.
- The annual cost to institutions of delivering the pilot to these students can be estimated by dividing the total (running) cost of delivering the pilot to this cohort over its two-year duration by two.

$$Annual\ cost\ of\ BMP\ per\ setting = \frac{Average\ cost\ of\ delivering\ BMP\ over\ 2\ years}{2} = \frac{£2{,}021}{2} = £1{,}010.5$$

Step 2: calculate the per-student cost of BMP during Year 1. This is done by dividing the Year 1 cost calculated above by the average number of eligible students for BMP per institution, in a single cohort.[41]

$$\frac{Annual\ cost\ per\ institution\ of\ delivering\ BMP}{Average\ number\ of\ eligible\ pupils\ per\ institution} = \frac{£1{,}010.5}{123} = £8$$

Step 3: scale the previous costs for Years 2 and 3 to account for the fact that in subsequent years of implementation, an additional cohort of eligible students would be exposed to BMP. That is, after the first year, in each year of BMP implementation there would be two eligible cohorts of students covered by the funding. This implies multiplying the total annual cost per year and the total average number of eligible students per setting by two (assuming a static cohort size each year).

The findings are shown in Table 46.

Table 45: Annual per student, per institution cost to deliver BMP, over 3 years

	Year 1	Year 2	Year 3
Staff time for decision-making and administration (recurring cost)	£1,010.50	£2,021	£2,021
Total Cost per institution per year	£1,010.50	£2,021	£2,021
Number of students per institution per year	123	246	246
Total cost per student, per school, per year	£8.22	£8.22	£8.22

Costs are based on self-reported responses to the third survey wave, which asked responds for an estimate of staff time spent on delivering the BMP and of the cost of delivery (based on staff time and average hourly pay). n = 88 (A = 25; B = 31; C = 31). Outliers were excluded (one respondent reported spending £175,000).

In practice, the BMP pilot lasted for two years, costing around £8 in staff time per year per eligible student. Since the entire cost of delivery relates to running costs (there were no set-up costs), the annual per student cost is the same regardless of whether we examine this over a one-, two-, or three-year time horizon. Therefore, there is no scaling required to convert the annual per student cost over one year calculated in Step 2 into an annual per student cost over three years.

Table 46 also presents an analogous calculation for each funding model individually. We find that costs in terms of staff time are somewhat higher in funding Models B and C. This is consistent with the findings of the Responsiveness section above, indicating that respondents in Model C institutions found it the most challenging to make decisions over how to allocate funds. However, we note that sample sizes are relatively low for this per-funding model analysis and we therefore apply some caution in drawing inference from these results.

Table 46: Per student cost of delivering the BMP

			Year 1	Year 2 and Year 3

[41] To perform this calculation, we used the average number of students as recorded in the original BMP sample frame, averaged over all institutions that participated in the wave three survey.

Item	Type of cost		Annual cost of BMP per institution	Number of students	Total cost per student, per year	Annual cost of BMP per institution	Number of students	Total cost per student, per year
Staff time spent on BMP decision-making and administration	Running cost (school average)	**Overall**	£1,010.5	123	£8	£2,021	246	£8
		Model A	£864	137	£6	£1,728	274	£6
		Model B	£1,197	128	£9	£2,394	256	£9
		Model C	£936	106	£9	£1,872	212	£9

The BMP pilot ran for two years. We report costs over three years for consistency with EEF guidelines. Costs for a hypothetical third year of implementation are therefore a projection based on actual reported costs over the first two years. Running costs in terms of staff time are measured using wave three of the institution survey. Student numbers used to derive the per-student costs are taken from the BMP sample file and assume an additional student cohort receiving the BMP in Years 2 and 3. Note that the cost data available from the institutions survey captures only the total costs of delivery across the two years of the pilot, which we have annualised to estimate the yearly cost. In practice, it is plausible that costs in terms of staff time may drop year-on-year as settings become increasingly familiar with the programme and gain experience. However, we do not have specific data available to establish whether this is indeed true.

Average per student cost to DfE

In addition to the costs of BMP for institutions presented above, we also explore here the costs of the programme for DfE in terms of the amount of funding provided per student. We do not have access to DfE data on the amounts of BMP funding actually provided over the course of the pilot; however, we can present a simple estimate of the average per student cost by calculating total spend according to the intended funding allocations under each model, and the number of eligible students.

