EVALUATION OF YEAR 1 OF THE ACADEMIC MENTORING PROGRAMME: IMPACT EVALUATION FOR YEAR 11

Evaluation Report: An exploration of impact in Year 11

This impact evaluation was carried out by NFER and the University of Westminster

Report authors

NFER: Palak Roy, Andrew Smith, Pippa Lord, Roland Marden and Ben Styles

University of Westminster: Veruska Oppedisano, Min Zhang and Richard Dorsett

October 2022

The overarching evaluation was commissioned by the Education Endowment Foundation (EEF) and undertaken by a consortium led by NFER

Public

Contents

About the evaluators

This impact evaluation of the first year of the Academic Mentoring (AM) programme was undertaken by the National Foundation for Educational Research (NFER) and the University of Westminster (UoW).

The NFER is the leading independent provider of education research, and holds the status of Independent Research Organisation (IRO) from UK Research and Innovation (UKRI). Our unique position and approach delivers evidence-based insights designed to enable education policy makers and practitioners to take action to improve outcomes for children and young people. Our key topic areas are: accountability, assessment, classroom practice, education to employment, social mobility, school funding, school workforce and systems and structures. As a not-for-profit organisation, we re-invest any surplus funds into self-funded research and development to further contribute to the science and knowledge of education research **www.nfer.ac.uk**@TheNFER.

The UoW is a diverse international education institution situated in the heart of London. The university champions sustainability, social responsibility, and inclusivity through its work and activities. The evaluators are affiliated to the Centre for Employment Research (CER) at the UoW, which focuses on three broad fields of research: skills, labour markets and programme evaluation; employment relations and employee voice; and equality and diversity.

Evaluation contact details:
Pippa Lord, Research Director
National Foundation for Educational Research, The Mere, Upton Park, Slough, Berks SL1 2DQ
p.lord@nfer.ac.uk

Acknowledgements

We are grateful to Rachel Hayes who led the impact evaluation of the AM programme in her previous role at NFER. We would like to thank the following members of NFER's operations team who collated the data received from Teach First for this impact evaluation: Kathryn Hurd, Jishi Jose, Guido Miani, Shazia Ishaq and Kinnery Koria. We would like to thank Greta Morando (formerly at University of Westminster) who carried out the statistical analysis in the early stages of the project. Thanks are also due to Emma Hawkins and Vrinder Atwal for their efficiencies in coordinating meetings and minutes throughout the project.

We would also like to thank the team responsible for the National Pupil Database and the team at the Office for National Statistics for their assistance with our applications to access datasets via the Secure Research Service.

We are grateful to Teach First and the Department for Education for their advice over the course of the evaluation and sharing their insights from the process evaluation.

We are grateful to members of the Evaluation Advisory Group: Sandra McNally (LSE, University of Surrey), Steve Higgins (Durham University), Bob Slavin (Johns Hopkins University), Heather Rolfe (Former Head of Research at Demos, ex-NIESR), Fatima Husain (SQW), and John Jerrim (UCL IoE).

Finally, we would like to thank the team at the EEF for their support throughout the impact evaluation, with particular thanks to Arnaud Vaganay, Katharina Keck, Amy Clark and Emily Yeomans.

Notice

About year 1 of the National Tutoring Programme (NTP) Academic Mentoring (AM) programme

The National Tutoring Programme (NTP) Academic Mentoring (AM) programme was designed to help disadvantaged pupils catch up on missed learning by providing trained academic mentors to deliver 1:1 and small group tutoring in disadvantaged schools. Schools could decide for themselves, which pupils would receive the support from academic mentors. However, the programme encouraged participating schools to select pupils to receive mentoring from disadvantaged households or those whose education had been disproportionately impacted by Covid-19.

There was also a second strand to the first year of the NTP—Tuition Partners—which offered tutoring support for pupils as a response to the Covid-19 pandemic. The Tuition Partners programme was overseen by the Education Endowment Foundation (EEF) and tuition was provided by a range of approved tutoring organisations.

This evaluation report focuses specifically on year 1 of the AM programme, part of the NTP. The AM programme (2020/21) was managed by Teach First, who oversaw the provision of tutors (known as 'academic mentors') to schools, including recruiting, training, and placing them in schools.

About this study

The EEF commissioned an independent impact evaluation of year 1 of AM and this was carried out by the National Foundation for Educational Research (NFER) and the University of Westminster (UoW). The evaluation aimed to quantify the impact of the first year of the AM programme on pupil attainment/learning outcomes, and how this varied by different models of mentoring, pupil, and school characteristics.

A separate process evaluation of AM (2020/21) was conducted and published in-house by Teach First (Teach First, 2021). Their evaluation used a mixed-methods approach and aimed to investigate the recruitment and training of academic mentors, the deployment of academic mentors in school, and the perceived outcomes of the AM programme for pupils and schools.

About this report

This report covers the findings from the analysis of the impact of the first year of the AM programme (2020/21) for Year 11. The outcome data used in this report are the Teacher Assessed Grades (TAGs) awarded to Year 11 pupils in the summer of 2021. As TAGs have not been used in other studies by the EEF to measure impact, we ran a number of checks on the TAGs prior to the main impact analysis to inform our approach to the analysis and interpretation of the results. The checks are outlined briefly in the report and presented in detail in the Appendix. The report outlines the impact of AM on learning outcomes for Year 11 pupils, through a number of estimators of impact, in both English and maths, and explores how this impact varies by school and pupil characteristics, and by different models of mentoring (e.g., face-to-face vs. online, different mentor: pupil ratios). Due to the use of TAGs as an outcome measure, the Year 11 analyses presented in this report should be considered as exploratory.

How to cite this volume: Roy, P., Oppedisano, V., Smith, A., Zhang, M., Lord, P., Marden R., Styles, B. and Dorsett, R. (2022). *Evaluation of Year 1 of the Academic Mentoring Programme: Impact Evaluation for Year 11. Evaluation Report: An exploration of impact in Year 11.* NFER and University of Westminster. London: Education Endowment Foundation.

Other evaluation reports on the first year of the NTP

A suite of evaluation reports is also available about year 1 of the Tuition Partners programme (2020/21), commissioned by the EEF and conducted by an independent evaluation consortium led by NFER with Kantar Public and the UoW. The suite includes four volumes on: an **implementation and process evaluation** (IPE) of Tuition Partners (2020/21); a **primary school sample impact evaluation** of Tuition Partners (2020/21); a **Year 11 impact evaluation** of Tuition Partners (2020/21); and an **overarching synthesis**, which draws together the IPE and impact evaluation findings from the three evaluation strands of the Tuition Partners (2020/21).

Figure 1 (on page 10) illustrates the NTP programme as part of the government's catch-up strategy, its pillars (in 2020/21), and the components of the evaluation in 2020/21.

Executive summary

The programme

The National Tutoring Programme (NTP) Academic Mentoring (AM) programme (2020/21) was designed to help disadvantaged pupils 'catch up' on missed learning by providing trained academic mentors to deliver one to one and small group tutoring in schools. This evaluation covers year 1 of the AM programme as delivered by Teach First from November 2020 to July 2021 (delivery was in three waves starting 26th October 2020, 15th January 2021 and 22nd February 2021). AM was one arm of the NTP. The NTP aimed to support teachers and schools in providing a sustained response to the Covid-19 pandemic and to provide a longer -term contribution to closing the attainment gap between disadvantaged pupils and their peers. The NTP was part of a wider government response to the pandemic, funded by the Department for Education (DfE) and was originally developed by the Education Endowment Foundation (EEF), Nesta, Impetus, The Sutton Trust, Teach First, and with the support of the KPMG Foundation.

The DfE appointed Teach First to manage the provision of mentors (referred to as 'academic mentors') to schools; recruiting, training and placing them in schools. The mentor worked in the school setting as an employee of the school. It was expected that each academic mentor would work with at least 50 pupils between the date they started in school and the end of the academic year. Mentoring was provided online and/or face-to-face; and was one to one, or in groups of 2-4 pupils; and available in English/literacy, maths, science, humanities, and modern foreign languages. Mentoring was expected to be delivered in schools during normal teaching time, as well as before or after school. In certain circumstances, mentoring could be delivered online with pupil(s) at home.

The AM programme was targeted at state-maintained primary and secondary schools serving disadvantaged populations. 89% of the schools met Teach First's priority criteria, which is based on the proportion of children living in income deprived families (IDACI) and whether the school is in an area of chronic and persistent underperformance (AEA). The remaining 11% of schools had an above average proportion of pupils eligible for Pupil Premium (Teach First, 2021). Participating schools could decide which pupils received support from academic mentors. However, the programme encouraged them to select pupils from disadvantaged households or those whose education had been disproportionately impacted by Covid-19. Pupils in Years 1–11 were eligible (5–16 years old). The programme aimed to reach a minimum of 900 schools and 50,000 children, with 1,000 academic mentors. By the end of February 2021, it had surpassed targets having trained and placed 1,124 academic mentors in 946 schools and delivered mentoring sessions to 103,862 pupils, 49% of whom were identified by mentors as being eligible for Pupil Premium of Free School Meals (FSM), and 23% of whom were identified as having a special educational need or disability.

The AM programme was initiated and delivered at a time of great pressure for schools when the education system had been disrupted by a series of school closures to most pupils and was contending with ongoing widespread pupil and staff absences. Covid-19 related issues disrupted the anticipated operation of academic mentoring during the year. The AM programme involved initial training and ongoing support from Teach First as intended but there was greater variation in schools' deployment of mentors during the latter stages of the Autumn Term 2020/21, and during the January to March 2021 period of school closures to most pupils.

The evaluation

This evaluation report presents the analysis of the impact of the AM programme on maths and English attainment outcomes for Year 11 pupils only—who represent a very small proportion of individuals targeted by the AM programme. Originally, it was planned to evaluate impact across all year groups (Years 1 – 11) at primary and secondary level using schools' standardised assessment data from Renaissance Learning (RL) assessments and, in addition, to evaluate the impact for Year 6 pupils using Key Stage (KS) 2 data. However, these analyses could not go ahead as KS2 assessments were cancelled in summer 2021 (related to the ongoing Covid-19 pandemic) and because the number of schools providing agreement to use their RL data was insufficient to warrant impact analyses. Data was only available for pupils in Year 11. Since GCSEs could not go ahead as planned in 2021, the data was in the form of Teacher Assessed Grades (TAGs), which had not previously been used as an outcome measurement tool. Checks were therefore undertaken to explore if TAGs would be suitable as an outcome measure. The only analysis that could proceed was therefore exploratory.

The evaluation uses a quasi-experimental design (QED), in which a group of secondary schools and Year 11 pupils who did not receive the AM programme were selected for comparison with schools and pupils who received the AM programme. Comparison schools were selected by matching schools that were similar in important, observable regards to the schools that participated in AM. The evaluation included analysis on the availability of AM for pupils who were

eligible for Pupil Premium (a key focus of the overall NTP), and all pupils, as these groups could be identified for both the AM and non-AM schools. In addition, the evaluation aimed to analyse the impact on pupils who received AM by predicting their participation and identifying a comparison group of pupils with similar characteristics. Analysis was based on data about Year 11 pupils' attainment and characteristics from the National Pupil Database (NPD) merged with data provided by Teach First about pupils' participation in AM. In total, 159 AM schools (8,977 Year 11 pupils eligible for Pupil Premium) and an equal number of comparison schools (8,419 Year 11 pupils eligible for Pupil Premium) were included in the final analysis. The evaluation assessed impact in English and maths using Teacher Assessed Grades (TAGs) from 2021. Where appropriate, this impact evaluation refers to important implementation features from the implementation and process evaluation (IPE) conducted by Teach First themselves. However, there is no independent IPE data to draw on in the interpretation of the impact results.

Of the Year 11 pupils selected for Academic Mentoring in this evaluation, 46% of them were eligible for Pupil Premium, however, despite this it is important to note that the number of Year 11 Pupil Premium-eligible pupils selected for AM in AM schools was small as a proportion of all Year 11 Pupil Premium-eligible pupils, and the number of these Year 11 Pupil Premium-eligible pupils receiving AM in maths and/or English (as opposed to other subjects), was smaller still. The same is the case when considering the whole year group of Year 11 pupils – the number receiving AM was small as a proportion of all Year 11 pupils. This means that in the analysis, the number of Year 11 pupils who actually received AM in maths and/or English was heavily 'diluted' by the number of pupils who did not.

The primary impact findings must be therefore treated with a high degree of caution. The analysis was subject to very high dilution; a large proportion of the pupils eligible for Pupil Premium included in the analysis in AM schools were not selected for AM. This was due to limited programme reach and a tendency for teachers to allocate both non-Pupil Premium and Pupil Premium eligible pupils to the programme. This dilution means that, in order to detect an effect, either the effect would need to be very strong amongst the very small proportion of Year 11 pupils eligible for Pupil Premium who were selected for mentoring (and there was no indication that this was the case elsewhere in our analysis), and/or there would need to be strong spillover effects amongst the rest of the Year 11 pupils eligible for Pupil Premium. Although the programme Theory of Change includes such a mechanism, it is unlikely to be relevant at the dilution levels seen.

With such high dilution, it is hard to detect whether AM had an effect on those who received mentoring in the analyses focusing on pupils eligible for Pupil Premium and on all pupils. It is not possible to conclude whether a lack of observed impact is due to the small proportion of disadvantaged pupils who received mentoring, or because AM did not work for those who received it.

An additional challenge was that it was not possible to construct a comparison group of similar Year 11 pupils in non-AM to schools to those who received mentoring in AM schools, based on observable, pupil-level characteristics, and this impact analysis did not go ahead. Schools used information such as classroom assessments to select pupils into the programme that was not observable in the available datasets, suggesting that pupil-level selection was driven by unobserved dimensions.

These constraints, both of very high dilution and not being able to identify a comparison group with similar pupil characteristics, mean that the evaluation is unable to conclude, with any certainty, whether or not AM had an impact on the English or mathematics attainment outcomes of those pupils who received it. The report must be considered in the light of these caveats.

Table 1: Summary of findings

Finding
Initial checks on the data indicated that the TAGs would be suitable as an outcome measure for some exploratory analysis into the impact of AM, in the absence of any other outcome data. However, because the TAGs were a new and unique assessment for which there is no prior data to compare to, the findings reported below should be considered exploratory and should be interpreted with caution.
Year 11 pupils eligible for Pupil Premium in schools that received AM made, on average, similar progress in English compared to Year 11 pupils eligible for Pupil Premium in comparison schools (there was no evidence of an effect). In maths, Year 11 pupils eligible for Pupil Premium in schools that received AM made, on average, slightly more progress (equivalent to 1 months' additional progress) compared to Year 11 pupils eligible for Pupil Premium in comparison schools. However, there is uncertainty around this result; it is also consistent with a null (0 months) effect or an effect of slightly larger than 1 month's additional progress. A particular challenge in interpretation is that, on average, only 13% of Year 11 pupils eligible for Pupil Premium were selected for mentoring by schools, and only 4.2% of Year 11 pupils eligible for Pupil Premium were selected for mentoring in maths and 2.9% in English, meaning that the vast majority of pupils eligible for Pupil Premium included in the analysis did not receive mentoring. Therefore, this estimated impact of AM is severely diluted and it is unlikely any of these differences were due to AM.
When looking at all Year 11 pupils, pupils in schools that received AM made, on average, similar progress in English and maths compared to all Year 11 pupils in comparison schools (there was no evidence of an effect). However, this finding was similarly subject to severe dilution: on average only 10% of Year 11 pupils in the analysed schools were selected for mentoring, with 3.4% in maths and 2.1% in English, and therefore it is hard to detect any effect that may (or may not) have been present.
Within schools that offered AM to Year 11 pupils, there was no association between the number of completed mentoring sessions in maths and Year 11 outcomes in maths, or between the number of completed mentoring sessions in English and Year 11 outcomes in English. These results are associations and not necessarily causal.

EEF security rating

The security of findings is usually described through a padlock classification assigned to the primary outcome of the trial (considering any threats to validity, challenges with design, balance and attrition). Evaluators also conduct exploratory analysis on other outcomes, which are not awarded a padlock rating and are considered of lower security than the primary outcome. At the point of commissioning, the primary outcome for this trial was GCSE outcomes. During the course of delivery it became clear that exams in summer 2021 would not go ahead as planned and were replaced with TAGs. Given uncertainty around the use of TAGs as a research outcome measure, this analysis has been considered exploratory and therefore, not awarded a padlock rating. These results should be treated with greater caution than results that have been assigned a padlock rating.

Additional findings

As the TAGs were a new and unique assessment approach, checks were conducted on the data to investigate whether it would be appropriate to use them as an outcome measure; this was purely to review their suitability for this study and is not a comment or reflection on the TAGs as an assessment mechanism. It does not appear that there were systematic differences in grading between AM and non-AM schools over the exam years analysed, although it is important to note that they are not able to detect with certainty whether there is any systematic bias for the purpose of the evaluation, as teachers may base their assessment on pupils' achievement prior to receiving mentoring and/or they may grade AM pupils differently because they received mentoring. The check on dosage indicated that there may be insufficient sensitivity to detect changes induced by the AM programme in the sample of AM schools. The analysis therefore proceeded on an exploratory basis.

Although there was evidence of a small positive impact for pupils eligible for Pupil Premium in maths (there was no evidence of an impact for pupils eligible for Pupil Premium in English), there was no association between the number of AM sessions in maths or English that pupils received and their attainment in maths or English respectively. Although the lack of association between number of sessions received and outcomes might suggest that AM did not work for those who received it, this could also be explained by a lack of sensitivity in TAGs in AM analysis. It is also worth noting that Year 11 pupils selected for AM received on average 8 sessions in maths and 7 sessions in English (prior to the TAG cut-off date). This is fewer than the programme's suggestion of at least one session a week for a term for each mentoring group.

We recommend that in future years of the AM programme, efforts are made to evaluate the mentoring with a pupil-randomised design, for example by randomising the order in which schools and groups of pupils receive the mentoring, or by varying the number of hours of mentoring to establish the optimum dosage.

Introduction

Background

In March 2020, the government asked all schools in England to close to the majority of pupils in response to the Covid-19 pandemic. Re-opening for some year groups was possible during June and July 2020, but full re-opening was not possible until September 2020. Research highlighted that children were behind in their learning, with attainment gaps and issues relating to access to remote learning provision felt to be more acute in the most deprived schools (EEF, 2020; Cullinane and Montacute, 2020; UCL, 2020; Sharp *et al*., 2020). The government launched a one-off universal **£650 million catch-up premium** for the academic year 2020/21, to support schools to provide catch-up activities to help pupils make up for lost teaching time.

The government also launched a **National Tutoring Programme (NTP)** to provide additional, targeted support for those children who needed the most help (for example, the disadvantaged and vulnerable groups that would have been affected most). The year 1 of the NTP (2020/21) was developed by the Education Endowment Foundation (EEF), Nesta, Impetus, The Sutton Trust, and Teach First, and with the support of the KPMG Foundation. The first year of the NTP was made up of two pillars: the Tuition Partners programme (which provided tutoring support to pupils and was delivered by the EEF), and the Academic Mentoring (AM) programme (in which mentors were placed in schools to work with small groups of pupils; managed by Teach First).[1]

In their review of the evidence on Covid-19 disruptions and the impact on attainment, the EEF highlighted tuition as a route for providing support—in addition to high quality teaching and learning in the classroom. There is a large body of evidence that 1:1 tutoring (EEF, 2021a) and small group tuition (EEF, 2021b) are effective (with average effect sizes of 5 months and 4 months, respectively)—particularly where they are targeted at pupils' specific needs. Meta-analyses show positive impacts of tutoring on learning outcomes to the order of 0.3 standard deviations, and that tutoring can be particularly effective for disadvantaged pupils (Torgerson *et al.*, 2018 and Dietrichson *et al.*, 2017). Given the unprecedented circumstances, researchers also highlighted that 'recovery' or 'catch up' research should take into account context, and in particular 'lockdowns', recovery strategies, and moderating features (such as online access).