For Model A this is simply £500 per eligible student:

$$\frac{\#\, model\ A\ students * £500}{\#\, model\ A\ students}$$

For Model B this is:

$$\frac{\#\, model\ B\ students * £250 + (\#\, model\ B\ students\ that\ achieved\ a\ pass) * £250}{\#\, model\ B\ students}$$

For Model C this is:

$$\frac{(\#\, model\ C\ students\ that\ achieved\ a\ pass) * £500}{\#\, model\ C\ students}$$

An overall cost can then be calculated as:

$$\frac{(avg\ cost\ per\ A\ student\ * \#A\ students) + (avg\ cost\ per\ B\ student * \#B\ students) + (avg\ cost\ per\ C\ student * \#C\ students)}{\#\, model\ A\ students + \#\, model\ B\ students + \#\, model\ C\ students}$$

Since the impact evaluation results did not find any evidence of differences in average attainment between funding arms, we use a constant achievement rate of 12% to estimate the number of eligible students in Model B and Model C schools that achieved the required pass (as per Table 15). We use the original sample frame dataset to calculate the number of eligible students per participating institution. Findings are reported in Table 47. This shows that the amount of money disbursed per student in Model C is, on average, substantially lower than that per student in Model A due to low numbers of students achieving the required pass.

Table 47: Estimated cost per student of BMP for DfE

	Overall	**Model A**	**Model B**	**Model C**

Average per student cost of BMP financial incentives	£279	£500	£280	£60

Underlying calculations are as follows:

Model A

Total number of schools: 140 (see Figure 2); average number of eligible students per school: 116 (this is based on all Model A schools in the original sample frame data for which a number of eligible students was available, comprising 135 records).

Model B

Total number of schools: 145 (see Figure 2); average number of eligible students per school: 113 (this is based on all Model B schools in the original sample frame data for which a number of eligible students was available, comprising 143 records). Using an average attainment rate of 12%, we have average per student cost of [(113*145*£250) + (113*145*0.12*£250)]/(113*145) = £280 per student.

Model C

Total number of schools: 149 (see Figure 2); average number of eligible students per school: 110 (this is based on all Model C schools in the original sample frame data for which a number of eligible students was available, comprising 142 records). Using an average attainment rate of 12%, we have average per student cost of (110*149*0.12*£500)/(110*149) = £60 per student.

Conclusions

Table 48: Key conclusions

Key conclusions
1. There was no evidence that BMP had an impact on the likelihood that eligible students passed a level 2 maths resit exam by November 2019. This is the case for BMP as a whole, and for each funding model separately.
2. There was no evidence of a differential impact of BMP on level 2 maths resit attainment according to students' previous attainment, free school meals eligibility, institution size, or the amount of BMP funding spent by participating institutions.
3. There was no evidence that BMP had an impact on the likelihood that eligible students sit a GCSE exam rather than a Functional Skills Level 2 exam or no maths resit at all.
4. Perceived benefits for students included improved understanding of maths, confidence and motivation, increased attendance in lessons and attainment through additional teaching and contact hours, smaller classes, employing more experienced staff, and motivating staff through opportunities for development and progression. Perceived benefits for teaching staff included lower workload and pressure, improved opportunities for development and progression, and increased confidence.
5. Models A and B were affected by severe implementation issues as 'guaranteed' payments were delivered late; ultimately not disbursed until March 2019. Settings in Model C reported that payment-by-results made it harder to finance support upfront and uncertainty about how much funding settings would receive made it difficult to allocate and spend resources.

Impact evaluation and IPE integration

Evidence to support the logic model

This evidence provided by this trial is mixed: moderate to strong evidence was found relating to some areas of the BMP logic model whereas in other areas the evidence was weak.

Areas of the logic model with strong or moderate evidence

Resources

The BMP pilot delivered the intended amounts of funding to participating institutions, although with significant delays to guaranteed payments. All eligible institutions that agreed to take part in BMP were successfully randomised to one of three possible intervention models and received an amount of funding that was in line with their treatment condition. Heads of maths generally reported finding the information they received about the BMP to be sufficient.

Use of funding—supported activities

Settings chose to spend BMP funding on a variety of measures that were anticipated in the logic model. The most commonly cited were to increase teaching staff and to increase support staff. In follow-up interviews conducted in October 2020, heads of maths also reported using leftover funding to help reduce class sizes or introduce specialised staff to support resitters in one to one or group settings.

Outcomes for target students

The IPE found that participating settings felt positively about the potential for BMP to lead to beneficial outcomes. The majority of respondents in the final IPE survey wave reported at least one benefit for resitters. The most common given was 'understanding of maths improved as a result of activities funded by the BMP'. A majority of respondents also stated that 'resitters were more confident' and 'more motivated' in maths, and just over half that 'resitters enjoyed maths more' and 'resitters' attendance in maths lessons' and 'overall wellbeing' improved. Heads of maths at institutions in Models A and B also observed that attendance had improved and that students were more proactive and engaged with maths.

Other outcomes

The IPE also found some evidence of perceived improvements in teacher- and institution-level outcomes. Perceived benefits for teaching staff included lower workload and pressure, improved opportunities for development and progression (for example, through training and the creation of new roles), and increased confidence. Institutions were seen to benefit from teaching staff's increased motivation, which translated in improved retention rates.

Assumptions

Heads of maths were generally supportive of the pilot programme, noting that BMP aligned with objectives already held by their institution and that there had previously been little in place to support post-16 students resitting maths. Heads of maths also reported a view that BMP helped address some existing challenges, such as limited staffing capacity, financial strain from cutbacks, low levels of student achievement, and associated low retention.

Areas of the logic model with mixed or weak evidence

Resources

The guaranteed payments made under Models A and B were provided later than planned. Upfront payments should have been paid at the start of the 2018/2019 academic year but they were not disbursed until March 2019. This raised concerns that funding would be used to support students in the following cohort, who entered post-16 education in 2019/2020. However, a limitation of this trial is that it was not possible to evaluate any impact of the BMP pilot on students who enrolled in post-16 education in 2019/2020 due to the impact of Covid-19 (see Limitations section).

Use of funding—supported activities

The IPE did not find evidence that BMP funding was spent on some of the measures anticipated in the logic model, such as parent engagement activities and travel allowances for students. Other measures that had been trialled were not considered effective, such as financial incentives for students. While there was evidence of spending on mental health and wellbeing support, the perceived effects of this were not discussed. IPE findings also showed that BMP funding was not used to its full extent by all institutions in the two years of the intervention. Spending was the lowest for institutions in Model C: wave two of the survey (administered in November 2019) found that these had spent an average (mean) of 16% of their maximum possible funding while institutions in Models A and B had spent 60% and 41% respectively.

Use of funding—timing and allocation issues

Respondents allocated to the payment-by-results model reported finding it difficult to allocate funds due to uncertainty over how much money the setting would ultimately receive, and a need to source alternative funds to make upfront investments. In this context, institutions raised that they preferred funding to be provided upfront and found the payment by results models unfair due to the financial risk associated with spending money that they might not receive.

Another barrier to spending the BMP funds was uncertainty over the timing of payments, which was experienced by all models. Respondents in Model A were the least likely to report that the payment date was clear, which is consistent with the delays in disbursement of the upfront payment.

Outcomes for target students

The positive perceptions of BMP impact found in the IPE were not reflected in the impact evaluation results, the majority of which were statistically indistinguishable from zero. This was the case for the overall BMP intervention as well as for each funding model in isolation. We also did not find any evidence of differential impacts for different subgroups. The impact evaluation did return two positive results from the quantitative impact estimation—one for a robustness check of the overall impact of the pilot on student attainment and the other for a robustness check of its impact on the likelihood of eligible students being entered for a GCSE exam—however, these positive results are modest in magnitude and were not consistent with findings from the other analysis models carried out. The QED methodology was also affected by some limitations, discussed below, which weaken our confidence in the strength of this evidence.

Other outcomes

The IPE found examples of BMP funding being used to support students outside the intended target group, such as students sitting their GCSE maths exams for the first time and on other institutions within multi-academy trusts. One reason for this was a lack of clarity among some heads of maths about who the intended target group were, and a desire to help reduce the number of students that would need to resit maths the following year. Some IPE respondents also noted that resitters entering post-16 education in the 2019/2020 academic year benefited more from BMP than those directly exposed to BMP in 2018/2019.