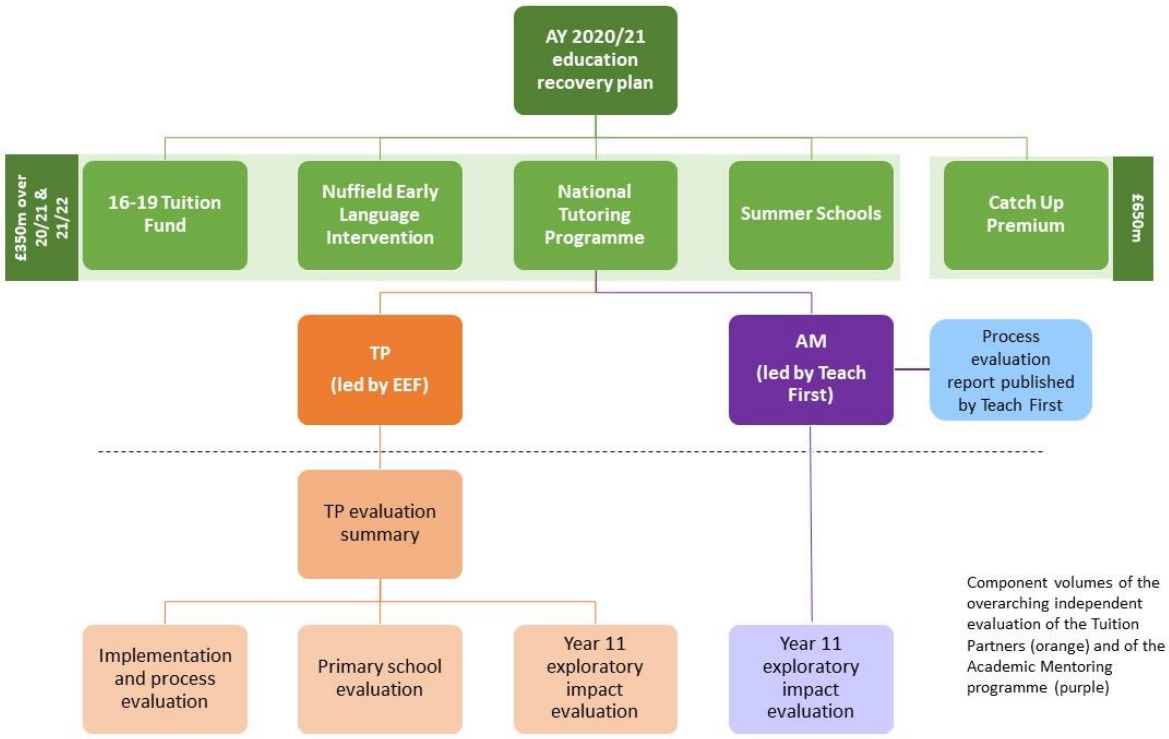
This evaluation report covers the impact of the first year of the AM programme (2020/21) on Year 11 pupils. The first year of the AM programme (2020/21) was managed by Teach First, who oversaw the provision of tutors (known as 'academic mentors') to schools, including recruiting, training, and placing them in schools. The EEF commissioned an independent impact evaluation of year 1 of the AM and this was carried out by the National Foundation for Educational Research (NFER) and the University of Westminster (UoW). A separate process evaluation of year 1 of the AM was conducted and published in-house by Teach First (Teach First, 2021).

The evaluation of year 1 of the Tuition Partners programme (2020/21) is covered in a separate suite of reports, commissioned by the EEF and conducted by an independent evaluation consortium led by NFER with Kantar Public and the UoW.

Figure 1 illustrates the NTP programme as part of the government's catch-up strategy, its pillars (in 2020/21), and the components of the evaluation in 2020/21.

[1] Note, school-led tutoring was not a feature of year 1 of the NTP.

Figure 1: The National Tutoring Programme (NTP) pillars and evaluation (2020/21)



Intervention: about the first year of the AM programme

The first year of the AM programme aimed to help disadvantaged pupils 'catch up' on missed learning caused by school closures in academic years 2019/20 and 2020/21 by providing trained academic mentors to deliver 1:1 and small group tutoring in schools. Schools had to meet certain criteria for disadvantage to be eligible for the programme but they could choose which pupils would receive the support from academic mentors. However, Teach First advised schools to select pupils from either disadvantaged households or those whose education had been disproportionately impacted by Covid-19 disruptions (Teach First, 2021).

Teach First managed the provision of tutors (referred to as 'academic mentors') to schools; recruiting, training, and placing them in schools. Teach First matched academic mentors to schools, considering the individual school's needs. The mentor then worked in the school setting as a full-time employee of the school.

The AM programme was underpinned by a logic model, in which it was expected that pupils receiving mentoring would make accelerated academic progress (compared to other comparable pupils that did not receive mentoring): 'Mentored secondary pupils make 3 months additional progress in a particular subject than comparable pupils that do not receive academic mentoring' (see the logic model in Teach First, 2021, p. 12).

Table 2 outlines the key features of the first year of the AM programme, using similar headings to those in the EEF's TIDieR[2] framework:

[2] TIDieR stands for Template for Intervention Description and Replication.

Table 2: Summary of year 1 of the AM programme (2020/21)

Aspect	Description
Programme	Academic Mentoring (AM)
Why (rationale)	Research shows that pupils' learning has been affected by school closures due to COVID-19 (Cullinane and Montacute, 2020; EEF, 2020). The AM programme is designed to provide additional support to schools to help disadvantaged pupils whose education has been most affected. There is a large body of evidence that tutoring and small group tuition is effective, particularly where it is targeted at pupils' specific needs, and that it can be particularly effective for disadvantaged pupils (Dietrichson *et al.*, 2017; EEF, 2018b, 2018a; Nickow, Oreopoulos and Quan, 2020; Torgerson *et al.*, 2018).
Who (recipients)	State-maintained primary and secondary schools in England. Schools had to meet certain disadvantaged criteria to qualify for the AM programme. • IDACI of 40 or greater (40% of pupils live in the three most deprived deciles). • Or IDACI of 35 to 40 and an Achieving Excellence Areas (AEA)[3] score of 4 to 6. • Or IDACI of 30 to 35 and an AEA score of 5 or 6. • Or IDACI of 25 to 30 and an AEA score of 6. • Schools with an above average (23.5) Pupil Premium rate if there is an available local academic mentor. The AM programme aimed to place 1,000 academic mentors in schools[4]. Schools could receive no more than two mentors. Schools were advised that academic mentors should support pupils from disadvantaged households or those whose education had been disproportionately impacted by COVID-19. The AM programme aimed for academic mentors to reach a total of 50,000 pupils in the 2020/21 academic year.[5]
What (materials)	The AM programme was an in-school tutoring service provided as part of year 1 of the National Tutoring Programme (NTP) (2020/21). The aim of the AM programme was to help disadvantaged pupils catch up on missed learning caused by school closure in academic years 2019/20 and 2020/21.
What (procedures)	Academic mentors provided support tailored to individual school needs with a focus on subject-specific work (1:1 or small groups), revision lessons, and additional support for those not in school (for example those who were clinically vulnerable).
Who (provider)	Academic mentors were required to be either graduates with a degree at 2.2 or above in a subject related to the subject they mentored in, or teachers with Qualified Teacher Status (QTS). Maths and English GCSE at grade C or above was required for all mentors. All academic mentors completed a short initial training programme delivered by Teach First (one week for QTS, 2 weeks for non-QTS) prior to starting tutoring. The training course covered three key areas: essential content (such as safeguarding and well-being); teaching content (such as planning for small group teaching and assessment); and subject-specific pedagogy. In addition to the initial training, the mentors were also required to attend one training per term. They also received ongoing support from the Curriculum and Training Leads (CTLs) who were assigned to the mentors based on their specialist subject and delivered a monthly check-in call as well as further ad hoc support on demand.
How (format)	Academic mentoring was delivered either in person or online. Each school could employ up to two mentors in the following subject areas: **Primary**: • Numeracy • Literacy **Secondary**: • Maths • English • Science • Humanities • Modern foreign languages The large majority of academic mentoring sessions were teaching subject content or study skills related to the subject they were mentoring in with very few that focused on revision, general study skills, or classroom support. Teach First provided guidance to schools to encourage a structured approach to tutoring sessions and close liaison between class teacher and academic mentor. Teach First also provided academic mentors with regular support from Curriculum and Training Leads who advised on best practice.

[3] The AEA analysis seeks to combine indicators, which show current educational performance with indicators, which show capacity to improve to define areas, which are most in need of support. AEA in categories 5 and 6, areas identified as 'cold spots' by the Social Mobility Commission and areas where a large proportion of schools are rated 'Requires Improvement' or 'Inadequate', or where standards are poor.
[4] The 900 schools referred to in the Study Plan was an estimated required number of schools.
[5] Mentors were placed in 946 schools and the mentoring sessions were provided to 103,862 children.

Where (location)	Academic mentoring was expected to be delivered in schools during normal teaching time as well as before and after school. In certain circumstances, mentoring could be delivered online with pupil(s) who were learning from home, for example those who were shielding/medically vulnerable.
When and how much (dosage)	The AM programme intended to place academic mentors in schools at three points, October 2020, January 2021, and February 2021. The aim was to place a total of 1,000 mentors. Each mentor was expected to work with at least 50 pupils between September 2020 and July 2021. Programme guidance recommended that pupils received 1:1/small group tutoring sessions at least once a week and lasting for at least a term.
Tailoring (adaptation)	As a result of the national lockdown in January to March 2021, when schools only remained open to children of key workers and vulnerable pupils, AM sessions were delivered online for pupils that remained at home during that time.

The AM programme: implementation summary

According to Teach First's process evaluation report, 2,355 schools applied to join the AM programme. From these applications, 946 schools were accepted onto the AM programme, representing 12% of total eligible schools in England. 89% of the participating schools met Teach First's priority criteria, which is based on the proportion of children living in income deprived families (IDACI) and whether the school is in an area of chronic and persistent underperformance (AEA). The remaining 11% of schools had an above average proportion of pupils eligible for Pupil Premium which was another eligibility criteria (Teach First, 2021). Academic Mentors were employed directly by schools, who were given the discretion to decide which pupils should access tuition, provided that tuition was delivered to a minimum of 50 pupils across the year. However, they were 'strongly encouraged' by Teach First 'to consider those from disadvantaged households or those whose education has been disproportionately impacted by COVID-19, such as children with Special Educational Needs and Disabilities or children with a social worker' (Teach First School Partnership Agreement).

As per the process evaluation report by Teach First (Teach First, 2021), by February 2021, the AM programme had trained and placed 1,124 academic mentors in 946 schools, exceeding the target of 1,000 placements, and delivered mentoring sessions to 103,862 pupils, exceeding the target of 50,000. The report cites that 49% of these pupils were identified by mentors as being eligible for Pupil Premium and 23% of them were identified as having a special educational need or disability. The report also indicated that these factors were considered alongside others, particularly pupils' missed learning since the start of the pandemic. A range of subjects were taught by academic mentors. In primary schools, two-thirds of the mentoring sessions were in literacy and a third were in numeracy. In secondary schools, where mentoring was conducted in a broader range of core subjects, nearly two-third of sessions were either in English or maths, with a much lower proportion in other subjects.

Schools generally followed the recommended practice of short duration and high frequency, with some variation according to the phase of school and content of sessions. Three-quarters of the groups were between two and seven pupils with a typical group size of between two and four pupils. Session length varied by phase with shorter sessions in primary schools (15–30 minutes) and longer sessions in secondary schools (30–60 minutes). Pupils received sessions regularly in primary schools, with two-thirds meeting three times a week or more. In secondary schools, most groups mirrored class lessons for the subject, with more than two-third of sessions being held once or twice a week.

Covid-related issues disrupted the normal operation of academic mentoring during the year. The AM programme involved initial training and ongoing support from Teach First as intended but there was greater variation in schools' deployment of mentors during the latter stages of the Autumn Term 2020/21, and during the January to March 2021 period of restricted attendance in schools. Teach First's process evaluation found that whilst the vast majority (80%) of mentors were deployed in the way advised by Teach First, a smaller proportion (20%) of academic mentors were used in other ways for a portion of their time in role, for example to provide teaching cover or to assist with teaching key worker and vulnerable children attending school (Teach First, 2021).

About the evaluation and its objectives

This evaluation aimed to quantify the overall impact of the first year of the AM programme (2020/21) on pupil attainment/learning outcomes, and how this varied by pupil and school characteristics, and by model of mentoring.

Originally, we intended to evaluate the impact for both primary and secondary school pupils, using schools' assessment data from Renaissance Learning (RL) assessments for pupils in Years 1–6 (primary school), and in Years 7–10 (secondary schools). However, a change in data sharing arrangements made by RL late in the academic year meant that schools were required to provide further opt-in consent. Despite considerable efforts to re-contact schools to gain

their further consent, the number of schools providing agreement was insufficient to warrant impact analysis on the primary and secondary evaluation samples using RL data. We also originally intended to analyse the impact of AM on Year 6 pupils' attainment by using Key Stage (KS) 2 national assessment data for all primary schools where Year 6s were involved in AM. However, due to the cancellation of the KS2 national assessments in 2021, and as there was no alternative national-level data available, this Year 6 analysis was removed from the evaluation. Hence, this evaluation report focuses on the impact of the AM programme on English and maths for Year 11 pupils only—who represent a very small proportion of individuals targeted by the AM programme.

As GCSE exams were cancelled in 2021 due to ongoing Covid-19 disruptions, we amended the outcome measure for the Year 11 analysis to Teacher Assessed Grades (TAGs) that were implemented for 2021. As this was a new and unique assessment practice for 2021, we ran some initial checks on the data, prior to running the Year 11 impact analysis, to investigate whether it would be appropriate to use TAGs as an outcome measure (e.g., in terms of sensitivity and reliability). This was purely to review their suitability as an impact-measurement tool for this study. The Year 11 analysis conducted in this study is exploratory in nature because TAGs have not been used as a research outcome measure before.

The evaluation aimed to determine what difference, if any, AM made to attainment outcomes, in both English and maths, for Year 11 pupils. This was investigated through a number of estimators of impact, using a quasi-experimental design (QED). Given the urgency of addressing missed learning in schools and supporting academic catch up, it was not considered ethical to randomise schools to receive, or not receive, AM. Moreover, given the timing of the evaluation commission, the QED approach was appropriate as delivery was already underway. When designing this impact evaluation, a number of issues were considered including: defining the research questions (RQs) that could be answered; the appropriateness of a QED and how best to operationalise a comparison group design; scale and scope (note, the evaluation involves data provided by academic mentors about all participants in Year 11 for analysing and reporting on impact); burden on schools; use of the National Pupil Database (NPD) data and other assessment data. These issues are discussed further in the 'Methods' section.

The primary research question (RQ1) focuses on all PP-eligible pupils as a way of identifying would-be participants and avoiding selection bias. Any effect of mentoring would be 'diluted' among all the PP-eligible pupils (as not all would take part in AM), but this was outweighed by being able to identify a majority-type of potential participants in both intervention and comparison groups. As not all of the PP-eligible pupils (nor indeed all of the pupils in Year 11) would be selected for AM, this RQ therefore considered the impact of the availability of AM and not the impact of actual participation. Similarly, RQ2 focuses on all Year 11 pupils, hence it considers the impact of the availability of AM and not actual participation. These issues are discussed in further detail in the 'Pupil-level selection' section.

The RQs evaluated in this report are for Year 11 pupils and are as follows:

Outcome analysis:

RQ1: What is the impact of AM availability on Year 11 PP-eligible pupils' attainment in secondary schools where AM was available to Year 11 pupils?

Further analysis:

RQ2: What is the impact of AM availability on predicted participants' attainment in secondary schools where AM was available to Year 11 pupils?
RQ3: What is the impact of AM availability on all pupils' attainment in secondary schools where AM was available to Year 11 pupils?

Moderator analysis:

RQ4: How does the impact of AM availability vary among Year 11 PP-eligible pupils, by school and pupil characteristics?
RQ5: How do outcomes vary among Year 11 AM pupils, by model of mentoring?

The study plan (versions 1 and 2) can be accessed on the **EEF website**.

The study plan contains the details of the changes made to the evaluation sample as outlined above.

Ethics

The study adhered to NFER's Code of Practice, and was approved by NFER's Code of Practice group at project set up in September 2020. The proposal was approved by the Westminster Business School Ethics Committee in June 2021.

Data protection

The Department for Education (DfE), the EEF, the evaluator (the collaboration between NFER and the UoW) were joint data controllers for the evaluation. Teach First was the data controller for AM delivery.
All work conducted for the evaluation was compliant with the Data Protection Act (DPA) 2018 and General Data Protection Regulation (GDPR). NFER has ISO27001 and Cyber Essentials Plus certifications and registration with the Information Commissioner's Office.

To carry out the evaluation, it was necessary to use and share personal data about pupils (both those who take up the offer and those who do not), as well as key staff members at participating schools and AM staff delivering the mentoring, so that they can be asked about delivery. The evaluator put in place appropriate measures to prevent pupils' personal information from being accidentally lost, used or accessed in an unauthorised way, altered, or disclosed. In addition, each organisation involved would limit access to pupils' personal information to their staff members who have a business need to see it. Any data shared between the school, Teach First, the EEF, the evaluator, and the DfE was via a secure portal.

The EEF, Teach First, and NFER identified the following legal basis for processing personal data:

GDPR Article 6 (1) (f), which states:

> *Legitimate interests: the processing is necessary for your (or a third party's) legitimate interests unless there is a good reason to protect the individual's personal data which overrides those legitimate interests.*

We carried out a legitimate interest assessment, which demonstrated that the evaluation fulfilled the Evaluator's core business purposes (undertaking research, evaluation, and information activities). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence for about the most effective ways of providing catch-up tuition. The evaluation cannot be done without processing personal data but processing does not override the data subject's interests.

The UoW identified the following legal basis:

GDPR Article 6 (1) (e), which states:

> *Public task: the processing is necessary for you to perform a task in the public interest or for your official functions, and the task or function has a clear basis in law.*

A separate legal basis is identified for processing special data. The legal basis for processing special data for the evaluation of the AM programme was:

GDPR Article 9 (2) (j), which states:

> *Archiving, research and statistics (with a basis in law): processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.*

Rights and retention periods

Parents (and KS4 pupils) could withdraw their child from the programme and/or from their data being processed, until it was added to the EEF archive. If they withdrew from the programme or evaluation (i.e., decided not to engage with the AM programme or the evaluation), the Evaluator would still use the evaluation data that the school provided up to that point and link it to the NPD unless the parent/KS4 pupil indicated otherwise.

Three months after the publication of this evaluation report, all of the pseudonymised matched data (pupil data only) will be added to the EEF archive, which is managed by Fischer Family Trust (FFT) on behalf of the EEF and hosted by the Office for National Statistics (ONS). This will enable the EEF and other research teams to use the pseudonymised data as part of subsequent research through the ONS Approved Researcher Scheme, including analysing long-term outcomes through the NPD. This data may also be linked to other research datasets for the purpose of Covid-related educational research.

We will securely delete any personal data relating to the evaluation 1 year after the publication of the final report. Teach First will securely delete any personal data collected for the evaluation alone at the end of the AM programme, when final grants have been paid (expected to be August 2021). Teach First may keep personal data collected as part of the delivery of their mentoring services for longer—this is covered in the privacy notice they provide. Once data has been archived, it is held in the EEF archive until it is no longer needed for research purposes.

Linking to NPD and use of Secure Research Service (SRS)

NFER securely submitted the pupil data to the NPD team to be matched to the pupil data held on the NPD. The UoW and NFER accessed the matched NPD data for analysis through the SRS secure online system. The SRS system does not allow users to remove or copy data from its servers. In this way, the team from the UoW did not have access to any identifiable data.

The project met the ONS 'five safes' in the following ways:

- Safe people: all researchers accessing the project's data via the SRS are Accredited Researchers and hold a 'basic disclosure' certificate that is no more than 2 years old.

- Safe projects: the project meets the conditions for accessing personal-level data. A full request to the NPD team was submitted, outlining the appropriate and ethical use of the data, and the public benefit of the research (to contribute to the evidence base on tutoring, and inform future tutoring programmes). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence about the most effective ways of providing catch-up tuition. The evaluation could be done without processing personal data but processing does not override the data subject's interests.

- The research team and the EEF were committed to publishing the results of the study.

- Safe settings: all researchers working on the NPD data only accessed the data via the SRS secure online system. Our organisations obtained safe room connectivity/homeworking agreements to have SRS remote connectivity access.

- Safe outputs: All outputs were checked by the ONS team to ensure that the outputs do not allow identification of individuals. Outputs were checked against the Intended Permitted Outputs and be subject to standard ONS disclosure rules.

- Safe data: the data request includes data variables of identifiability risk level 3 (pupil matching reference, PMR), as the DfE will match the data we collect with the NPD data. The PMR (meaningless identifier) replaces the unique pupil number (UPN) when the data are matched and then archived to minimise the risks of identification. Our researchers will only analyse de-identified data in the SRS.