Assumptions

There was mixed evidence for whether the amount of BMP funding was enough to support improved maths attainment: 58% of respondents interviewed for the IPE agreed that the funding was sufficient, with those in Model C the least likely to agree.

Interpretation

Overall, the impact evaluation findings indicate that the provision of additional funding in the form of a student premium does not translate into an improvement in maths level 2 resit attainment among students resitting these exams in post-16 education. The IPE findings show mixed evidence on the results of the intervention.

The lack of impact found by this pilot should be interpreted in the context of some methodological limitations (set out in the following section) and some challenges to delivery. Key implementation difficulties for this pilot included a seven-month delay in disbursing guaranteed payments and challenges experienced by schools in being able to allocate and spend money assigned on a payment-by-results basis. These issues may provide a partial explanation as to why no impact was found on student outcomes, despite the perception of heads of maths that BMP was in line with their objectives and helped to address an otherwise unmet need for support.

The impact evaluation results from this pilot do not confirm a link between education spending and attainment that has been demonstrated in other studies, such as Nicoletti and Rabe (2012) and OECD analysis of PISA results (OECD, 2012). Our results are more in line with findings from a recent systematic review of studies analysing the impact of cash transfer programmes on education outcomes (Baird et al., 2014), which only found a small impact of less than 0.1 standard deviations on test scores.

This pilot adds to a nascent body of evidence on the impact of PBR models. The findings from the IPE point to some evidence of ‘creaming and parking’, whereby funding support was prioritised toward students considered to be very close to getting a pass at the expense of learners who were considered less likely to achieve a pass.

We can also reflect on the results of this pilot in relation to the ten ‘success factors’ identified by the NAO for PBR schemes (see the Background section). The trial findings showed that Model C, the full PBR funding arm, was not effective. This is consistent with a failure of one of those factors: ‘providers are likely to respond to financial incentives’. The IPE results found evidence that respondents from Model C found it difficult to respond to the financial incentives provided under Model C due to a difficulty in financing upfront payments and uncertainty over how much funding they would ultimately be eligible for. These difficulties are reflected in the finding that settings from Model C were found to have spent the lowest amount of allocated BMP funding, compared to the other two models that included a guaranteed payment component.

Limitations and lessons learned

The evaluation was affected by several limitations.

Effects of Covid-19

The onset of Covid-19 led the evaluation team to modify the methodology in ways that may have constrained the ability of this trial to detect a positive impact of the BMP.

First, Covid-19 led to a decision to shorten the time horizon over which outcomes were evaluated. The original intention was to analyse the impact on maths level 2 resit outcomes up to summer 2020. As a result of Covid-19, GCSE exams were cancelled in summer 2020 and grades were instead awarded to students based on their teachers’ assessment. Due to concerns about potential inconsistency in grading for the summer 2020 resits, a joint decision was taken to instead focus the impact analysis on student achievement up to November 2019. This change could mean that the impact evaluation did not assess BMP for long enough to detect impacts that might have otherwise emerged by summer 2020.

An implication of this decision was that we also excluded the cohort of students who started post-16 education in 2019/2020 from the impact analysis. This is because the removal of the summer 2020 exam period from the methodology meant that the duration of the impact analysis was no longer long enough to detect any possible impacts for this cohort.

Difficulty in establishing a comparison group

The quasi-experimental methodology was not successful at identifying a strong comparison group for settings and students exposed to BMP. The analysis models to some extent reduce differences between the intervention and comparison groups but are not able to produce an entirely well-balanced analysis sample. For the primary weighting analysis and two-stage matching robustness check, achieving an acceptable level of covariate balance requires removing a large number of BMP settings from the analysis sample. This weakens statistical power as well as the extent to which the findings can be considered representative of all BMP settings. A key reason for this is that BMP settings were allocated to areas with a different AEA classification to all areas from which the comparison group was drawn, since BMP was deliberately targeted towards areas with higher need. This fundamentally limits the extent of overlap in institution- and area-level covariates between BMP institutions and those in the available comparison group. In other words, it becomes very difficult to identify post-16 institutions that can be considered as similar to BMP settings in all respects apart from in their exposure to the intervention.

Attrition from the trial

Participants were lost from the trial at various stages. First, some institutions left the trial between randomisation and the signing of MOUs. Among the 824 institutions initially randomised, 53% (469) agreed to participate. Among the 434 institutions remaining in the trial, only 424 were successfully linked to student records from our NPD extract. This meant that there were ten BMP settings that were removed from the analysis as we were unable to find any eligible Year 12 students in our NPD data extract. Attrition from a trial can lead to various implications: for this trial, however, we do not find evidence that attrition has led to substantial limitations.

One inevitable implication of attrition is reduced statistical power to detect changes in outcomes. In this trial, sample sizes remain large even after adjusting for attrition. This means that updated power calculations that account for attrition continue to show favourable estimated effect sizes.

A second implication of attrition is that it affects the interpretation of the QED estimates that form the majority of the impact estimation conducted for the trial. The QED analysis for this trial involves constructing a non-BMP comparison group that resembles the BMP intervention group, after attrition, as closely as possible. This means that the final impact estimates reflect the average impact of BMP funding for the settings and students that remain in the trial after attrition. To the extent that attrition changes the average attributes of the intervention group sample, this will have an influence on the interpretation of the QED estimates. The characteristics of the BMP intervention group to which our impact estimates relate are described in detail in the Impact Evaluation Results: Student and School Characteristics section.

There is one analysis in this report that directly invokes the randomised allocation of the BMP funding models. This is a robustness check for RQ3 of the IE, exploring the relative effectiveness of the three randomly allocated funding models. For this experimental analysis, attrition introduces a risk of bias. This would occur if there were factors determining whether a setting withdrew from the trial that are systematically related to both treatment allocation and final outcomes. We cannot rule out the risk that attrition does lead to bias for this piece of additional reporting. However, we note that levels of attrition are very similar across the three funding models. This allays concerns about this risk as it suggests that differential attrition across the three funding models has not occurred.

Future research and publications

The pilot would merit further consideration if issues around the late disbursement of guaranteed funds can be addressed. We would suggest a randomised methodology for a future trial to enable a robust assessment of the programme's impact. The eligibility criteria for the BMP intervention in this case precluded the ability to identify a robust comparison group to form the basis of strong QED estimation.

References

Abadie, A. and Gardeazabal, J. (2003) ‘The Economic Costs of Conflict: A Case Study of the Basque Country’, *American Economic Review*, 93 (1), pp. 113–132.

Abadie, A., Diamond, A. and Hainmueller, J. (2010) ‘Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program, *Journal of the American Statistical Association*, 105 (490), pp. 493–505.

Abadie, A., Diamond, A. and Hainmueller, J. (2015) ‘Comparative Politics and the Synthetic Control Method’, *American Journal of Political Science*, 59 (2), pp. 495–510.

Abadie, A. and Spiess, J. (2022) ‘Robust post-matching inference’, *Journal of the American Statistical Association*, 117 (538), pp. 983–995.

Alferes, V. R. (2012) ‘Methods of Randomization in Experimental Design’, in *Quantitative Applications in the Social Sciences*, Sage (p. 171f).

Angrist, J. D. and Pischke, J. (2009) *Mostly Harmless Econometrics: An Empiricist’s Companion*, Princeton: Princeton University Press.

Arpino, B. and 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), pp. 2074–091.

Austin, P. (2011) ‘An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies’, *Multivariate Behavioral Research*, 46 (3), pp. 399–424.

Austin, P. (2011) ‘Optimal Caliper Widths for Propensity-Score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies’, *Pharmaceutical Statistics*, 10, pp. 150–161.

Austin, P. (2014) ‘A Comparison of 12 Algorithms for Matching on the Propensity Score’, *Statistics in Medicine*, 33, pp. 1057–069.

Austin, P., Grootendorst, P. and Anderson, G. (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), pp. 734–753.

Baird, S., Ferreira, F., Özler, B. and Woolcock, M. (2014) ‘Conditional, Unconditional and Everything in Between: A Systematic Review of the Effects of Cash Transfer Programs on Schooling Outcomes, *Journal of Development Effectiveness*, 6 (1), pp. 1–43.

Baser, O. (2006) ‘Too Much Ado About Propensity Score Models? Comparing Methods of Propensity Score Matching’, *Value in Health*, 9 (6).

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’. Overseas Development Institute.

Caliendo, M. and Kopeinig, S. (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: https://publications.parliament.uk/pa/cm201516/cmselect/cmpubacc/327/327.pdf

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009) ‘Dealing with Limited Overlap in Estimation of Average Treatment Effects’, *Biometrika*, 96 (1), pp. 187–199.