Project team

The impact evaluation was delivered by the following team from NFER and the UoW:

- Pippa Lord, Trials Director and Consortium Lead (NFER)
- Palak Roy, Senior Research Manager (NFER)
- Roland Marden, Senior Research Manager (NFER)
- Kinnery Koria, Project Manager (NFER)
- Andrew Smith, Senior Evaluation Analyst (NFER)
- Veruska Oppedisano, Senior Lecturer (UoW)

- Min Zhang, Research Fellow (UoW)
- Richard Dorsett, Professor of Economic Evaluation (UoW)
- Ben Styles, Head of Classroom Practice and Workforce (NFER)
- Rachel Hayes, Senior Research Manager (in her previous role at NFER)
- Kathryn Hurd, Head of Survey Operations, Research, and Product Operations (NFER)
- Greta Morando, Research Fellow (UoW)

Methods

Evaluation design

Table 3: Evaluation design

Design		Matching/Instrumental Variables
Unit of analysis		Pupils eligible for Pupil Premium (PP; identified as pupils eligible for Free School Meals [FSM] in the previous 6 years, NPD var: FSM ever in 6 years) in Year 11
Stratification variable (s) (if applicable)		n/a
Number of units to be included in analysis (intervention, comparison)		318 secondary schools (159 intervention, 159 comparison schools). 17,396 PP-eligible pupils[6] (8,977 from intervention schools and 8,419 from comparison schools)
Primary outcome	Variable	Attainment in English, attainment in maths
	Measure (instrument, scale, source)	TAGs replacing GCSEs in 2021, NPD
Secondary outcome(s)	Variable(s)	n/a
	Measure(s) (instrument, scale, source)	n/a
Baseline for primary outcome	Variable	Attainment in English, attainment in maths
	Measure (instrument, scale, source)	KS2 SATs in 2016 from the NPD
Baseline for secondary outcome(s)	Variable	n/a
	Measure (instrument, scale, source)	n/a

KS, Key Stage; n/a, not applicable; NPD, National Pupil Database; SATs, Standard Assessment Tests; TAGs, Teacher Assessed Grades.

The analysis presented in this report is focused only on Year 11 pupils using as outcomes the TAGs awarded in 2021.

The challenge, as with any quasi-experimental impact evaluation, is that the selection of schools and pupils into the AM programme is unlikely to be random. We used propensity score matching to control for school selection into AM by constructing a matched comparison group of non-AM schools that was similar in important ways to the AM schools (details below in 'Propensity score matching' in 'Statistical analysis' section). This assumed that sufficient school characteristics could be observed to control for selection (the 'selection on observables' or 'conditional independence' assumption). It is this type of selection that Weidmann and Miratrix (2020) consider, providing evidence that simple matching approaches may work well for this purpose. The counterfactual is assumed to be a 'business as usual'. However, in the prevailing climate, it was likely that pupils who did not receive AM were provided with other forms of support by schools, and these may have involved 1:1 or small group support. In the sample of AM schools, we controlled

[6] Note, the difference between pupils selected for AM and pupils analysed. Table 2 contains information about the PP-eligible pupils included in the analysis. As noted elsewhere, this is not the same as the pupils that participated in AM.

for schools providing tutoring through the Tuition Partners pillar, but not for other forms of tutoring. We excluded the secondary schools that delivered Tuition Partners from the sample of eligible comparison schools.

We originally intended to perform an instrumental variable (IV) analysis (see the 'Statistical analysis' section below for a full account). However, this approach was unable to address the pupil-level selection bias (it could only address the school-level selection issue already addressed by the matching strategy). Evidence from Weidmann and Miratrix (2020) suggests that controlling for observable characteristics addresses the problem of school-level selection and, since the IV strategy would not have addressed the very high dilution in the primary analysis (see the section on 'Pupil-level selection'), we did not continue with this approach. Instead, we directed our analysis resources to solving the dilution problem rather than pursuing the IV analysis.

Mentoring was targeted at pupils in disadvantaged schools. Since the group of targeted pupils could not be identified within non-AM schools (i.e., comparison schools), we attempted to avoid pupil-level selection bias by focusing on pupils eligible for Pupil Premium since these were expected to represent a key group of the mentored pupils and are identifiable in both AM and non-AM schools.

Participant selection

How schools join the AM programme: AM schools

Schools were eligible to apply for an academic mentor if they have either:

- an Income Deprivation Affecting Children Index (IDACI) of 40 or greater (40% of pupils live in the three most deprived deciles);
- an IDACI of 35 to 40 and an Achieving Excellence Areas (AEA) score of 4 to 6;
- an IDACI of 30 to 35 and an AEA score of 5 or 6;
- an IDACI of 25 to 30 and an AEA score of 6; or
- if a school does not meet the criteria above, but has an above average Pupil Premium, set at 23.5%, if there is an available local academic mentor.

Participating schools were able to identify, which of their pupils they felt would most benefit from academic mentoring. Pupils could be in Year 1 right through to Year 11. AM did not have prescribed conditions on the characteristics of pupils who received the intervention, although the AM programme encouraged schools participating in the programme to select pupils to receive mentoring who came from disadvantaged households or whose education had been disproportionately impacted by Covid-19. This report focuses only on AM for Year 11s.

The Year 11 analysis focuses on secondary schools where at least one pupil in Year 11 received AM. In order to be an 'AM' (intervention) school in the analysis, the mentoring needed to have started before the assessment date. For the TAGs this was slightly complicated because the TAGs were not tests that were completed on a particular date. Therefore, we set a cut-off date of 11 June 2022, which was one week before the deadline for schools to submit their TAGs to the examination boards (assuming some time for schools to complete internal moderation and/or quality assurance processes). .

In the study plan, we intended to use the sample of 367 secondary schools doing AM for the main analysis.[7] However, as not all schools had pupil-level information on whether mentoring was delivered to Year 11 pupils, we decided to focus instead on the sample of schools for which pupil attendance data, including whether mentoring was delivered to Year 11 pupils, was available.

A total of 45% of 367 of secondary schools had usable pupil-level data, which left us with 164 schools, as described in the 'Participant flow including losses and exclusions' section. We selected a similar number of comparison non-AM schools using propensity score matching (see 'Propensity score matching' in 'Statistical analysis' section).

[7] The number of secondary schools that delivered AM was 367 by the end of the academic year 2020/21. However, at the time of calculating the minimum detectable effect size (MDES), in the spring 2022, the data available indicated that the number of secondary schools delivering AM was 366.

Outcome measures

Baseline measures

We used maths and reading KS2 data from the NPD as the baseline for Year 11.

Primary outcome

The analysis used TAGs for 2021 that were made available in the NPD. The evaluation assessed the impact of mentoring on pupils' attainment, in particular PP-eligible pupils' attainment, as the purpose of the AM programme (2020/21) was to help disadvantaged pupils to 'catch up' on missed learning due to the Covid-19 pandemic and the restrictions on schools in 2020 and 2021.

Pupil-level AM data

Teach First provided NFER with a list of 367 schools with school-level information. Of these, 164 schools provided also pupil-level information on AM delivery. Note, the evaluators did not provide quality assurance of this data. After collating and cleaning the data, NFER shared these lists with the ONS. The data included school identifiers to allow the NPD team to match them to the NPD data. Once matched, the NPD removed the school identifiers and retained a meaningless identifier.

Sample size

We used cluster randomised trial power calculations to provide an indication of the MDES for RQ1, although the intervention relies on a quasi-experimental approach. We allowed for clustering of pupils within schools.

During the design phase, the assumed sample was based on a level of take-up of 10 PP-eligible pupils per school[8]. This reflected our intention to focus primarily on PP-eligible pupils. We assumed an intra-cluster correlation (ICC) of 0.15 and pre-post correlations of approximately 0.5. Note, since our primary analysis focuses on disadvantaged pupils, we do not produce separate estimates for all pupils.

Our sample size calculations refer to the population of 366 unique secondary schools doing AM by spring 2022. Power calculations suggested that with 366 AM secondary schools and a minimum of 366 non-AM schools the MDES for Year 11 pupils would be 0.07. By the end of the academic year, 367 secondary schools had delivered AM to Year 11 pupils.

After cleaning the data and performing the matching, the final sample was composed of 159 AM schools with pupil-level information that had at least one Year 11 pupil who was selected for AM and had a mentoring session before 11 June 2021.

Statistical analysis

Propensity score matching

We used matching to control for school selection into the AM programme by constructing a matched comparison group of non-AM schools that was similar in important, observable regards to the AM schools in the Year 11 sample.

To create a sample of comparison schools, we followed the procedure outlined in the study plan and used a slightly updated list of variables listed in the study plan below for reference and in Appendix B:

- IDACI;
- AEA score;
- interaction between IDACI and AEA scores;
- KS1 to KS2 value added attainment, at district level;
- management/school type secondary—Community, Academy, Foundation, and Free;
- schools, Sponsored Academies, Voluntary school, Studio school, and University Technical College;
- region (London, Government Office Region, and regional dummies);

[8] The number is based on the data on pupils receiving AM provided by Teach First. On average there were 20 pupils per secondary school doing AM. About 50% of them are PP-eligible pupils, which left us with 10 pupils. The assumption was based on all PP-eligible pupils in the secondary schools to be targeted, as the initial analysis was planned on pupils in all year groups in secondary schools. We did not update the take up number after the cancellation of the secondary school-level analysis.

- Free School Meals (FSM)—percentage eligible in previous year; and
- KS4 maths and English in 2019/20.[9]

Matching method

Matching was used only for a subsample of the participating schools. Comparison schools were selected from the sample of Year 11 schools, excluding all 846 secondary schools in the AM school-level population file and 367 secondary schools that delivered academic mentoring without pupil-level information on the year groups that received mentoring (see 'Participant flow including losses and exclusions' section).

The propensity score matching method matched treated units to comparison units using propensity scores. A unit's propensity score is its probability of being in the treated group given its values for the matching variables. This was estimated by fitting a probit regression model to a dataset that included all treated units and all potential comparison units, where school treatment status is the dependent variable and school characteristics in the dataset are the predictors. Results from the probit model are reported in Table A1 in Appendix A. The balance table comparing observable characteristics between AM intervention schools and all eligible comparison schools is reported in Table A2 in Appendix A. The table shows that most of the observable characteristics are significantly different between AM intervention schools and eligible comparison schools before matching.

Each treated unit was then matched without replacement with a comparison unit with the closest possible propensity score (1:1 matching). Common support is shown in Figure A1 in Appendix A: five AM schools have very high propensity scores (i.e., extremely likely to be chosen to be AM schools). Propensity scores for AM and non-AM schools between 0.12 and 0.48 were kept, which consisted in dropping the bottom 70% of the distribution (non-AM schools, potential controls) and 0.04% from the top of the distribution for the analysis. As no non-AM schools scored higher than 0.48, these five schools are dropped due to lack of common support.

Matching was undertaken without replacement using the single nearest neighbour. Table A3a, A3b, and A3c in Appendix A report different specifications using different calipers. Matching quality is measured by differences in attainment between AM and matched non-AM schools before the intervention for matched and unmatched samples. The specification without calipers (Table A3a) indicates that the significant differences in KS2 and GCSEs between AM and non-AM schools before matching disappear after matching. Restricting the analysis with calipers (0.05 in Table A3b and 0.01 in Table A3c) did not substantially improve the matching quality but lead to an unnecessary loss of schools (four schools when using 0.01). Hence, our preferred specification did not use calipers.

In matching, the assumption of conditional independence requires that we can observe all covariates that jointly determine the selection process and outcomes. If sufficient school characteristics can be observed to control for selection of schools into the AM programme, simple matching approaches may work well for this purpose (Weidmann and Miratrix, 2020). While the conditional independence assumption cannot be tested, we can explore the extent to which matching balances the covariates we do observe between AM schools and comparison schools. This is discussed in the 'Pupil and school characteristics' section.

[9] In the study plan, we list GPS (grammar, punctuation, spelling) average grades in Year 2017. The variable available in our school-level file relates mostly to primary schools. Some secondary schools also have this data, but there is more than 95% missing for secondary schools. We removed this variable from the list and replaced it with maths and English KS4 in the 2 years before the intervention.

Preliminary analysis

To test whether AM schools are similar to the matched non-AM schools, we presented balance tables and conducted placebo tests on previous years' NPD data. If the selection of the control sample, controls adequately for unobserved factors, there should be no significant difference in attainment between AM schools and comparison schools prior to the intervention. The placebo testing is done for each of the preceding years 2016/17, 2017/18 and 2018/19, using results for GCSE. We also present the results using GCSE scores from 2019/20, although they were determined outside of the usual examination process (see sections 'Ex post 2: distribution of teacher assessed GCSE between AM and matched non-AM schools' and the 'Placebo tests').

We assessed the performance of the match by:

1. (a) Comparing observed characteristics of AM schools and their matched comparators before and after matching (Table A1 in Appendix and Table 7a), and (b) comparing KS2 of AM schools and their matched comparators (Table 7c). This was a school-level analysis, but outcomes are considered for PP-eligible pupils.
2. Comparing GCSE attainment outcomes in AM and matched non-AM schools in pre-AM years as a placebo test,[10] using ordinary least squares (OLS) regression, without weights. As the samples are balanced between AM and non-AM schools, weights are not needed. Results are reported in the 'Placebo tests' section.

We had planned to compare observed characteristics of AM and matched comparator schools using OLS regression, without weights and using entropy balancing and inverse probability weighting using school-level weights, but as the sample was already balanced, we did not conduct this comparison and continued the analysis without weights.

Outcome analysis

Regression (RQ1a): What is the impact of AM availability on PP-eligible pupils' attainment?

To estimate impacts, we regressed the pupil-level outcome on a 0/1 indicator for AM being available at the school level, i.e., if at least one pupil in Year 11 received AM in the sample of AM and matched non-AM schools.

We used KS2 results as the baseline measure of the outcome of interest. All school-level variables listed in the 'Preliminary analysis description' section and in Appendix B were used as controls. Pupil-level controls included background variables (gender, ethnicity, English as an Additional Language [EAL], care status, and special educational needs [SEN]). Pupil-level and school-level controls are listed in Appendix B. We also controlled for schools participating in the Tuition Partners programme. Residuals were clustered at the school level to account for any common school-specific unobservable component. Regression was based on PP-eligible pupils in the AM schools and matched comparator schools. The software used to run the model was Stata 17.

The coefficient on the AM indicator represented the estimated average treatment effect, on an 'intention-to-treat basis', though it should be noted that not all pupils targeted by the AM programme are in the sample of PP-eligible pupils considered here. This is estimated using OLS regressions.

Each estimator has two outcomes: maths; and English. We adjusted for multiple testing using the Romano-Wolf (2005a; 2005b) simulation approach, as implemented by the Stata program rwolf.ado (Clarke, Romano, and Wolf, 2020). Impact estimates are presented with their 95% confidence intervals (CIs) and Romano-Wolf p-values.

Instrumental variable (RQ1b): What is the impact of AM availability on PP-eligible pupils' attainment?

We initially planned to use IV techniques to provide estimates of AM that did not rely on the school-level selection on observables assumption. This approach builds on the AM eligibility criteria and is based on the intuition that, for schools close to the thresholds, eligibility is as good as randomly assigned. This in turn suggests that eligibility can be used as an instrument in an IV analysis of the impact of AM on outcomes. The conditional independence assumption required for matching to identify a treatment effect may not hold. Some necessary control variables, such as a school's propensity and motivation to improve the attainment of more disadvantaged pupils, are unmeasured or unknown. IV methods solve the problem of missing or unknown controls by requiring the conditional independence assumption to hold between the instrument and the outcome.

When we designed the evaluation, we expected that some eligible schools would not participate in AM and some ineligible schools might participate hence we planned to implement our estimator using IV (which we note is also robust

[10] When controlling for IDACI and AEA in the pre-year, we use values of IDACI and AEA in 2020/21 as these are the criteria according to which AM participation is assigned.

to the possibility that some ineligible schools participated). In fact, we used two instruments in our estimator: the combination of the IDACI and AEA thresholds; and the percentage of PP-eligible pupils above the average.

Evidence from an initial analysis on the AM data in March 2021 indicated that the two eligibility criteria were highly correlated with take-up of AM and could therefore serve as effective instruments (as opposed to 'weak' instruments). Estimation was based on both AM and all non-AM schools, and on the subgroup of PP-eligible pupils. Inspite of these advantages, the planned school-level IV was unable to address the pupil-level selection bias. It was only able to address school-level selection on unobservables, which cannot be addressed by propensity score matching. However, evidence from Weidmann and Miratrix (2020) suggests that controlling for observable characteristics addresses the problem of school-level selection.

For the same reason, the IV strategy would not have addressed very high dilution in the primary analysis (see the section on 'Pupil-level selection'). Therefore, we directed our analysis resources to solving this problem rather than pursuing the IV analysis. We considered several options to address the pupil-level selection bias. First, we considered using the time of sign-up as an instrument for AM participation. This would have included dosage (as it exogenously depends on time of enrolling) on the sample of AM schools only and the assumption would have been that time of enrolling is exogenous to performance. However, we knew from ex post 1 check that dosage was not at all correlated with the performance, which meant that even with a strong instrumental variable, the model would not be reliable. Second, we explored the option of restricting the sample to AM schools that targeted a larger proportion of PP-eligible pupils, to address the problem of dilution. However, we did not proceed with any of these analyses as the available data did not have a suitable distribution, and/or the restricted sample was too small, to create groups of schools or pupils with different thresholds for matching. In particular, 94% of AM schools (with pupil-level information) selected, on average, less than 35% of PP-eligible pupils for AM. Only nine schools selected more than 40% of PP-eligible pupils for AM. Having explored several options, we did not proceed with further analysis to address pupil-level selection.

Further analyses

RQ2: What is the impact of AM availability on the attainment of pupils predicted to participate?

Our approach to the primary analysis (RQ1) provides an estimate of the impact on a subgroup of the eligible population, PP-eligible pupils, which did not coincide with the group of children who received AM. RQ2 was therefore designed to involve modelling the probability of pupil participation in AM schools, using various markers of disadvantage recorded in the NPD (socio-economic status measured by eligibility for FSM/Pupil Premium, SEN, interaction with social service, prior attainment, EAL, and ethnicity). To do this, we used pupil-level AM attendance data provided for the analysed sample of 159 secondary schools delivering AM to Year 11. The results were used to predict participation in both AM and non-AM schools and the predictive quality of the model checked against actual participation. The ex post 3 Year 11 check reported in Appendix C provides details of the predictive power of the model. The analysis did not proceed as the predictive power of the model is low.

RQ3: What is the impact of the availability of AM on all pupils' attainment?

As another means of understanding the overall impact of AM, a fourth analysis focuses on the attainment of all pupils (rather than PP-eligible pupils or predicted AM pupils). Similar to RQ1, we regressed the pupil-level outcome on a 0/1 indicator for AM being available at the school level, for pupils at AM schools and matched non-AM schools. We anticipated that these estimates were likely to be smaller than RQ1 and RQ2 estimates, as the AM impact would be more diluted than RQ1 and RQ2 estimates, when considered across the entire year group with at least one pupil in receipt of mentoring. This estimator also captures the effect of spillover (peer) effects. The purpose of this is to capture the overall impact of AM. The regression analysis controls for the same school-level and pupil-level characteristics mentioned in the primary analysis.

Moderator analysis

These analyses explored variation in estimates according to school/pupil characteristics; and different models of mentoring.

RQ4: How does the impact of AM availability vary among PP-eligible pupils, by school and pupil characteristics?

Moderator analysis was conducted through interaction terms on the following categories of:

1. School characteristics: Ofsted (Office for Standards in Education) rating (high vs. low, defined as above/below median); proportion of FSM (high vs. low, defined as above/below median). These variables identified the context where AM was delivered and allowed analysis of whether AM has been more effective in disadvantaged

contexts. With respect to what is listed in the study plan, we also added type of school (academy vs. maintained) and school size (by quartile).

2. Pupil characteristics: prior attainment; SEN versus not; EAL versus not; ethnicity (defined according to the major ethnic groups: Asian; Black; White; Mixed; Any other; and Unknown ethnicity); and gender. These variables identify participants and allow analysis of whether AM has been more effective for children with specific demographics.
3. Other: geography (urban vs. rural; low vs. high IDACI, defined as above/below median): different geographical areas may have different provision of AM in terms of quantity and quality. If schools in more deprived areas have lower attainment outcomes, this may be correlated with the supply of AM in those areas, and the risk that low-quality mentors reach more disadvantaged schools.

Estimates are based on PP-eligible pupils in the analysed sample of AM schools and matched comparison non-AM schools.

RQ5: How do outcomes vary among AM pupils, by model of mentoring?

A descriptive analysis compared outcomes associated with different mentoring models and moderators among AM schools. We did not propose any impact analysis within RQ5 since we cannot observe the counterfactual treatment model among non-AM schools. Instead, this element of the analysis summarised mean attainment among participating pupils in AM schools according to the model of mentoring they experienced. We regressed English and maths attainment on the variables listed below for the sample of AM pupils who received mentoring in English and maths, respectively. In particular, we looked at the following variables at pupil level to assess heterogeneity:

1 The intervention: mode of delivery of completed sessions (online and face-to-face vs. no session in a specific subject) and mentor: pupil ratio (1:1 up to 1:10). We also explored whether the number of completed maths sessions was associated with outcomes in maths, and whether the number of completed English sessions was associated with outcomes in English.
2 Mentors: QTS qualification/not; shared gender of pupil and mentor. In addition to what is listed in the study plan, we also added tutor's graduation year.

Missing data analysis

We did not expect to have missing outcomes since all schools provided TAGs for Year 11 pupils. Where variables used for matching or as controls in regressions were missing, we controlled for them in the analysis by including a dummy for the missing observation. Pupils with missing TAGs and missing baseline assessments were excluded from the analysis. More information about missing data can be found in the results section.

Estimation of effect sizes

Estimates are presented as effect sizes, calculated using the Hedges' g formula (Hedges, 1981). Formally, the effect sizes are calculated as follows:

$$g^* = \frac{\Gamma((n_T + n_C - 2)/2)}{\sqrt{(n_T + n_C - 2)/2} \cdot \Gamma((n_T + n_C - 3)/2)} \cdot \frac{\beta_T}{\sqrt{\frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2}{n_T + n_C - 2}}}$$

Where n_T is the number of treatment group observations, n_c is the number of control group observations, $\Gamma()$ is the gamma function, β_T is the regression coefficient on the dummy variable indicating membership of the treatment group, S_T^2 is the variance of the outcome variable among the treated group and S_C^2 is the variance of the outcome variable among the control group.

Timeline

Table 4: Timeline

Dates	Activity	Staff responsible / leading
October 2020	Project set up, logic model development, materials development, study plan development	Evaluator
October 2020–July 2021	Mentoring period (whole programme)	Academic Mentors
5 January–8 March 2021	National lockdown period—many pupils learning from home, schools only open to children of key workers and vulnerable children	
April 2021	Submit the National Pupil Database (NPD) request	The University of Westminster (UoW)
Mid May 2021 to early July 2021	Teacher Assessed Grades for Year 11	Schools
Summer 2021	Study plan finalisation and publish	Evaluator
June 2021–September 2021	Renaissance Learning (RL) data share amendment and the National Foundation for Educational Research (NFER) and Teach First re-contacting schools to seek their agreement to share RL data	NFER, Teach First, and RL
September 2021	Data cleaning (Management Information/pupil data)	NFER
September 2021	Assessment data from RL sent to Teach First and shared with the Department for Education	Teach First
October 2021–November 2021	Decision to discontinue with Years 1–10 analyses	All
January 2022	Revised study plan (Year 11 only)	Evaluator
December 2021	NPD (unamended) data available and matched into dataset	NPD team/the UoW
May- 2022–June 2022	Impact analysis	The UoW and NFER
July 2022	Write first draft of report	All
July 2022–October 2022	Report revisions and report publication	All

Results

Participant flow including losses and exclusions

Figure 2 provides the details for the flow of participants through the study. AM schools that did not supply pupil-level data were dropped from the sample of eligible AM schools (n=179). All secondary schools that delivered AM to at least one pupil in Year 11 were considered initially. Secondary schools that participated in the Tuition Partners programme (n=846) were dropped from the sample of potential comparison schools. We did not remove from the sample of treatment schools AM schools that also delivered the Tuition Partners programme (30% of schools in the AM sample, see Table 7a and the 'Observable characteristics' subsection). However, we controlled for school-level participation in the Tuition Partners programme in the analyses (see Appendix B).

This left us with 188 AM schools considered for the analysis. 2,230 pupils had a blank pupil identifier and so could not be matched. This removed nine schools entirely from the sample of AM schools. Six AM schools could not be matched on the unique reference number (URN) to the main school-level file and they were removed. Nine schools among those that supplied pupil-level attendance data were not included because they did not meet the inclusion criteria listed as follows: we excluded from the sample pupils who took their first AM session after 11 June 2021 pupils whose first and last date of the AM sessions were missing, and pupils whose first date of the AM session was after the last date of the AM session, removed as errors in data recording was detected. There were 164 schools with at least one Year 11 pupil selected for AM, pupil data supplied and with mentoring sessions before 11 June 2021. This is out of a total of 367 secondary schools in the AM school-level population file.

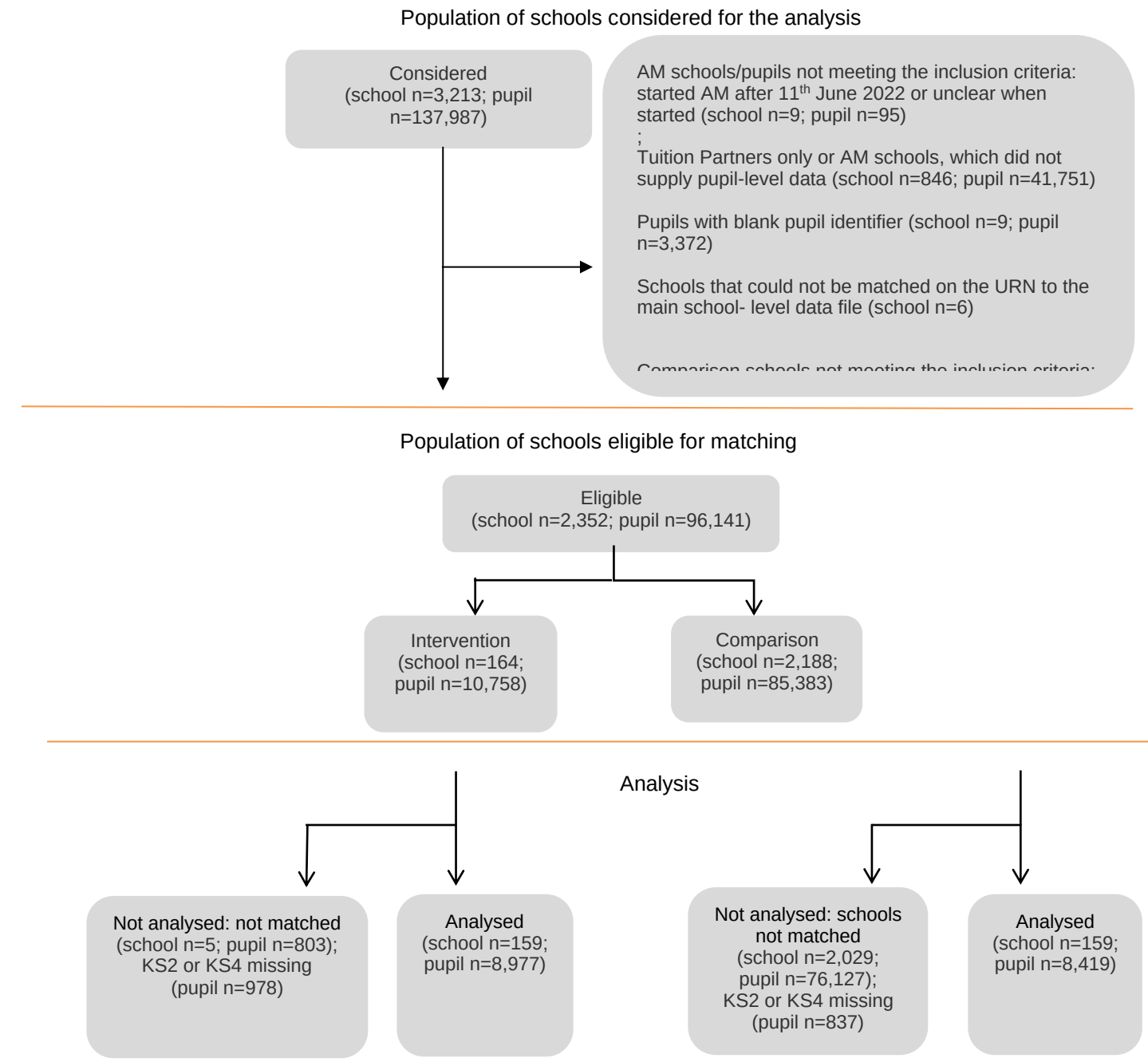
We then lost five schools of the remaining 164 AM schools at the matching stage, to improve the quality of the match with non-AM schools. In terms of the non-AM schools, a similar number of non-AM schools was selected by propensity score matching, implying 2,029 schools were excluded (about 93%).

In the resulting matched sample of schools analysed, about 10% of pupils in AM schools and 9% of pupils in non-AM schools were lost because of missing baseline KS2 results or TAGs in one of the two subjects. In particular, there are 926 pupils with missing baseline test results for KS2 and 83 pupils with missing TAGs in AM schools and 772 pupils with missing KS2 baseline test results, and 88 pupils with missing TAGs in non-AM schools.[11] For KS2 results the missing data occurred years before the intervention, and hence it should not lead to any bias. Pupils with missing primary baseline and/or outcome data were excluded from the analysis.

Our final analysed sample included 159 AM schools and 159 matched non-AM schools. As the main analysis was on the availability of AM in the school, we did not restrict the definition of AM schools in the maths and English samples to account for the subject where AM was actually received. In the AM sample of 159 schools that were included in the analyses, 40 schools (25%) offered AM in maths, 43 schools (26%) offered AM in English, and 77 schools (48%) offered AM in subjects other than English and maths (such as science, modern foreign languages, and others). The analysis focuses on maths and English attainment as outcomes and includes schools where AM was available to Year 11 pupils. In other words, it includes all AM schools irrespective of the subject pupils were mentored in. While this identifies the two subjects where most of the mentoring was delivered, it does not account for the 48% of schools that provided mentoring solely in other subjects.

[11] The numbers do not add up to the total number of pupils with missing KS2 results and/or TAGs because of some overlap, i.e., pupils with both KS2 results and TAGs missing.

Figure 2: Participant flow diagram for AM programme evaluation, Year 11 PP-eligible pupils



AM, Academic Mentoring; KS, Key Stage, PP, Pupil Premium; URN, unique reference number.

Table 5: Minimum detectable effect size at different stages, RQ1, PP-eligible pupils

		Study plan		Analysis	
		Maths	English	Maths	English
MDES		0.07	0.07	0.07	0.08
Pre-test/ post-test correlations	Levels 1 and 2 (pupil and school level)	0.50	0.50	0.49	0.35
Intra-cluster correlations (ICCs)	Level (school)	0.15	0.15	0.08	0.09
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
Average cluster size, Year 11 PP-eligible pupils		10	10	55	55
Number of schools	Intervention	366	366	159	159
	Comparison	366	366	159	159
	Total:	732	732	318	318
Number of pupils	Intervention	3,660	3,660	8,977	8,977
	Comparison	3,660	3,660	8,419	8,419
	Total:	7,320	7,320	17,396	17,396

MDES, minimum detectable effect size; PP, Pupil Premium; RQ, research question.

The MDES for RQ1 is the smallest impact that the analysis can reasonably be expected to be sensitive enough to register. It is measured in units of the standard deviation of the outcome. Following the convention of 80% power and 95% significance, the MDES reported in the study plan is 0.07 for both maths and English on the PP-eligible sample in Year 11. This is set out in Table 5 above, along with the other assumptions used.

As explained earlier, the analysis focused on 164 AM schools for which pupil-level attendance data was available. Five schools were dropped from the sample during the matching stage due to a lack of common support.

The top row of Table 5 presents the MDES for the PP-eligible sample. In addition to reflecting the number of schools and pupils on which impacts are based, the observed ICC and correlation between regressors and TAGs can now be included. The ICC is lower than that assumed at the design stage (0.08 for maths and 0.09 for English compared to 0.15) and the pupil and school-level pre-test/post-test correlation is lower (0.49 for maths and 0.35 for English compared to the assumed pre-post correlation of 0.7),[12] but the cluster size is higher (55 for maths and English compared to 10).[13] Together, these have the effect of not altering the MDES for maths and slightly increasing the MDES to 0.08 for English.

[12] The assumptions were based on GCSEs as there was no prior information on the correlation between observable characteristics and the TAGs.
[13] The assumptions we made in the study plan was based, by mistake, on the AM take up rather than on PP-eligible pupils, Hence the discrepancy in the two numbers.

Attrition

Table 6 reports pupil-level attrition. Attrition is due to sample loss because of missing KS2 results and TAGs (10% of pupils in the AM matched sample and 9% of pupils in the non-AM sample).

Table 6: Pupil-level attrition from the trial (RQ1)

		Intervention	Comparison	Total
Number of pupils	Matched	9,955	9,256	19,211
	Analysed	8,977	8,419	17,396
Pupil attrition (from matching to analysis)	Number	978	837	1,815
	Percentage	9.82	9.04	9.45

Source: Year 11 population data.

RQ, research question.

Pupil and school characteristics

School-level selection

Observable characteristics

Demographic data are presented in Table 7a below, with all figures rounded to three decimal places. Table 7a presents the comparison of observable characteristics between AM schools, matched non-AM schools, and the national averages.

It suggests that the matching produced a sample of matched non-AM schools that are more similar to AM schools than the sample of potential comparison schools.

The descriptives (Table 7a) indicate that most of the observable characteristics were balanced between the AM schools and matched non-AM schools. The significant differences were the following: the percentage of schools with 'Outstanding' Ofsted rating, significantly higher in the AM sample; the percentage of schools in urban areas, 8% higher in the AM sample, and, consequently, the percentage of schools in rural areas, 8% lower in the AM sample; the percentage of SEN students, 1.5% lower in the AM sample and the percentage of pupils with other ethnicity, less than 1% higher in the AM sample.[14] Although not significant, larger schools are also more likely to be AM schools. Measures of deprivation and disadvantage are well balanced between AM and matched non-AM schools.[15] Tuition Partners participation was also statistically different between AM and non-AM schools as, by construction, non-AM schools have been selected from the sample of non-Tuition Partners schools.[16] While we found that some relevant observable characteristics were significantly different between AM and matched non-AM schools, we controlled for all school-level variables listed in Table 7a in all regressions.

Comparison with the national average indicates that the AM and matched non-AM samples are different in terms of observable characteristics from the population of secondary schools. Schools in the AM and non-AM samples feature a slightly lower proportion of schools with 'Outstanding' and 'Missing' Ofsted ratings than the population of schools and

[14] For all the categorical variables we also performed a chi-squared test of the frequency between AM and non-AM schools. In all cases, except for the urban/rural variable, the test does not reject the null hypothesis that the variables are independent.

[15] It seems that AM treatment schools are more likely to be larger, urban schools and with better Ofsted rating. We explored if stratifying by AM eligibility criteria at the matching stage would improve the balance in observable characteristics, but that was not the case.

[16] We tried adding various combinations of additional covariates for the propensity score model, such as the proportion of SEN, EAL, the pupil count, rural, and Ofsted rating. While these resulted in a very small increase in the R-squared when estimating the propensity scores, there was no significant advantage in terms of matching quality. Hence, we kept the set of variables listed in the study plan and in Appendix B.

a slightly higher proportion of schools with 'Inadequate' Ofsted ratings than the population of schools. School types (e.g., Academies, maintained) are differently distributed between the AM and non-AM samples and the population of schools. The AM and non-AM samples show a higher percentage of Academic-sponsor led schools and a lower percentage of Free-mainstream schools than the population of schools.

There are slightly more schools in urban areas in the AM and non-AM samples than in the population of schools (90% in non-AM and 98% in AM schools vs. 83% in the population of schools), but zero missing observations in the urban/rural category for the AM and non-AM samples with respect to the population of schools. AM and non-AM samples feature a slightly different geographical representation, with fewer schools from the South East and the South West and more schools from London and the North East compared to the national sample. In terms of measures of disadvantage, the AM and non-AM schools have a higher percentage of FSM than the national sample (0.40% in the AM and non-AM samples vs. 0.28% in the national sample) and a higher percentage of EAL pupils (20% in non-AM and 22% in AM schools vs. 15% in the population of schools). The average IDACI score is higher in the AM and non-AM sample than in the population of schools, suggesting more of the AM and non-AM samples are more concentrated in more deprived areas. In terms of ethnic background, AM and non-AM schools have a higher percentage of Asian students than the population of schools (21% in non-AM and 23% in AM schools vs. 18% in the population of schools) and of Black students (16% in non-AM and 23% in AM schools vs. 18% in the population of schools).

Overall, it appears that the AM and matched non-AM schools represent a more disadvantaged group of schools than the population of schools, which is consistent with the target of the AM programme. AM and matched non-AM schools share similarity among almost all observable characteristics, making us confident to consider school-level selection to be accounted for by the matching procedure.

As the sample of 159 AM schools with pupil-level information is a subsample of the sample of secondary schools that delivered AM (n=367), we compared observable characteristics between the AM sample with pupil-level information and hence were selected for the analysis, and the population of AM secondary schools. Results, presented in Table 7b, indicated that there were no significant differences between the two samples, suggesting that the sample used in the analysis was representative of the population of AM secondary schools.

The evidence from the preliminary analysis indicates (see Table 7a and the 'Preliminary analysis' subsection) that the quality of the match is sufficiently good.

Table 7a: Baseline characteristics of Year 11 AM schools, matched comparison schools, and national proportions

Variable	Means: National data	Means:Matched comparison	SD: Matched comparison	Means: AM schools	SD: AM schools	Difference AM and matched non-AM	Std Difference AM and matched non-AM
School-level PP KS2 Maths scores 2020/21	100.544	100.614	(1.973)	100.679	(1.811)	0.064	0.023
School-level PP KS2 Read scores 2020/21	101.196	99.439	(1.931)	99.413	(1.924)	-0.026	-0.008
School-level PP KS4 English 2020/21	4.256	4.180	(0.589)	4.260	(0.651)	0.080	0.100
School-level PP KS4 Maths 2020/21	4.531	3.930	(0.657)	4.011	(0.629)	0.081	0.094
School-level ALL KS4 English 2019/20	5.079	4.638	(0.499)	4.693	(0.535)	0.056	0.080
School-level ALL KS4 English 2018/19	4.687	4.187	(0.532)	4.270	(0.546)	0.083	0.115
School-level ALL KS4 English 2017/18	4.686	4.252	(0.582)	4.289	(0.562)	0.038	0.053
School-level ALL KS4 Maths 2019/20	4.954	4.491	(0.550)	4.557	(0.566)	0.066	0.082
School-level ALL KS4 Maths 2018/19	4.636	4.099	(0.578)	4.189	(0.599)	0.090	0.108
School-level ALL KS4 Maths 2017/18	4.632	4.161	(0.604)	4.186	(0.668)	0.025	0.030
Total pupil counts	925.059	891.346	(318.488)	933.604	(328.301)	42.258	0.136
Pupils-to-teacher ratio 2018	16.240	15.754	(2.934)	15.677	(2.660)	-0.077	-0.033
Ofsted 2018: Outstanding	0.207	0.069	(0.255)	0.126	(0.333)	0.057*	0.133
Ofsted 2018: Good	0.525	0.560	(0.498)	0.553	(0.499)	-0.006	-0.013
Ofsted 2018: Inadequate	0.047	0.075	(0.265)	0.082	(0.275)	0.006	0.033
Ofsted 2018: Requires improvement	0.149	0.245	(0.432)	0.214	(0.411)	-0.031	-0.092
Ofsted 2018: Missing	0.072	0.050	(0.219)	0.025	(0.157)	-0.025	-0.105
School type: Academy-sponsor led	0.233	0.396	(0.491)	0.428	(0.496)	0.031	0.078
School type: Community school	0.099	0.113	(0.318)	0.101	(0.302)	-0.013	-0.041
School type: Voluntary aided/controlled school	0.076	0.057	(0.232)	0.050	(0.219)	-0.006	-0.024
School type: Foundation school	0.052	0.057	(0.232)	0.075	(0.265)	0.019	0.086
School type: Free school - Mainstream	0.480	0.314	(0.466)	0.296	(0.458)	-0.019	-0.038
School type: Others	0.061	0.063	(0.244)	0.050	(0.219)	-0.013	-0.072
Urban	0.828	0.893	(0.310)	0.975	(0.157)	0.082***	0.228
Rural	0.137	0.107	(0.310)	0.025	(0.157)	-0.082***	-0.25
Region: East Midlands	0.084	0.063	(0.244)	0.082	(0.275)	0.019	0.068
Region: East of England	0.108	0.107	(0.310)	0.082	(0.275)	-0.025	-0.080
Region: London	0.146	0.220	(0.416)	0.252	(0.435)	0.031	0.090
Region: North East	0.041	0.107	(0.310)	0.094	(0.293)	-0.013	-0.062
Region: North West	0.135	0.138	(0.346)	0.145	(0.353)	0.006	0.018
Region: South East	0.150	0.088	(0.284)	0.057	(0.232)	-0.031	-0.087
Region: South West	0.094	0.050	(0.219)	0.063	(0.244)	0.013	0.044
Region: West Midlands	0.113	0.119	(0.325)	0.113	(0.318)	-0.006	-0.020
Region: Yorkshire & the Humber	0.094	0.107	(0.310)	0.113	(0.318)	0.006	0.021
TP school - in addition to AM	0	0.000		0.302	(0.461)	0.302***	0.694
Census school-level % FSM Spring 2021	0.280	0.399	(0.126)	0.397	(0.115)	-0.002	-0.015
% EAL	0.152	0.198	(0.216)	0.220	(0.209)	0.021	0.118
% SEN	0.218	0.234	(0.077)	0.218	(0.065)	-0.015*	-0.187
% Female	0.498	0.488	(0.132)	0.493	(0.169)	0.004	0.026
Average AEA score	3.517	3.887	(1.818)	3.774	(1.739)	-0.113	-0.067
Average IDACI scores	0.038	0.046	(0.014)	0.048	(0.014)	0.002	0.105
% White British	0.117	0.103	(0.065)	0.099	(0.063)	-0.004	-0.055
% Asian	0.018	0.021	(0.038)	0.023	(0.037)	0.002	0.050
% Black	0.016	0.021	(0.033)	0.023	(0.031)	0.002	0.079
% Other ethnic	0.026	0.025	(0.022)	0.031	(0.024)	0.006**	0.243
% Not white	0.060	0.068	(0.067)	0.077	(0.067)	0.009	0.158
Observations		159		159			318

Source: Year 11 population data.

Note: the percentage of pupils with unknown ethnicity was suppressed because of low counts.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; AEA, Achieving Excellence Areas; EAL, English as an Additional Language; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; Ofsted, Office for Standards in Education; SEN, special education needs; KS, Key Stage; PP, Pupil Premium; SD, standard deviation; Std, standardized difference, TP, Tuition Partners.

Table 7b: Baseline characteristics of the population of AM secondary schools and AM secondary schools with pupil-level data

Variable	Means:All AM schools	SD:All AM schools	Means:AM schools in matched sample	SD:AM schools in matched sample	Difference between AM schools in matched sample and all AM schools	Std Difference
School-level PP KS2 Maths scores 2020/21	100.853	(1.788)	100.679	(1.811)	-0.174	-0.097
School-level PP KS2 Read scores 2020/21	99.619	(1.947)	99.413	(1.924)	-0.206	-0.106
School-level PP KS4 English 2020/21	4.290	(0.642)	4.260	(0.651)	-0.030	-0.044
School-level PP KS4 Maths 2020/21	4.016	(0.613)	4.011	(0.629)	-0.006	-0.025
School-level ALL KS4 English 2019/20	4.717	(0.542)	4.693	(0.535)	-0.024	-0.009
School-level ALL KS4 English 2018/19	4.284	(0.539)	4.270	(0.546)	-0.014	-0.018
School-level ALL KS4 English 2017/18	4.294	(0.559)	4.289	(0.562)	-0.005	-0.014
School-level ALL KS4 Maths 2019/20	4.567	(0.562)	4.557	(0.566)	-0.010	-0.038
School-level ALL KS4 Maths 2018/19	4.197	(0.585)	4.189	(0.599)	-0.008	-0.047
School-level ALL KS4 Maths 2017/18	4.210	(0.645)	4.186	(0.668)	-0.025	-0.009
Total pupil counts	907.163	(294.804)	933.771	(328.344)	26.608	0.087
Pupils-to-teacher ratio 2018	15.697	(2.980)	15.677	(2.660)	-0.021	-0.007
Ofsted 2018: Outstanding	0.108	(0.310)	0.126	(0.333)	0.018	0.057
Ofsted 2018: Good	0.538	(0.499)	0.553	(0.499)	0.016	0.031
Ofsted 2018: Inadequate	0.090	(0.287)	0.082	(0.275)	-0.008	-0.030
Ofsted 2018: Requires improvement	0.227	(0.419)	0.214	(0.411)	-0.013	-0.031
Ofsted 2018: Missing	0.038	(0.191)	0.025	(0.157)	-0.013	-0.070
School type: Academy-sponsor led	0.410	(0.493)	0.428	(0.496)	0.018	0.036
School type: Community school	0.096	(0.295)	0.101	(0.302)	0.005	0.016
School type: Voluntary aided/controlled school	0.073	(0.260)	0.050	(0.219)	-0.022	-0.090
School type: Foundation school	0.078	(0.269)	0.075	(0.265)	-0.003	-0.011
School type: Free school - Mainstream	0.308	(0.462)	0.296	(0.458)	-0.013	-0.027
School type: Others	0.035	(0.184)	0.050	(0.219)	0.015	0.079
TP school - in addition to AM	0.299	(0.459)	0.302	(0.461)	0.002	0.005
Urban	0.971	(0.168)	0.975	(0.157)	0.004	0.024
Rural	0.029	(0.168)	0.025	(0.157)	-0.004	0.024
Region: East Midlands	0.073	(0.260)	0.082	(0.275)	0.009	0.034
Region: East of England	0.070	(0.255)	0.082	(0.275)	0.012	0.046
Region: London	0.241	(0.428)	0.252	(0.435)	0.010	0.024
Region: North East	0.087	(0.283)	0.094	(0.293)	0.007	0.025
Region: North West	0.148	(0.356)	0.145	(0.353)	-0.004	-0.010
Region: South East	0.064	(0.245)	0.057	(0.232)	-0.007	-0.031
Region: South West	0.067	(0.250)	0.063	(0.244)	-0.004	-0.016
Region: West Midlands	0.142	(0.350)	0.113	(0.318)	-0.029	-0.086
Region: Yorkshire & the Humber	0.108	(0.310)	0.113	(0.318)	0.006	0.018
Census school-level % FSM Spring 2021	0.407	(0.115)	0.397	(0.115)	-0.010	-0.089
% EAL	0.227	(0.214)	0.220	(0.209)	-0.008	-0.036
% SEN	0.224	(0.072)	0.218	(0.065)	-0.006	-0.080
% Female	0.488	(0.165)	0.493	(0.169)	0.005	0.028
Average AEA score	3.733	(1.747)	3.774	(1.739)	0.041	0.024
Average IDACI scores	0.049	(0.014)	0.048	(0.014)	-0.000	-0.035
% White British	0.097	(0.062)	0.099	(0.063)	0.002	0.033
% Asian	0.024	(0.037)	0.023	(0.037)	-0.001	-0.019
% Black	0.024	(0.031)	0.023	(0.031)	-0.001	-0.037
% Other ethnic	0.031	(0.024)	0.031	(0.024)	-0.001	-0.022
% Not white	0.079	(0.066)	0.077	(0.067)	-0.002	-0.036
Observations	367		159		526	526

Source: Year 11 population data.
Note: the percentage of pupils with unknown ethnicity was suppressed because of low counts.
P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; AEA, Achieving Excellence Areas; EAL, English as an Additional Language; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; KS, Key Stage; Ofsted, Office for Standards in Education; PP, Pupil Premium; SD, standard deviation; SEN, special educational needs; Std, standardized difference; TP, Tuition Partners.

Baseline assessment

As well as looking at differences in background characteristics, we also ran checks to compare the samples in terms of the pupil-level and school-level KS2 results. The results are shown in Table 7c for maths and English, respectively.

When looking at school-level attainment, KS2 results for the 2020/21 cohort of pupils were well balanced between AM and non-AM schools, with no statistically significant differences between the two samples. Also pupil-level KS2 results were similarly distributed between AM and non-AM schools. Compared with the national average, AM and non-AM schools have similar KS2 results to the national population.

Table 7c: KS2 for 2021 Year 11s of AM schools, matched comparison schools, and national data

Variable	National sample	Means: Comparison	SD: Comparison	Means: AM schools	SD: AM schools	Difference between AM and matched non-AM	Std Difference
School-level PP KS2 Maths scores	100.544	100.614	(1.973)	100.679	(1.811)	0.064	0.023
School-level PP KS2 Read scores	101.196	99.439	(1.931)	99.413	(1.924)	-0.026	-0.008
Pupil-level PP KS2 Maths score	101.149	100.892	(6.718)	100.980	(6.743)	0.088	0.013
Pupil-level PP KS2 Read scores	100.082	99.433	(7.958)	99.491	(7.913)	0.058	0.007
Pupil level N	126,453	8,419		8,977		17,396	17,396

Source: Year 11 population data.

AM, Academic Mentoring; KS, Key Stage; N, number; PP, Pupil Premium; SD, standard deviation; Std, standardised difference.

Pupil-level selection

The AM programme was intended to reach disadvantaged schools. There were no prescribed conditions on the characteristics of pupils who received the intervention, although the AM programme encouraged schools participating in the programme to select pupils to receive mentoring who came from disadvantaged households or whose education had been disproportionately impacted by Covid-19. Schools were able to decide, which pupils would be selected for mentoring, and while there were no formal targets for who should be reached it was anticipated by Teach First that Pupil Premium eligibility would be one of the key markers of disadvantage that would likely inform selection of pupils into the AM programme.

The benefit of being able to use PP-eligible pupils in the analysis was to avoid the complication of pupil selection as a result of school decision or pupil choice.

For context, around 36% of pupils were eligible for Pupil Premium within the AM schools in our analysed sample (Table 8a: 36.45% of pupils). Of pupils selected for AM in our pupil-level Year 11 analysis, 46% were eligible for Pupil Premium. However, our analysis focuses on all PP-eligible pupils. Looking at it the other way around, only 13% of PP-eligible pupils were selected for AM in the intervention sample (Table 8a: 12.9% of PP-eligible pupils). Receiving AM was therefore, not a very common characteristic of PP-eligible pupils in Year 11 and our strategy to focus on PP-eligible pupils in the analysis is unlikely to identify the impact of the intervention on the population of pupils who actually received it. We refer to this issue as dilution: any effect of mentoring would be highly diluted among the PP-eligible pupils, as the analysis is on a group (PP-eligible pupils) where the majority did not participate in AM. If the proportion of PP-eligible pupils in receipt of AM had been higher, then our evaluation strategy would have avoided the complication of pupil selection as a result of school decision or pupil choice. (We had also planned to analyse based on predicted participation [RQ2], which would have provided an alternative approach to approximating the eligible group, however as reported later, this analysis could not proceed due to poor predictive power of the model.)

With only 13% of PP-eligible pupils selected for AM in the intervention sample, pupil-level selection from schools is an issue. Furthermore, Table 8a shows that the proportion of all pupils in the intervention group (AM schools) who were selected for AM is 10%. This suggests that dilution is also likely to be a problem in detecting significant impact of AM in both RQ1 and RQ3 (analysis on PP-eligible pupils and all pupils).

In addition to pupil-level and school-level selection, pupils' were selected into AM subjects. In the AM sample of 159 schools that were included in the analyses, 40 schools (25%) offered AM in maths, 43 schools (26%) offered AM in

English, and 77 schools (48%) offered AM in subjects other than English and maths (such as science, modern foreign languages, and others). It has to be acknowledged that this contributes further to dilution in the analyses.

For each subject, the dilution is worked out using two parameters: the number of PP-eligible pupils receiving AM in a given subject and the number of PP-eligible pupils in intervention schools. This results in 4.18% of PP-eligible pupils being selected for AM in maths ([Table 8b]). It also results in 2.87% of PP-eligible pupils being selected for AM in English ([Table 8b]). These are the extent of the pupil-level dilution for the RQ1 analyses.

Similarly, for all pupils, 3.41% of all pupils were selected for AM in maths and 2.1% of all were selected for AM in English ([Table 8b]). These are the extent of the pupil-level dilution for RQ3 analyses.

With such a high dilution for each subject, any impacts observed are unlikely to capture the effect that AM had on the pupils who received mentoring, given that the vast majority of pupils included in the analyses did not receive AM in a given subject.

We further explored the distribution of KS2 maths and English scores of PP-eligible pupils selected to participate in AM versus PP-eligible pupils not selected to participate in AM schools, as measured by quartiles of KS2 results, of AM pupils. The results, shown in Table 9 present the regression of TAGs on the interaction between a dummy variable equal to one for the pupil participating in AM and zero otherwise and three dummies for the quartile of the distribution of the KS2 (the base category being the lowest quartile) and a set of pupil-level and school-level controls. The purpose is to describe the ability composition of AM pupils and assess whether it is different from the ability composition of pupils not selected for AM. We performed the analysis on the sample of PP-eligible pupils, non-PP-eligible pupils, and all pupils in AM schools only. In the sample of PP-eligible pupils, for the English and maths samples, the analyses present non-statistically significant coefficients of the interaction between prior attainment quartile and AM participation, suggesting that selection of PP-eligible pupils in the AM programme was not driven by observable lower baseline assessment. In the samples of non-PP-eligible pupils and all pupils', the interaction between AM pupils and the third quartile of the KS2 distribution is negative and significant for both maths and English as well as the interaction between AM pupils and the fourth quartile of the KS2 distribution for maths and all pupils, suggesting that pupils in the middle and high distribution of baseline KS2 were less likely to be selected in AM in the samples of non-PP-eligible pupils and all pupils.

Table 8a: Percentage of PP-eligible pupils in the sample, of PP doing AM and of AM who are PP

	School counts	Pupil counts, intervention schools	Pupil counts, comparison schools	AM pupils counts, intervention schools	% of pupils who did AM, intervention schools	Pupil Premium-eligible pupils counts, intervention schools	% of Pupil Premium-eligible pupils, intervention schools	Pupil Premium-eligible pupils counts, comparison schools	% of Pupil Premium-eligible pupils, comparison schools	Pupil Premium-eligible pupils doing AM counts, intervention schools	Of pupils doing AM, % Pupil Premium-eligible pupils	Of Pupil Premium-eligible pupils, % of pupils doing AM
AM Y11 analysed sample	159 AM, 159 non-AM	24,631	23,342	2,523	10.24	8,977	36.45	8,419	36.07	1,159	45.94	12.91
AM Y11 population sample	164	28,128	n/a	2,898	10.30	10,758	38.25	n/a	n/a	1,115	38.47	10.36

Source: Year 11 population data.
AM, Academic Mentoring; PP, Pupil Premium.

Table 8b: Percentage of schools and Year 11 pupils receiving AM in maths and English

	PP-eligible pupils selected for AM	All pupils selected for AM	Pupil Premium-eligible pupils counts, intervention schools	Pupil counts, intervention schools	Of Pupil Premium-eligible pupils, % of pupils doing AM in specific subject	Of all pupils, % of pupils doing AM in specific subject
Maths	375	839	8,977	24,631	4.18	3.41
English	258	518	8,977	24,631	2.87	2.10

Source: Year 11 population data.

Table 9: Interaction of quartiles of KS2 with AM pupils' status

	PP-eligible pupils						Non-PP-eligible pupils						All pupils					
	Math			English			Math			English			Math			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
Receiving AM X KS2 Maths quartiles:																		
Math baseline 2nd quart#AM	-0.048	0.113	0.675				-0.109	0.118	0.356				-0.078	0.084	0.354			
Math baseline 3rd quart#AM	-0.211	0.130	0.107				-0.344**	0.123	0.006				-0.299**	0.096	0.002			
Math baseline 4th quart#AM	-0.230	0.187	0.222				-0.277	0.156	0.078				-0.284*	0.134	0.036			
Receiving AM X KS2 English quartiles:																		
English baseline 2nd quart#AM				-0.139	0.101	0.171				-0.153	0.111	0.171				-0.153	0.078	0.051
English baseline 3rd quart#AM				-0.109	0.130	0.405				-0.271*	0.115	0.020				-0.221*	0.092	0.017
English baseline 4th quart#AM				-0.164	0.163	0.315				-0.186	0.161	0.248				-0.231	0.121	0.058
Constant	11.152	11.313	0.326	17.137	14.864	0.251	22.253	13.399	0.099	16.380	12.795	0.202	18.289	11.786	0.123	16.656	12.875	0.198
N	8977			8977			15616			15616			24631			24631		
R-squared	0.477			0.348			0.513			0.374			0.524			0.388		

Source: Year 11 AM schools data.

First quartile is the lowest.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; KS, Key Stage; N, number; PP, Pupil Premium; quart, quarter; SE, standard error.

Preliminary analysis

This section describes the process we followed in defining the matched comparison sample we tested in the preliminary placebo analysis.

Placebo tests

Placebo testing was conducted on the sample of AM and matched comparison non-AM schools on KS4 maths and English assessments for the academic years 2016/17 to 2019/20 and it is reported in Table 10.

In Table 10, for each subject and year, we present pupil-level regressions on KS4 scores controlling for AM status, baseline KS2 results, and the variables listed in the 'Methods' section and in Appendix B. Results, across all three specifications, indicated that, before the intervention, AM schools did not show significantly different KS4 scores than matched non-AM schools, except in the academic year 2016/17, when English and maths KS4 were slightly but significantly lower in AM schools than in non-AM ones. In all the other cases, coefficients were close to zero and not significant. The fact that in the 3 years before the intervention AM schools did not present significantly different KS4 scores than non-AM schools provides support for the use of the matched comparison group as a means of estimating the counterfactual.

Table 10: Placebo test on AM schools compared to non-AM ones in KS4 maths and English in pre-intervention years, PP-eligible pupils

	KS4 Maths 2019				KS4 Maths 2018				KS4 Maths 2017				KS4 Maths 2016			
	Coef	S.E.	P-value	N	Coef	S.E.	P-value	N	Coef	S.E.	P-value	N	Coef	S.E.	P-value	N
AM schools	0.000	0.001	0.482	17,403	0.001	0.001	0.283	16,372	-0.000	0.001	0.356	14,792	-0.021**	0.007	0.002	13,808
	KS4 English 2019				KS4 English 2018				KS4 English 2017				KS4 English 2016			
AM schools	0.001	0.001	0.484	17,403	0.001	0.001	0.234	16,372	-0.000	0.001	0.926	14,792	0.007	0.006	0.261	13,808

Source: Year 11 population data.

Controls are listed in Appendix B.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; KS, Key Stage; N, number; PP, Pupil Premium; S.E., standard error.

Outcomes and analysis

Results of the Year 11 checks

In January 2021, the government announced, in the context of new national restrictions, that exams in summer 2021 could not go ahead as planned[17]. As a result of this, the secondary school analysis planned to use the TAGs awarded in 2021 instead. Several considerations about the appropriateness of using TAGs as an outcome measure that we needed to consider are listed below:

- Consideration 1: That TAGs may be distributed differently compared to previous years, particularly around the grade 3/4 boundary.
- Consideration 2: That teachers' knowledge of which pupils had been selected for AM may have led to bias (conscious or unconscious) in their awarding of TAGs to these pupils. This could lead to positive bias (as teachers know these pupils have had additional support), or negative bias (as these pupils have been previously identified as struggling).
- Consideration 3: Uncertainties around whether the TAGs would reflect pupils' performance after the tutoring. Schools may have used work produced over the year to reach their final TAGs, rather than performance in a test at a fixed time point.
- Consideration 4: That the assessments may not be sensitive enough to change as a result of pupils having received mentoring. This concern is linked to the three prior concerns, with all of these potentially affecting the measure's sensitivity to change.

Therefore, we conducted some checks (ex ante) before we started the impact analysis and (ex post) while performing the impact analysis to inform the presence of any of the above concerns, described in Appendix C.

The checks indicated that there were some differences in the way that TAGs were awarded; however, they were not able to detect with certainty the presence of any bias. Ex post 1 check indicated that a higher number of mentoring sessions received was not correlated with achieving better English and maths TAGs. This finding suggests the TAGs may not have had significant sensitivity to capture the change induced by AM (also see the results in subsection 'Weighting/regression RQ1' in the 'Results' section). Alternatively, it may also mean that AM did not have an impact on pupils, or that lower-ability pupils received more sessions. Ex ante 2 check indicated that there may have been negative bias in teachers grading AM pupils which means they were graded lower than other pupils with the knowledge that they required mentoring and/or negative selection of pupils into AM which means lower performing pupils were being selected for AM programme, however this test was conducted without a comparison group. We proceeded with the analysis on an exploratory basis, mindful of the caveats associated with doing so. Therefore, the findings need to be treated with caution.

Amount of mentoring

During ex post 1 check, we also ran a frequency distribution for pupil-level dosage data for each subject. Among the AM Year 11 pupils, the average number of sessions was 7.7 for maths and 6.7 for English. More than half of pupils received less than five sessions in maths, and less than four sessions in English, with the average (eight and seven sessions, respectively) being driven by a minority of outliers who received many more sessions. This is fewer than the AM programme's suggestion of at least one session a week for a term (e.g., 13 sessions). See Figures C3 and C4 in Appendix C for full results.

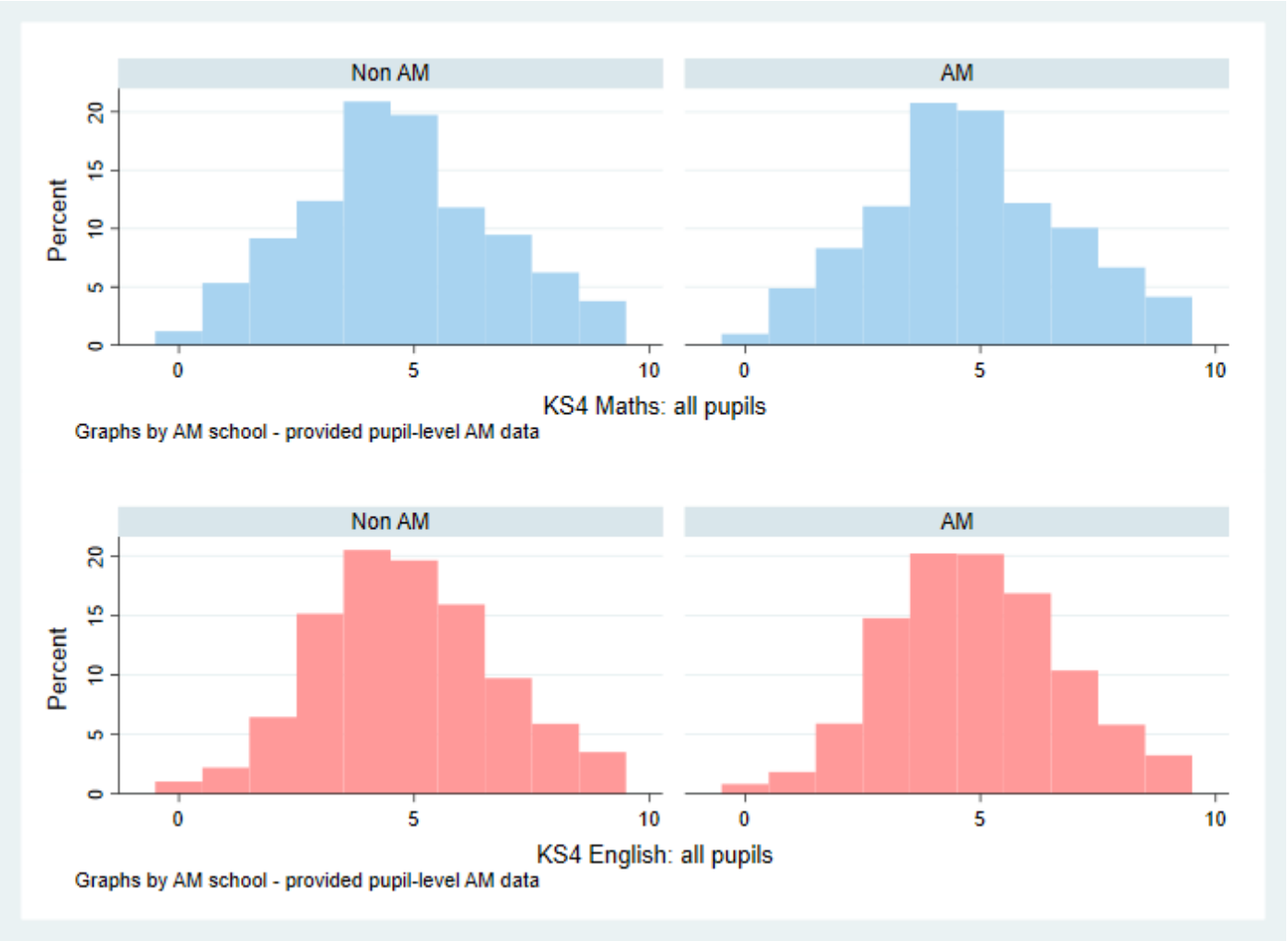
Analyses

Outcome analysis

Sets of histograms provide a graphical inspection of the distributions of the outcome measure (TAGs in 2020/21) in the analysed AM and non-AM schools (see Figure 3a for maths and Figure 3b for English). Maths and English TAGs are similarly distributed for PP-eligible pupils and for all pupils between AM and non-AM schools.

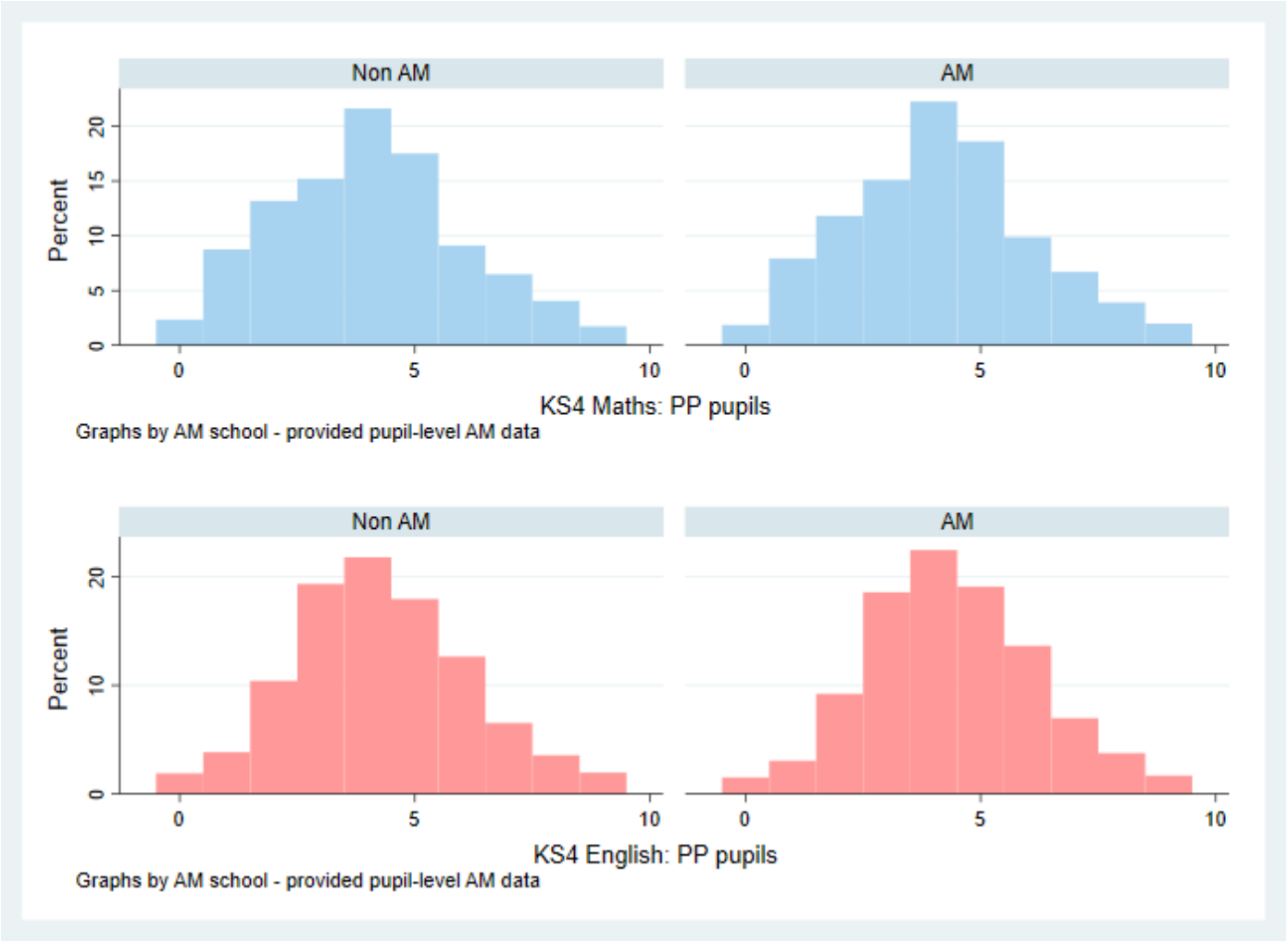
[17] https://www.gov.uk/government/publications/submission-of-teacher-assessed-grades-summer-2021-info-for-teachers/information-for-heads-of-centre-heads-of-department-and-teachers-on-the-submission-of-teacher-assessed-grades-summer-2021-html

Figure 3a: Histograms of the distributions of maths and English Teacher Assessed Grades, all pupils, in AM schools and matched non-AM ones



AM, Academic Mentoring; KS, Key Stage.

Figure 3b: Histograms of the distributions of maths and English Teacher Assessed Grades, PP-eligible pupils, in AM schools and matched non-AM ones



AM, Academic Mentoring; KS, Key Stage; PP, Pupil Premium.

Matching RQ1a: What is the impact of AM availability on PP-eligible pupils' attainment?

We present the results of AM on PP-eligible pupils by measuring AM with a 0/1 indicator for AM being available at the school level.

Results in Table 11 show the impact of AM availability in maths and in English on PP-eligible pupils under the conditional independence assumption. For English the coefficients are zero and the effect size of AM is 0.024 (CI: -0.033 to 0.082) (see Table 15 for effect sizes), ruling out large effects of AM availability. This is the equivalent of zero additional months progress in English. For maths, the coefficient is positive and significant at the 5% level. The effect size for maths is 0.053 (CI: 0.007 to 0.099) (see Table 15). This is the equivalent of one additional month progress. However, there is uncertainty around this result; it is also being consistent with a null (0 months) effect or an effect of slightly larger than 1 month's additional progress.

The specification presenting the Romano-Wolf correction for multiple testing indicates less evidence against the null hypothesis, as the coefficient is significant at the 6% level. The result indicates that AM availability in a school is associated with an increase of the average maths grade by 0.106. However, the considerably high level of dilution also means that a large proportion of PP-eligible pupils were included in the analysis who did not receive mentoring. Therefore, this estimated impact of AM is severely diluted and it is unlikely any difference seen was due to AM.

Table 11: Impact of AM measured with a 0/1 dummy indicating the availability of AM or not on the population of PP-eligible pupils

	1						2 - RW correction					
	Maths			English			Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
AM schools	0.106*	0.047	0.024	0.045	0.054	0.403	0.106	0.047	0.0594	0.050	0.054	0.3168
Constant	-5.230	7.950	0.511	-2.507	7.719	0.746	-4.397	7.924	0.579	-4.515	7.744	0.560
N	17396			17396			17396			17396		

Source: Year 11 population data (n=318 schools).

Note: School-level clustered residuals. Controls are listed in Appendix B. Column 2, reports the p-value associated with the Romano-Wolf (RW) correction.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; N, number; PP, Pupil Premium; S.E., standard error.

Further analyses

RQ2: What is the impact of AM availability on the attainment of pupils predicted to participate?

Before proceeding with RQ2, we estimated the participation equation on all pupils in AM schools to assess its predictive power. Results in Table C7 (see the 'Ex post 3 check' section), show the pupil-level participation equations for AM maths and AM English, estimated with a Logit model and a Linear Probability Model. The R-squared indicates that only less than 2.5% (3.4% in the Logit model) of total variability in AM participation is captured by observable characteristics. This suggests that schools selected pupils to participate in AM based on a number of variables that are not observable to us in the dataset, for example, motivation. The numbers in Table C8 suggest that the quality of the predictive model is sufficiently low not to warrant its use in predicting participation. Therefore, we were unable to proceed with the impact estimates for pupils predicted to participate in AM.

RQ3: What is the impact of the availability of AM on all pupils' attainment?

Similar to the approach for RQ1 (PP-eligible pupils) we present the results of a 0/1 indicator for AM being available at the school level.

Results in Table 12 show that the impact of the availability of AM was not significant in affecting maths and English TAGs. Compared to the RQ1 specification, the coefficient of maths is slightly lower (0.08 as opposed to 0.106) and the standard error slightly lower (0.042 compared to 0.047), hence the effect, although similar, is not significant. Applying the Romano-Wolf multiple hypothesis testing correction confirmed the results. The effect size for English is -0.005 (-0.055 to 0.044) (see Table 15). For maths, the effect size is 0.039 (CI: -0.001 to 0.079) (see Table 15). For both subjects this is the equivalent of zero additional months' progress.

The results on the full sample of the population have the disadvantage of considering all pupils, which includes all AM pupils whether they were PP-eligible pupils or not and suffers from the dilution issue as only 1.3% of all pupils were selected for AM in maths and 0.8% of all pupils were selected for AM in English. This means a large proportion of pupils were included in this analysis who did not receive mentoring and therefore it is hard to detect any effect that may (or may not) be present.

Table 12: Impact of AM measured with a 0/1 dummy indicating the availability of AM or not, on all pupils

	1						2 - RW correction					
	Maths			English			Maths			English		
	Coef	S.E.	p value	Coef	S.E.	p value	Coef	S.E.	p value	Coef	S.E.	p value
AM schools	0.080	0.042	0.059	-0.011	0.047	0.820	0.079	0.042	0.0792	-0.005	0.048	0.9604
Constant	-3.275	6.792	0.630	-0.135	6.658	0.984	-2.834	6.786	0.677	-2.229	6.783	0.743
N	47973			47973			47973			47973		

Source: Year 11 population data (n=318 schools).

Note: School-level clustered residuals. Controls are listed in Appendix B. Column 2 reports the p-value associated with the Romano-Wolf (RW) correction.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; N, number; S.E., standard error.

Moderator analysis

RQ4: How does the impact of AM availability vary among PP-eligible pupils, by school and pupil characteristics?

Table 13a and Table 13b present the results of the interaction between AM schools and a set of school-level, pupil-level, and geographic characteristics on PP-eligible pupils' TAGs. Results were separately produced for maths and English. The tables report the marginal effects to facilitate the interpretation of the interactions. The marginal effects were calculated from predictions of a fit model at fixed values of some covariates and averaging over the remaining covariates.

There are different caveats to this analysis. First, results are not causal. Second, the majority of PP-eligible pupils did not receive the intervention so the sample selected does not coincide with the individuals who received the intervention. Third, as we are testing multiple hypotheses, some of the coefficients could be statistically significant by chance. Results should be interpreted with caution and we do not recommend drawing any conclusions or recommendations from this RQ.

The results were the following for the maths sample:

- Pupils in AM schools with a high proportion of FSM pupils were associated with slightly higher maths TAGs compared to non-AM schools (coefficient 0.11).

- Pupils in AM schools with the lowest quartile of pupils counts (below 25% of the distribution of counts in the analysed sample) were associated with slightly higher maths TAGs compared to non-AM schools (coefficient 0.16).

- Female students, students with SEN, and students without EAL were associated with higher maths TAGs in AM schools compared to the same pupils in non-AM schools (coefficients: 0.13 for female; 0.16 for SEN; and 0.14 for non-EAL).

- Students with unknown and other ethnicity were associated with higher maths TAGs in AM schools compared to non-AM schools (coefficients: 0.20 for Any other; 0.44 for Unknown ethnicity).

- Students with KS2 results above 100 were associated with higher TAGs in AM schools compared to non-AM schools (coefficients: 0.1 for KS2 results higher than 100; 0.12 for KS2 higher than 105; 0.15 for KS2 higher than 110; and 0.17 for KS2 higher than 115).

- AM schools in urban areas and in above median IDACI scores were associated with slightly higher maths TAGs compared to non-AM schools (coefficient: 0.10 in both cases).

For English, the significant interactions are the following:

- AM schools with missing Ofsted ratings were associated with higher English TAGs compared to non-AM schools with missing Ofsted ratings (0.37). Missing Ofsted ratings may be positively correlated with TAGs in AM schools.

- AM schools with the lowest quartile of pupils counts (below 25% of the distribution of counts in the analysed sample) were associated with slightly higher English TAGs compared to non-AM schools (0.15).

- AM schools in rural areas were associated with lower English TAGs compared to non-AM schools (-0.40).

Table 13a: How does the impact of AM availability on maths vary among PP-eligible pupils, by school and pupil characteristics?

	Interact w/ school-level			Interact w/ pupil-level			Interact w/ geography		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
Ofsted High vs Low									
Ofsted Low	0.104	0.099	0.292						
Ofsted High	0.082	0.058	0.156						
Ofsted Missing	0.353	0.186	0.059						
School %FSM High vs Low:									
School %FSM high vs low=0	0.077	0.086	0.370						
School %FSM high vs low=1	0.107	0.056	0.059						
Maintained vs Non-maintained schools:									
Academy/Free school	0.092	0.053	0.081						
Maintained school	0.128	0.094	0.176						
Pupil counts Q4:									
Q1	0.156*	0.075	0.040						
Q2	0.114	0.072	0.116						
Q3	0.081	0.095	0.392						
Q4	0.041	0.140	0.769						
Female=0				0.132*	0.051	0.010			
Female=1				0.071	0.059	0.228			
Non-SEN				0.087	0.049	0.078			
SEN				0.163*	0.073	0.027			
Non-EAL				0.138**	0.050	0.007			
EAL				-0.009	0.076	0.906			
EAL unknown				-0.590	0.394	0.135			
White British				0.060	0.058	0.299			
Asian				0.100	0.095	0.291			
Black				0.122	0.094	0.194			
Other ethnic				0.202**	0.075	0.007			
Unknown ethnic				0.443*	0.182	0.015			
KS2 Maths scores 90-115:									
90				0.047	0.068	0.487			
95				0.072	0.054	0.189			
100				0.096*	0.048	0.045			
105				0.121*	0.051	0.018			
110				0.146*	0.062	0.019			
115				0.170*	0.077	0.028			
Rural vs Urban:									
urban							0.101*	0.049	0.041
rural							0.100	0.190	0.600
IDACI scores:									
Low							0.114	0.072	0.114
High							0.099*	0.050	0.049
IDACI Missing							-0.490	0.418	0.242
Observations	17,396			17,396			17,396		

Source: Year 11 population data (n=318 schools).

Note: School-level clustered residuals. Controls are listed in Appendix B. The table presents estimated margins, i.e., the marginal effect of each interaction, estimated using counterfactual analysis.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; EAL, English as an Additional Language; Coef, coefficient; FSM, Free School Meals; KS, Key Stage; N, number; Ofsted, Office for Standards in Education; PP, Pupil Premium; Q, Quartile; S.E., standard error; SEN, special educational needs.; IDACI, Income Deprivation Affecting Children Index.

Table 13b: How does the impact of AM availability on English vary among PP-eligible pupils, by school and pupil characteristics?

	Interact w/ school-level			Interact w/ pupil-level			Interact w/ geography		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
Ofsted High vs Low									
Ofsted Low	-0.066	0.101	0.515						
Ofsted High	0.070	0.060	0.245						
Ofsted Missing	0.370*	0.163	0.023						
School %FSM High vs Low:									
School %FSM high vs low=0	0.053	0.078	0.500						
School %FSM high vs low=1	0.039	0.064	0.539						
Pupil counts Q4:									
Q1	0.151*	0.073	0.040						
Q2	0.008	0.088	0.927						
Q3	0.058	0.109	0.593						
Q4	-0.065	0.136	0.633						
Female=0				0.058	0.061	0.343			
Female=1				0.037	0.061	0.538			
Non-SEN				0.045	0.057	0.435			
SEN				0.060	0.071	0.400			
Non-EAL				0.049	0.055	0.378			
EAL				0.046	0.089	0.605			
EAL unknown				-0.132	0.332	0.691			
White British				0.006	0.061	0.917			
Asian				0.142	0.103	0.169			
Black				0.058	0.094	0.542			
Other ethnic				0.082	0.085	0.334			
Unknown ethnic				0.236	0.205	0.251			
KS2 Read scores 90-115:									
90				0.108	0.060	0.072			
95				0.076	0.054	0.162			
100				0.044	0.054	0.417			
105				0.012	0.059	0.842			
110				-0.020	0.069	0.767			
115				-0.053	0.081	0.516			
Rural vs Urban:									
urban							0.060	0.055	0.279
rural							-0.400*	0.158	0.012
IDACI scores:									
Low							-0.004	0.078	0.959
High							0.052	0.056	0.350
IDACI Missing							-0.075	0.656	0.909
Observations	17396			17396			17396		

Source: Year 11 population data (n=318 schools).

Note: School-level clustered residuals. Controls are listed in Appendix B. The table presents estimated margins, i.e., the marginal effect of each interaction, estimated using counterfactual analysis.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; EAL, English as an Additional Language; Coef, coefficient; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; KS, Key Stage; N, number; Ofsted, Office for Standards in Education; PP, Pupil Premium; Q, Quartile; S.E., standard error; SEN, special educational needs; IDACI, Income Deprivation Affecting Children Index.

RQ5: How do outcomes vary among AM pupils, by model of mentoring?

This analysis is based only on the AM pupils in AM schools as here we are exploring differences in delivery (in contrast to previous RQs that compare AM schools with the comparison non-AM schools). Consequently, the results reported here should not be considered causal associations, but instead as descriptive findings that may be indicative. We did not control for the pupil: mentor ratios as indicated in the study plan, as the variable had many missing values. We controlled for the same set of controls used in previous specifications.

Results, reported in Table 14, indicate that:

- Face-to-face delivery is associated with worse English TAGs than online delivery.
- Male mentors are associated with better English TAGs. Similar analysis for maths showed no significant difference.

- Academic Mentors with a QTS were associated with better maths TAGs.

As shown in Table 14, we found no association between the number of sessions pupils received and their outcomes in English or maths. Among the AM Year 11 pupils, the average number of sessions was 7.7 for maths and 6.7 for English (see Table C5 and Figures C3 and C4 in Appendix C for a similar analysis and histograms of dosage).

Table 14: Estimates of KS4 maths and English grades by model of tutoring, among AM pupils

	Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value
Total Maths sessions completed	0.007	0.009	0.442			
Baseline online						
F2F	0.279	0.197	0.165			
Online and F2F	-0.010	0.273	0.970			
Total English sessions completed				-0.003	0.011	0.799
Baseline online						
F2F				-0.874*	0.398	0.034
Online and F2F				-0.364	0.270	0.184
Mentor Male	0.571	0.491	0.251	0.764*	0.284	0.010
Tutor: same gender as pupil	0.029	0.140	0.839	0.156	0.128	0.232
Graduation year	0.045	0.023	0.062	0.026	0.016	0.112
Qualified Teacher Status	1.751***	0.399	0.000	0.170	0.658	0.797
N	785			482		
R-squared	0.621			0.423		

Source: Year 11 population data, AM schools (n=159 schools).

Note: School-level clustered residuals. Controls are listed in Appendix B.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; F2F, face-to-face; KS, Key Stage; N, number; S.E., standard error.

Missing data analysis

The missing data analysis indicates that few pupils were dropped from the analysis because of missing TAGs or missing KS2 maths and English results. There were 926 pupils with missing data for KS2 and 83 pupils with missing data for TAGs in AM schools and 772 pupils with missing KS2 results, 88 pupils with missing TAGs in non-AM schools. The majority of missing data is concentrated in KS2 results, which occurred years before the intervention, and hence it should not lead to any bias.

Estimation of effect sizes

The estimates of RQ1 were an effect size of 0.0532 for maths (95% CI: 0.0072 to 0.0991) and 0.0245 for English (95% CI: -0.0329 to 0.0820) as shown in Table 15. The effect size for maths reached a conventional level of significance, while it did not for English.

Table 15: Primary analysis, RQ1 and RQ3

	Intervention group	Comparison group	Effect size		
Outcome	n (missing)	n (missing)	Total n (missing)	Hedges g (95% CI)	p-value
Maths Teacher Assessed Grade (TAG) (RQ1)	8,977 (1,781)	8,419 (76,964)	17,396 (78,745)	0.053 (0.007 to 0.099)	0.023
English TAG (RQ1)	8,977 (1,781)	8,419 (76,964)	17,396 (78,745)	0.025 (–0.033 to 0.082)	0.402
Maths TAG (RQ3)	24,631 (2,737)	23,342 (2,238)	47,973 (5,065)	0.039 (–0.001 to 0.079)	0.058

English TAG (RQ3)	24,631 (2,737)	23,342 (2,238)	47,973 (5,065)	–0.006 (–0.055 to 0.044)	0.820

Source: Year 11 population data.

Note: The missing values are calculated as pupils eligible for matching who are not analysed. The numbers include pupils in schools that were not matched.

CI, confidence interval; n, number; RQ, research question.

Conclusion

Table 16: Summary of findings

Finding
Initial checks on the data indicated that the TAGs would be suitable as an outcome measure for some exploratory analysis into the impact of AM, in the absence of any other outcome data. However, because the TAGs were a new and unique assessment for which there is no prior data to compare to, the findings reported below should be considered exploratory and should be interpreted with caution.
Year 11 pupils eligible for Pupil Premium in schools that received AM made, on average, similar progress in English compared to Year 11 pupils eligible for Pupil Premium in comparison schools (there was no evidence of an effect). In maths, Year 11 pupils eligible for Pupil Premium in schools that received AM made, on average, slightly more progress (equivalent to 1 months' additional progress) compared to Year 11 pupils eligible for Pupil Premium in comparison schools. However, there is uncertainty around this result; it is also consistent with a null (0 months) effect or an effect of slightly larger than 1 month's additional progress. A particular challenge in interpretation is that, on average, only 13% of Year 11 pupils eligible for Pupil Premium were selected for mentoring by schools, and only 4.2% of Year 11 pupils eligible for Pupil Premium were selected for mentoring in maths and 2.9% in English, meaning that the vast majority of pupils eligible for Pupil Premium included in the analysis did not receive mentoring. Therefore, this estimated impact of AM is severely diluted and it is unlikely any of these differences were due to AM.
When looking at all Year 11 pupils, pupils in schools that received AM made, on average, similar progress in English and maths compared to all Year 11 pupils in comparison schools (there was no evidence of an effect). However, this finding was similarly subject to severe dilution: on average only 10% of Year 11 pupils in the analysed schools were selected for mentoring, with 3.4% in maths and 2.1% in English, and therefore it is hard to detect any effect that may (or may not) have been present.
Within schools that offered AM to Year 11 pupils, there was no association between the number of completed mentoring sessions in maths and Year 11 outcomes in maths, or between the number of completed mentoring sessions in English and Year 11 outcomes in English. These results are associations and not necessarily causal.

Impact evaluation and IPE integration

Interpretation

The findings reported here should be considered exploratory due to the use of the TAGs as an outcome measure. We ran some initial checks on the data to investigate whether it would be appropriate to use them as an outcome measure and to review their suitability for this study. This is in no way a comment or reflection on the TAGs as an assessment mechanism. There is some evidence from one of our tests that there may have been negative bias and/or negative selection of pupils into AM (which means lower performing pupils were being selected for the AM programme), however this test was conducted without a comparison group. It does not appear that there were systematic differences in grading between AM and non-AM schools over the exam years analysed (2021, 2020, 2019, and 2018), although it is not possible to confirm that this is certainly the case. The check on dosage indicates that there may not be sufficient sensitivity to pick up changes induced by the AM programme in the sample of AM schools. We proceeded with the analysis on an exploratory basis. The use of the TAGs is discussed further in the 'Limitations and lessons learned' section below.

This study aimed to evaluate the impact of the AM programme on pupil attainment and was designed to do so using a QED involving several estimators of impact. The primary analysis result suggests a small positive impact of the availability of AM in Year 11 on maths TAGs for PP-eligible pupils. The coefficient for the specification with all pupils is slightly lower and not significant (and equivalent to zero month's additional progress compared with the comparison group). The impact of availability of AM in Year 11 on English TAGs is not significant in both specifications (PP-eligible pupils and all pupils). However, these impact findings must be treated with a high degree of caution. These analyses were subject to very high dilution (that is, they were registering the outcomes of the vast majority of PP-eligible pupils who were not selected for AM in English or in maths). We note, 77 schools (48%) offered AM in subjects other than English and maths (such as science, modern foreign languages, and others).

The issue of dilution is important to understand here (also see the discussion on limitations, below). It has not been possible to precisely identify the counterfactual at a pupil level, that is, it was not possible, using the data available, to accurately select a group of pupils who would have participated in AM from non-AM schools despite efforts to do so. For a pupil-level intervention such as AM, this is a major challenge to its evaluation. The analysis was conducted on proxy groups that may give an indication of impact, but as the vast majority of the pupils in the analysed intervention group (whether considering the Pupil Premium analysis or all pupils analysis) did not in fact participate in AM, the dilution of any effect is likely to have meant the analyses were underpowered.

As noted, we detected a small positive effect for maths in the analysis on the availability of AM to PP-eligible pupils in Year 11. However, with such high dilution, the effect would need to be strong among the small proportion of Year 11 PP-eligible pupils who were selected for tutoring (and we do not have an indication that this was the case elsewhere in our analysis), and/or there would need to be strong spillover effects among the rest of the Year 11 PP-eligible pupils (although the programme Theory of Change includes such a mechanism, it is unlikely to be relevant at the dilution levels seen). Another possibility is that the Year 11 PP-eligible pupils not selected for AM may have received some other form of maths support that was having an effect. However, we are not aware of this from the AM IPE report, and this would need to be more so, indeed differentially so, to any such support being provided in the comparison group (and we have no reason to suspect this).

However, with such high dilution, it is hard to detect whether AM had an effect on those who received mentoring in the analyses focusing on PP-eligible pupils and on all pupils. This does not detract from the analysis being relevant from a policy perspective. However, we cannot conclude whether a lack of observed impact is due to the small proportion of disadvantaged pupils who received mentoring, or because AM did not work for those who received it.

Furthermore, we found no association between the number of AM sessions in maths or English that pupils received and their attainment in maths or English, respectively. It is also worth considering the number of AM sessions that Year 11 pupils received. Year 11 pupils selected for AM received on average eight sessions in maths and seven sessions in English (prior to the TAG cut-off date) (and more than half of Year 11 pupils received less than five sessions in maths, and less than four sessions in English). This is fewer than the AM programme's suggestion of at least one session a week for a term and thus may also have made it hard to detect an effect of mentoring. Indeed, for maths, the effect detected seems even harder to explain—as pupils received fewer sessions than suggested.

We note that the AM programme did not specify an optimum number of sessions or hours of mentoring. Although the lack of association between number of sessions received and outcomes might suggest that AM did not work for those who received it, this could also be explained by a lack of sensitivity in TAGs in AM analysis. We also note that the AM data recorded the number of sessions rather than the number of hours received. While this is a limitation in the data, we would assume that with Year 11s, a mentoring session might take place during a normal secondary school lesson time, and so analysing the number of sessions (rather than hours) is not unduly problematic. However, given that the number of sessions may not necessarily reflect the number of hours received, we recommend further research into the optimum dosage needed to improve pupils' attainment.

The analysis suggested a number of other moderators were associated with outcomes. For example, the findings indicate that academic mentors with a QTS (rather than without) are associated with better maths TAGs, and male (rather than female) academic mentors are associated with better English TAGs. Face-to-face delivery is associated with worse English TAGs when compared to online delivery in English. Given the exploratory nature of this analysis, and the limitations (discussed below), we recommend further research into the different features of mentoring and how they might be associated with outcomes.

The AM programme was initiated and delivered at a time of great pressure for schools, when the education system had been disrupted by a series of school closures to most pupils, and was contending with ongoing widespread pupil and staff absences. Covid-related issues disrupted the anticipated operation of academic mentoring during the year. The AM programme involved initial training and ongoing support from Teach First as intended but there was greater variation in schools' deployment of mentors during the latter stages of the Autumn Term 2020/21, and during the January to March 2021 period of restricted attendance in schools. Staff absences due to Covid-19 were particularly high and academic mentors with QTS were used in some schools to provide teaching cover or to assist with teaching of key worker and vulnerable children attending school. However, Teach First's process evaluation suggested only 20% of Academic Mentors were used in this way for a portion of their time in role (Teach First, 2021).

The AM programme was backed by central investment and support, but it was not the only way schools chose to support their pupils, and we have not been able to account for other initiatives and practices that non-AM schools may have been deploying to support their pupils. Our study did not explore any learning recovery strategies in place in the comparison/non-AM schools. However, we know from other studies (Rose *et al*., 2021; Nelson *et al*., 2021; Harland *et al*., 2022) that schools across England were putting in place a range of recovery strategies and support, and so it is likely that all schools in the evaluation may have been recovering to some extent. This makes it harder to isolate the specific effect of the mentoring support, which may have been part of a mix of support that schools were putting in place

(schools could also use the 'one-off universal' catch-up premium[18] for learning recovery and comparison schools may well have been using tutoring sourced from outside the NTP).

As noted above, this analysis is exploratory and the results should be interpreted as such—the limitations of the research are discussed in the next section.

Limitations and lessons learned

This evaluation focuses on the impact of the AM programme for Year 11 pupils. We note that not all of the original planned analysis could go ahead. Originally, we had planned to evaluate the impact for Year 6 pupils (using KS2 data), and to evaluate the impact at primary and secondary school using schools' standardised assessment data from RL assessments. However, as KS2 assessments were cancelled in summer 2020, and there was no national alternative data, we removed the Year 6 analysis from the study. We also removed the RL evaluation samples (primary school Year 1–6 and secondary school Year 7–10) from the study design; after a change in data sharing arrangements by RL, despite considerable efforts to re-contact schools, the number of schools providing agreement was insufficient to warrant impact analyses on the evaluation samples using RL data.

The study therefore only includes Year 11 pupils. We had originally intended to use GCSE grades as the outcome measure. However, GCSEs were cancelled in summer 2020 and replaced by TAGs, and so we amended our study design to include a number of considerations prior to using the TAGs.

The first study limitation concerns the validity of TAGs as an outcome measure for research. There were several considerations about the appropriateness of using TAGs as an outcome measure that we have made. The considerations arose because of our intended use of the TAGs as an outcome measure: the TAGs were not introduced with this use in mind and given the speed of their introduction necessitated by the circumstances, information that we would usually draw on to determine the reliability and validity of an outcome measure was not available. However, in the absence of any other outcome data we felt that some exploratory analysis, as reported here, was better than not trying to estimate impact at all. To mitigate against this, we carried out some initial checks on the data as to its likely suitability for our purpose.

First, the checks performed suggest that grades were, on average, higher in 2021 than in previous years, which was in line with data published by Ofqual (2021). However, once we control for school-level observable characteristics, they do not point towards systematic differences across AM and non-AM schools in the way they allocated TAGs. Second, the checks indicate that the TAGs may not have significant sensitivity to capture change induced by the dosage (we note that dosage was recorded in terms of number of sessions, which may not necessarily reflect the number of hours received). The tests point also towards negative pupil selection into the AM programme, in terms of attainment or negative bias in awarding grades. This is unsurprising given the purpose of AM and is not a concern for the Pupil Premium or 'all pupils' analysis.

The second study limitation concerns the issue of dilution, which is an important one for this analysis. Our original design introduced a range of RQs designed to complement each other as a counterbalance in the event of different participation selection mechanisms employed by schools. We anticipated that, due to the focus on supporting disadvantaged pupils and the guidance provided to schools, that a high proportion of PP-eligible pupils would be selected for AM. We also planned to predict, which pupils would participate in AM using the data available. We anticipated that one or other of these would enable us to identify a good counterfactual in the comparison schools. However, in the event, neither of the strategies were successful due to the way pupils were selected for AM. All participating schools had an above average proportion of pupils eligible for Pupil Premium, and of pupils selected for AM in our pupil-level Year 11 analysis 46% were eligible for Pupil Premium. However, despite this the analysis suffered from severe dilution. This is because the analysis was based on all Year 11 PP-eligible pupils—an identifiable group in the data—and of Year 11 PP-eligible pupils, only a small proportion were selected for AM. The maths analysis accounts for pupils in the 40 schools that offered AM in maths and the English analysis accounts for pupils in the 43 schools who offered mentoring in English. On average, 4.18% of PP-eligible pupils were selected for AM in maths in AM schools; and 2.87% of PP-eligible pupils were selected for AM in English in AM schools. (We also note that these low proportions are driven by the extent to which PP-eligible/non-PP-eligible pupils were selected for AM, and also by the total number of Year 11 pupils selected for AM in each school.) In addition to this, we were not able to identify the pupils who would have participated in AM in the non-AM schools because although the participating schools fulfilled the school-level eligibility criteria, they used information to select pupils into the programme that is not observable in the datasets, suggesting that pupil-level

[18] In the academic year 2020/21, there was a one-off universal £650 million catch-up premium provided by the government to support schools to provide catch-up activities to help pupils make up for lost teaching time.

selection was driven by unobservable dimensions and thus could not be accounted for in the analysis. Therefore, the analysis reports the impact on attainment of the availability of AM on specific groups of pupils that we are able to identify in both intervention and comparison schools (specifically PP-eligible pupils and all pupils). Taken together, this means that our estimates are for groups of pupils that do not directly align with the group of pupils that participated in AM.

The third study limitation concerns the study design: neither schools nor pupils were randomly assigned to treatment and control groups. Given the urgency of addressing missed learning in schools and supporting catch up and recovery, it would not have been considered ethical to randomise. Moreover, given the timing of the evaluation commission, the QED approach was appropriate as delivery was already underway. QEDs are the next best impact evaluation tool, but they have challenges and limitations, chiefly relating to creating a suitable comparison group. In this evaluation, the schools that signed up to the intervention represented the treatment group, and the comparison group was selected through a matching procedure on the basis of observable characteristics. One consequence of this is that not all demographic characteristics and outcomes were balanced at the baseline. However, we controlled for these imbalances in the outcome models. Another possible consequence of this QED is that unobserved characteristics may have affected the treatment efficacy instead of, or in addition to, the AM intervention. Given our evaluation design was based on recent research by Weidmann and Miratrix (2020) we are reasonably confident that we removed school-level selection bias in our comparisons. However, our inability to address pupil-level selection bias severely limits the conclusions we can draw.

This analysis is based on the participation and monitoring data supplied by academic mentors to Teach First and shared with the evaluator. However, not all academic mentors returned data to Teach First, and where it was supplied there were some gaps and inconsistencies in the dataset. In particular, the number of schools in the analysis is lower than we anticipated at the study plan stage because pupil-level data were submitted only for a subsample of schools that received AM, limiting the sample of schools deemed eligible for the analysis.

Future research and publications

As noted in the ‘Limitations and lessons learned’ section, no study could have been fully equipped to deal with the changes to national assessments or to the way in which schools actually selected pupils to participate in AM. Given the exploratory nature of this analysis, and the large-scale nature of the intervention, future evaluations might look to explore the presence of more firm evidence on the impact of AM.

It would be important to address pupil-level selection, if possible, specifically to ensure the selection of pupils is not endogenously determined. Pupil selection for mentoring was not simply influenced by Pupil Premium status nor by observable pupils’ characteristics. This caused issues for the evaluation’s ability to identify a suitable counterfactual group of pupils; similar pupils who did not participate in mentoring. If the period of lockdowns and disruption to education is at an end, there may be the potential for randomised controlled trials to be conducted, which would help avoid the issue of selection on unobservable characteristics, which is so ingrained here.

We recommend that in future years of the AM programme, efforts are made to evaluate the mentoring with a randomised design, for example by randomising the order in which schools and groups of pupils within year groups receive it, or by varying the number of hours of mentoring to establish the optimum dosage. This could be part of a programme of future evaluation to help explore, which models of tutoring/mentoring are most effective for which pupils and in what circumstances. It would be important to include both quantitative assessments of impact and IPE to understand different models and contexts.

References

Clarke, D., Romano, J. and Wolf, M., 2020. The Romano–Wolf multiple-hypothesis correction in Stata. *The Stata Journal*, 20(4), pp.812–843. https://doi.org/10.1177/1536867X20976314.

Cullinane, C. and Montacute, R., 2020. *COVID-19 and social mobility impact brief #1: school closures*. [online] Available at: <https://www.suttontrust.com/wp-content/uploads/2021/01/School-Shutdown-Covid-19.pdf> [Accessed 29 July 2022].

Dietrichson, J., Bøg, M., Filges, T. and Klint Jørgensen, A.-M., 2017. Academic interventions for elementary and middle school students with low socioeconomic status: a systematic review and meta-analysis. *Review of Educational Research*, 87(2), pp.243–282. https://doi.org/10.3102/0034654316687036.

Educational Endowment Foundation, 2020. *Impact of school closures on the attainment gap: rapid evidence assessment*. [online] Available at: <https://educationendowmentfoundation.org.uk/public/files/EEF_(2020)_-_Impact_of_School_Closures_on_the_Attainment_Gap.pdf>.

Educational Endowment Foundation, 2021a. *One to one tuition*. [online] EEF. Available at: <https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/one-to-one-tuition> [Accessed 29 July 2022].

Educational Endowment Foundation, 2021b. *Small group tuition*. [online] EEF. Available at: <https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/small-group-tuition> [Accessed 29 July 2022].

Harland, J., Bradley, E., Staunton, R., Lord, P., Sims, D. and Styles, B., 2021. *National Tutoring Programme – Tuition Partners reach and engagement randomised controlled trial: testing different types of recruitment emails*. [online] Available at: <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/national-tutoring-programme> [Accessed 29 July 2022].

Harland, J., Fletcher, L., Morton, C., Lord, P. and Styles, B., 2022. *Learning recovery in Yorkshire and the Humber*. [online] Available at: <https://www.nfer.ac.uk/media/4901/learning_recovery_in_yorkshire_and_the_humber.pdf> [Accessed 29 July 2022].

Hedges, L., 1981. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6(2), pp.107–128. https://doi.org/10.3102/10769986006002107.

Nelson, J., Lynch, S. and Sharp, C., 2021. *Recovery during a pandemic: the ongoing impacts of Covid-19 on schools serving deprived communities*. [online] Available at: <https://www.nfer.ac.uk/media/4614/recovery_during_a_pandemic_the_ongoing_impacts_of_covid_19_on_schools_serving_deprived_communities.pdf> [Accessed 29 July 2022].

Nickow, A., Oreopoulos, P. and Quan, V., 2020. *The impressive effects of tutoring on Prek-12 learning: a systematic Review and meta-analysis of the experimental evidence*. [National Bureau of Economic Research working paper] Available at: <https://www.nber.org/papers/w27476> [Accessed 26 September 2022].

Ofqual, 2021. Guide to GCSE results for England, 2021. *GOV.UK*. [online] 12 Aug. Available at: <https://www.gov.uk/government/news/guide-to-gcse-results-for-england-2021> [Accessed 29 July 2022].

Romano, J. and Wolf, M., 2005a. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), pp.94–108. https://doi.org/10.1198/016214504000000539.

Romano, J. and Wolf, M., 2005b. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), pp.1237–1282. https://doi.org/10.1111/j.1468-0262.2005.00615.x.

Rose, S., Badr, K., Fletcher, L., Paxman, T., Lord, P., Rutt, S., Styles, B. and Twist, L., 2021. *Impact of school closures and subsequent support strategies on attainment and socio-emotional wellbeing in Key Stage 1*. [online] Available at: <https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/Impact-on-KS1-Closures-Report.pdf?v=1638448453> [Accessed 29 July 2022].

Sharp, C., Nelson, J., Lucas, M., Julius, J., McCrone, T. and Sims, D., 2020. *The challenges facing schools and pupils in September 2020*. [online] Available at: <https://www.nfer.ac.uk/media/4119/schools_responses_to_covid_19_the_challenges_facing_schools_and_pupils_in_september_2020.pdf> [Accessed 29 July 2022].

Teach First, 2021. *Academic Mentoring Programme (AMP) Phase 1 process report 2020/21*. [online] Available at: <https://www.teachfirst.org.uk/sites/default/files/2021-11/AMP%20process%20report_v5_Nov2021_0.pdf> [Accessed 29 July 2022].

Torgerson, C., Bell, K., Coleman, E., Elliott, L., Fairhurst, C., Gascoine, L., Hewitt, C. and Torgerson, D., 2018. *Tutor Trust : Affordable primary tuition. Evaluation report and executive summary*. [online] Available at: <https://dro.dur.ac.uk/26952/1/26952.pdf?DDD29+vrfd57+d700tmt> [Accessed 29 July 2022].

UCL, 2020. *Briefing note: inequalities in resources in the home learning environment*. [online] Available at: <https://discovery.ucl.ac.uk/id/eprint/10114836/1/cepeobn2.pdf> [Accessed 29 July 2022].

Weidmann, B. and Miratrix, L., 2020. Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *Journal of Policy Analysis and Management*, 40(3), pp.964–986. https://doi.org/10.1002/pam.22236.

Appendix A

Table A1: Probit of AM participation

	Coef	S.E.	p-value
AM school - provided pupil-level AM data			
AEA score=2	0.034	0.214	0.874
AEA score=3	0.212	0.239	0.374
AEA score=4	0.411	0.264	0.120
AEA score=5	0.642*	0.311	0.039
AEA score=6	0.480	0.371	0.196
IDACI scores from Teach First	0.016**	0.005	0.004
Census school-level % FSM Spring 2021	0.462	0.733	0.528
IDACI x AEA	-0.001	0.001	0.362
School-level ALL KS4 English 2019/20	-0.445**	0.155	0.004
School-level ALL KS4 Maths 2019/20	0.184	0.147	0.211
East of England	0.157	0.212	0.460
London	0.647**	0.250	0.010
North East	0.425	0.245	0.083
North West	-0.008	0.207	0.969
South East	-0.235	0.229	0.304
South West	0.194	0.235	0.410
West Midlands	-0.074	0.200	0.710
Yorkshire and the Humber	-0.010	0.205	0.960
Community school	-0.130	0.167	0.437
Voluntary aided school	-0.363	0.210	0.084
Foundation school	0.057	0.193	0.767
Free school - Mainstream	-0.176	0.115	0.128
Special free school	-0.067	0.239	0.779
Free school UTC	-0.837	0.470	0.075
KS1-to-KS2 value added 2018 at local district level	0.049	0.152	0.747
Constant	-6.090	15.123	0.687
Observations	2215		
Pseudo R-squared	0.171		

Source: Year 11 population data.
Missing variables of all controls included but not listed.
P-values: * <0.1; ** <0.05; *** <0.001.
AEA, Academic Excellence Areas; AM, Academic Mentoring; Coef, coefficient; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; KS, Key Stage; S.E., standard error; UTC, university technical college.

Table A2: Baseline characteristics of Year 11 AM schools, eligible comparison schools and national proportions

Variable	Means: AM schools	SD: AM schools	Means: Eligible non AM	Eligible non-AM	Difference between AM schools and all eligible non-AM schools	Std Difference
School-level PP KS2 Maths scores 2020/21	100.630	(1.870)	100.679	(1.811)	0.049	0.027
School-level PP KS2 Read scores 2020/21	99.362	(1.867)	99.413	(1.924)	0.051	0.028
School-level PP KS4 English 2020/21	4.108	(0.616)	4.260	(0.651)	0.152***	0.25
School-level PP KS4 Maths 2020/21	3.821	(0.638)	4.011	(0.629)	0.190***	0.309
School-level ALL KS4 English 2019/20	5.139	(0.770)	4.693	(0.535)	-0.445***	-0.974
School-level ALL KS4 English 2018/19	4.755	(0.834)	4.258	(0.547)	-0.496***	-0.988
School-level ALL KS4 English 2017/18	4.757	(0.838)	4.244	(0.572)	-0.513***	-0.998
School-level ALL KS4 Maths 2019/20	5.021	(0.901)	4.557	(0.566)	-0.464***	-0.911
School-level ALL KS4 Maths 2018/19	4.717	(0.959)	4.176	(0.601)	-0.541***	-0.955
School-level ALL KS4 Maths 2017/18	4.715	(0.980)	4.134	(0.681)	-0.581***	-0.949
Pupil counts	917.773	(329.494)	933.604	(328.301)	15.831	0.048
Pupils-to-teacher ratio 2018	16.306	(2.683)	15.668	(2.676)	-0.639***	-0.233
Ofsted 2018: Outstanding	0.062	(0.241)	0.126	(0.333)	0.064***	0.199
Ofsted 2018: Good	0.577	(0.494)	0.553	(0.499)	-0.024	-0.048
Ofsted 2018: Inadequate	0.055	(0.228)	0.082	(0.275)	0.027	0.111
Ofsted 2018: Requires improvement	0.242	(0.428)	0.214	(0.411)	-0.028	-0.069
Ofsted 2018: Missing	0.064	(0.244)	0.025	(0.157)	-0.038*	-0.191
School type: Academy-sponsor led	0.428	(0.495)	0.428	(0.496)	0.000	0.000
School type: Community school	0.084	(0.277)	0.101	(0.302)	0.017	0.056
School type: Voluntary aided/controlled school	0.072	(0.259)	0.050	(0.219)	-0.022	-0.100
School type: Foundation school	0.065	(0.247)	0.075	(0.265)	0.011	0.040
School type: Free school - Mainstream	0.325	(0.469)	0.296	(0.458)	-0.030	-0.063
School type: Others	0.026	(0.160)	0.050	(0.219)	0.024*	0.123
Urban	0.900	(0.300)	0.975	(0.157)	0.075***	0.372
Rural	0.100	(0.300)	0.025	(0.157)	-0.075***	-0.372
Region: East Midlands	0.097	(0.296)	0.082	(0.275)	-0.015	-0.062
Region: East of England	0.084	(0.277)	0.082	(0.275)	-0.002	-0.006
Region: London	0.189	(0.392)	0.252	(0.435)	0.062*	0.148
Region: North East	0.137	(0.344)	0.094	(0.293)	-0.043	-0.138
Region: North West	0.077	(0.267)	0.145	(0.353)	0.068***	0.198
Region: South East	0.156	(0.363)	0.057	(0.232)	-0.099***	-0.384
Region: South West	0.024	(0.152)	0.063	(0.244)	0.039***	0.181
Region: West Midlands	0.111	(0.314)	0.113	(0.318)	0.003	0.008
Region: Yorkshire & the Humber	0.126	(0.332)	0.113	(0.318)	-0.013	-0.040
TP schools (w/ at least 1 TP pupils)			0.302	(0.461)		0.83
Census school-level % FSM Spring 2021	0.260	(0.141)	0.397	(0.115)	0.137***	1.042
% EAL	0.137	(0.174)	0.220	(0.209)	0.083***	0.418
% SEN	0.217	(0.091)	0.218	(0.065)	0.001	0.017
% Female	0.498	(0.181)	0.493	(0.169)	-0.005	-0.036
Average AEA score	3.534	(1.664)	3.774	(1.739)	0.239*	0.145
Average IDACI scores	0.036	(0.020)	0.048	(0.014)	0.012***	0.731
% White British	0.122	(0.070)	0.099	(0.063)	-0.023***	-0.357
% Asian	0.016	(0.032)	0.023	(0.037)	0.006**	0.184
% Black	0.014	(0.028)	0.023	(0.031)	0.009***	0.33
% Other ethnic	0.025	(0.025)	0.031	(0.024)	0.006***	0.278
% Not white	0.004	(0.010)				
% Not white	0.059	(0.059)	0.077	(0.067)	0.018***	0.268
Observations	2,188		159		2,352	2,352

Source: Year 11 population data.

Note: the percentage of pupils with unknown ethnicity was suppressed because of low counts.

P-values: * <0.1; ** <0.05; *** <0.001.

AEA, Academic Excellence Areas; AM, Academic Mentoring; EAL, English as an Additional Language; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; KS, Key Stage; Ofsted, Office for Standards in Education; PP, Pupil Premium; SD, standard deviation; SEN, special educational needs; Std, ???; TP, Tuition Partners.

Figure A1: Common support

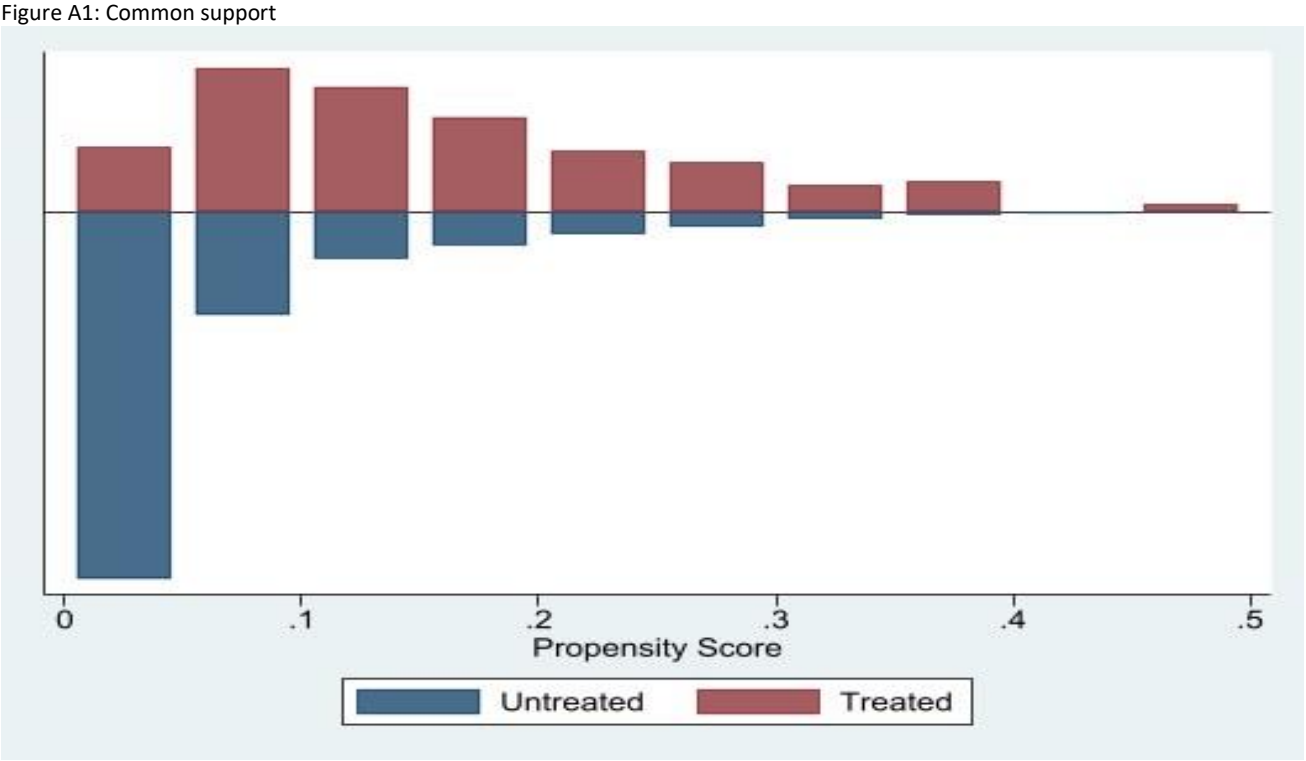


Table A3a: Summary statistics of previous years' GCSEs and previous years' and current year KS2, before and after matching, no caliper

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
KS4 English 2019/20	Unmatched	4.69	5.15	-0.46	0.06	-7.44
	ATT	4.69	4.67	0.02	0.06	0.37
KS4 Maths 2019/20	Unmatched	4.56	5.04	-0.48	0.07	-6.67
	ATT	4.56	4.51	0.05	0.07	0.73
KS4 English 2018/19	Unmatched	4.27	4.76	-0.49	0.07	-7.35
	ATT	4.27	4.27	0.00	0.07	-0.07
KS4 Maths 2018/19	Unmatched	4.19	4.72	-0.53	0.08	-6.99
	ATT	4.19	4.20	-0.01	0.07	-0.15
KS4 English 2017/18	Unmatched	4.29	4.76	-0.47	0.06	-7.23
	ATT	4.29	4.32	-0.03	0.06	-0.51
KS4 Maths 2017/18	Unmatched	4.19	4.72	-0.53	0.08	-7.01
	ATT	4.19	4.26	-0.08	0.08	-1.05
KS4 English 2016/17	Unmatched	4.32	4.81	-0.49	0.06	-7.76
	ATT	4.32	4.37	-0.05	0.06	-0.81
KS4 Maths 2016/17	Unmatched	4.20	4.74	-0.54	0.07	-7.38
	ATT	4.20	4.25	-0.05	0.07	-0.78
KS2 Reading 2020/21	Unmatched	100.84	102.80	-1.97	0.26	-7.68
	ATT	100.84	101.16	-0.33	0.22	-1.49
KS2 Maths 2020/21	Unmatched	102.05	103.19	-1.14	0.23	-4.99
	ATT	102.05	102.22	-0.17	0.19	-0.86
KS2 Reading 2019/20	Unmatched	29.51	31.48	-1.96	0.26	-7.48
	ATT	29.51	29.58	-0.07	0.24	-0.3
KS2 Maths 2019/20	Unmatched	68.61	71.85	-3.23	0.56	-5.75
	ATT	68.61	68.92	-0.31	0.52	-0.59
KS2 Reading 2018/19	Unmatched	29.53	31.40	-1.88	0.26	-7.3
	ATT	29.53	29.63	-0.10	0.23	-0.44
KS2 Maths 2018/19	Unmatched	68.48	72.01	-3.54	0.58	-6.1
	ATT	68.48	68.85	-0.38	0.54	-0.7
KS2 Reading 2017/18	Unmatched	30.97	32.98	-2.01	0.29	-6.94
	ATT	30.97	31.04	-0.07	0.28	-0.26
KS2 Maths 2017/18	Unmatched	68.21	71.20	-2.98	0.62	-4.83
	ATT	68.21	68.39	-0.18	0.55	-0.32
KS2 Reading 2016/17	Unmatched	29.66	31.82	-2.16	0.27	-8.01
	ATT	29.66	30.03	-0.37	0.25	-1.48
KS2 Maths 2016/17	Unmatched	67.24	71.05	-3.81	0.58	-6.62
	ATT	67.24	67.76	-0.52	0.53	-0.99

Treatment assignment	Off support	On support	Total
Untreated	0	2,056	2,056
Treated	0	159	159
Total	0	2,215	2,215

ATT, average treatment effect on the treated; KS, Key Stage; S.E., standard error; T-stat, T-statistics

Table A3b: Summary statistics of previous years' GCSEs and previous years' and current year KS2, before and after matching, caliper 0.05

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
KS4 English 2019/20	Unmatched	4.69	5.15	-0.46	0.06	-7.44
	ATT	4.69	4.67	0.02	0.06	0.37
KS4 Maths 2019/20	Unmatched	4.56	5.04	-0.48	0.07	-6.67
	ATT	4.56	4.51	0.05	0.07	0.73
KS4 English 2018/19	Unmatched	4.27	4.76	-0.49	0.07	-7.35
	ATT	4.27	4.27	0.00	0.07	-0.07
KS4 Maths 2018/19	Unmatched	4.19	4.72	-0.53	0.08	-6.99
	ATT	4.19	4.20	-0.01	0.07	-0.15
KS4 English 2017/18	Unmatched	4.29	4.76	-0.47	0.06	-7.23
	ATT	4.29	4.32	-0.03	0.06	-0.51
KS4 Maths 2017/18	Unmatched	4.19	4.72	-0.53	0.08	-7.01
	ATT	4.19	4.26	-0.08	0.08	-1.05
KS4 English 2016/17	Unmatched	4.32	4.81	-0.49	0.06	-7.76
	ATT	4.32	4.37	-0.05	0.06	-0.81
KS4 Maths 2016/17	Unmatched	4.20	4.74	-0.54	0.07	-7.38
	ATT	4.20	4.25	-0.05	0.07	-0.78
KS2 Reading 2020/21	Unmatched	100.84	102.80	-1.97	0.26	-7.68
	ATT	100.84	101.16	-0.33	0.22	-1.49
KS2 Maths 2020/21	Unmatched	102.05	103.19	-1.14	0.23	-4.99
	ATT	102.05	102.22	-0.17	0.19	-0.86
KS2 Reading 2019/20	Unmatched	29.51	31.48	-1.96	0.26	-7.48
	ATT	29.51	29.58	-0.07	0.24	-0.3
KS2 Maths 2019/20	Unmatched	68.61	71.85	-3.23	0.56	-5.75
	ATT	68.61	68.92	-0.31	0.52	-0.59
KS2 Reading 2018/19	Unmatched	29.53	31.40	-1.88	0.26	-7.3
	ATT	29.53	29.63	-0.10	0.23	-0.44
KS2 Maths 2018/19	Unmatched	68.48	72.01	-3.54	0.58	-6.1
	ATT	68.48	68.85	-0.38	0.54	-0.7
KS2 Reading 2017/18	Unmatched	30.97	32.98	-2.01	0.29	-6.94
	ATT	30.97	31.04	-0.07	0.28	-0.26
KS2 Maths 2017/18	Unmatched	68.21	71.20	-2.98	0.62	-4.83
	ATT	68.21	68.39	-0.18	0.55	-0.32
KS2 Reading 2016/17	Unmatched	29.66	31.82	-2.16	0.27	-8.01
	ATT	29.66	30.03	-0.37	0.25	-1.48
KS2 Maths 2016/17	Unmatched	67.24	71.05	-3.81	0.58	-6.62
	ATT	67.24	67.76	-0.52	0.53	-0.99

Treatment assignment	Off support	On support	Total
Untreated	0	2,056	2,056
Treated	0	159	159
Total	0	2,215	2,215

ATT, average treatment effect on the treated; KS, Key Stage; S.E., standard error; T-stat, T-statistics

Table A3c: Summary statistics of previous years' GCSEs and previous years' and current year KS2, before and after matching, caliper 0.01

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
KS4 English 2019/20	Unmatched	4.69	5.15	-0.46	0.06	-7.44
	ATT	4.71	4.68	0.03	0.06	0.54
KS4 Maths 2019/20	Unmatched	4.56	5.04	-0.48	0.07	-6.67
	ATT	4.57	4.52	0.05	0.07	0.77
KS4 English 2018/19	Unmatched	4.27	4.76	-0.49	0.07	-7.35
	ATT	4.29	4.29	0.00	0.07	0.03
KS4 Maths 2018/19	Unmatched	4.19	4.72	-0.53	0.08	-6.99
	ATT	4.20	4.21	-0.01	0.07	-0.14
KS4 English 2017/18	Unmatched	4.29	4.76	-0.47	0.06	-7.23
	ATT	4.30	4.34	-0.04	0.06	-0.58
KS4 Maths 2017/18	Unmatched	4.19	4.72	-0.53	0.08	-7.01
	ATT	4.19	4.28	-0.09	0.08	-1.15
KS4 English 2016/17	Unmatched	4.32	4.81	-0.49	0.06	-7.76
	ATT	4.33	4.38	-0.05	0.06	-0.79
KS4 Maths 2016/17	Unmatched	4.