Cummings, P. (2011) ‘Estimating Adjusted Risk Ratios for Matched and Unmatched Data: An Update’, *Stata Journal*, 11 (2), pp.290–298.

DCLG (2015) ‘Qualitative Evaluation of the London Homelessness Social Impact Bond: Second Interim Report’, London: Department for Communities and Local Government.

Demidenko, E. (2007) ‘Sample Size Determination for Logistic Regression Revisited’, *Statistics in Medicine*, 26, pp. 3385–397. http://dx.doi.org/10.1002/sim.2771

DfE (2017) ‘School Funding and Pupil Outcomes: a Literature Review and Regression Analysis’, London: Department for Education.

Dominguez-Reig, G. and Robinson, D. (2019) ‘Education Policy Institute: 16–19 Education Funding Trends and Implications’. 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 (2018) 'The Attainment Gap 2017', London: Education Endowment Foundation. **https://educationendowmentfoundation.org.uk/support-for-schools/bitesize-support/closing-the-attainment-gap**

EEF (2022) 'Statistical Analysis Guidance for EEF Evaluations', London: Education Endowment Foundation.

Fox, C. and Albertson, K. (2012) 'Is Payment by Results the Most Efficient Way to Address the Challenges Faced by the Criminal Justice Sector?', *Probation*, 59 (4), pp. 355–373.

Funk, M. J, Westreich, D., Wiesen, C., Stürmer T., Brookhart, A. and Davidian, M. (2011) 'Doubly Robust Estimation of Causal Effects', *American Journal of Epidemiology*, 173 (7), pp. 761–767.

Hong, G. and Raudenbush, S. W. (2006) 'Evaluating Kindergarten Retention Policy', *Journal of the American Statistical Association*, 101 (475), pp. 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. and Vignoles, A. (2006) 'Estimating the Relationship Between School Resources and Pupil Attainment at GCSE', Department for Education and Skills, Institute of Education.

King, G., Lucas, C. and Nielsen, R. (2017) 'The Balance-Sample Size Frontier in Matching Methods for Causal Inference', *American Journal of Political Science*, 61 (2), pp. 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, pp. 59–82.

Lee, D. and Lemieux, T. (2010) 'Regression Discontinuity Designs in Economics', *Journal of Economic Literature*, 48, pp. 281–355.

Li, F., Zaslavsky, A. M. and Landrum, M. B. (2013) 'Propensity Score Weighting with Multilevel Data', *Statistics in Medicine*, 32 (19), pp. 3373–387.

Linde, A. and Yarnold, P. (2016) 'Combining Machine Learning and Matching Techniques to Improve Causal Inference in Program Evaluation', *Journal of Evaluation in Clinical Practice*, 22 (6), pp. 868–874.

Lunceford, J. K. and Davidian, M. (2004) 'Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study', *Statistics in Medicine*, 23 (19), pp. 2937–960.

Lunt, M. (2013) 'Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance with Propensity Score Matching', *American Journal of Epidemiology*, 179 (2), pp. 226–235.

McConnell, B. and Vera-Hernandez, M. (2015) 'Going Beyond Simple Sample Size Calculations: a Practitioner's Guide', London: Institute for Fiscal Studies. https://www.ifs.org.uk/publications/7844

McCrary, J. (2008) 'Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test', *Journal of Econometrics*, 142 (2), pp. 698–714.

NAO (2015) 'Outcome-Based Payment Schemes: Government's Use of Payment by Results'. London: National Audit Office.

Nicoletti, C. and 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. and 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

OECD (2012), 'Does Money Buy Strong Performance in PISA?', PISA in Focus 1. **https://www.oecd.org/pisa/pisaproducts/pisainfocus/49685503.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', Coventry: Office of Qualifications and Examinations Regulation. 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'. 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. and 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), pp. 163–189.

Ritchie, J., Lewis, J., Nicholls, C. M. and Ormston, R. (eds) (2013) *Qualitative Research Practice: A Guide for Social Science Students and Researchers*, London: Sage.

Rodeiro, C. V. (2018) ‘Which Students Benefit from Retaking Mathematics and English GCSEs Post-16?’, *Research Papers in Education*, 20 (3), pp. 245–270.

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

Rosenbaum P. R. and Rubin D. B. (1983) ‘The Central Role of the Propensity Score in Observational Studies for Causal Effects’, *Biometrika*, 70, pp. 41–55.

Schochet, P. Z. (2013) ‘Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference’, *Educational and Behavioral Statistics*, 38 (3). http://journals.sagepub.com/doi/pdf/10.3102/1076998611432176

Sianesi, B. (2004) ‘An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s’, *Review of Economics and Statistics*, 86 (1), pp. 133–155.

Stuart, E. A. (2010) ‘Matching Methods for Causal Inference: A Review and a Look Forward’, *Stat Sci* 25 (1), pp. 1–21.

StataCorp (2021) Stata Statistical Software: Release 17, College Station, TX: StataCorp LLC.

StataCorp (2023) *Stata Causal Inference and Treatment-Effects Estimation Reference Manual. Release 18*. College Station, TX: Stata Press.

Thoemmes, F. J. and West, S. G. (2011) ‘The Use of Propensity Scores for Nonrandomized Designs with Clustered Data’, *Multivariate Behavioral Research*, 46 (3), pp. 514–543.

Webster, R. (2016) ‘Payment by Results: Lessons from the Literature’. http://russellwebster.com/PbRlitreview.pdf

Work and Pensions Committee. (2011) ‘Work Programme: Providers and Contracting Arrangements’, London: The Stationery Office.

Zubizarreta J. R. and 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), pp. 547–560.

Appendix A: EEF cost rating

Figure 21: Cost Rating

Cost rating	Description
£ £ £ £ £	*Very low*: less than £80 per student per year.
£ £ £ £ £	*Low*: up to about £200 per student per year.
£ £ £ £ £	*Moderate*: up to about £700 per student per year.
£ £ £ £ £	*High*: up to £1,200 per student per year.
£ £ £ £ £	*Very high*: over £1,200 per student per year.

Appendix B: Security classification of trial findings

OUTCOME: Likelihood of passing level 2 maths resit

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

Threats to validity	Risk rating	Comments
Threat 1: Confounding	High	Significant residual confounding noted in the report, due to pronounced area-level differences between BMP and non-BMP institutions.
Threat 2: Concurrent Interventions	No information	Among BMP settings, 18% reported receiving additional funding alongside the BMP to support improving attainment at GCSE. The IPE did not include comparison settings so no information is available on concurrent interventions or the extent to which business as usual provision differs between groups.
Threat 3: Experimental effects	Low	Funding was only available to institutions in Achieving Excellence Areas (AEA) categories 5 and 6 so no opportunity for spillover to comparison institutions in AEA category 4.
Threat 4: Implementation fidelity	High	Significant delays in payment to Groups A and B resulting in funding being spent on the subsequent student cohort from the one ultimately used to generate the primary impact estimate. Covid-19 also disrupted provision and prevented student outcomes from being measured in summer 2020, which shortened the evaluation window. Both factors likely lead to an underestimation of treatment impact.
Threat 5: Missing Data	Low	Although attrition in the experimental arms is very high (>50% of settings), comparison group matching was conducted on only the final analysis sample and is therefore unbiased by missing values. Incidence of missingness for both outcome and covariate data was low.
Threat 6: Measurement of Outcomes	Low	Analysis uses GSCE and Functional Skills attainment from national administrative datasets; teacher-assessed grades used during the Covid pandemic were excluded from the evaluation due to lack of comparability with previous exam scores.

Threat 7: Selective reporting	Low	Analysis is comprehensively reported and deviations from the study plan well explained. Interpretation of findings is appropriately cautious.

- **Initial padlock score:** 3 Padlocks – Quasi-experimental comparison group design dealing with observable selection issues, powered to observe an overall MDES of 0.1. Although 49% of settings withdrew from or could not be included in the evaluation after being randomised, as this attrition occurred prior to generating the matched comparison group the peer reviewers felt it was more appropriately regarded as recruitment attrition and not a risk to internal validity.
- **Reason for adjustment for threats to validity**: 2 Padlocks – Two high threats to validity: residual confounding and delayed BMP payments which lead to concerns about dilution of impact.
- **Final padlock score:** initial score adjusted for threats to validity = 1 Padlock

Further appendices:

Technical appendices are published as a separate document.

Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn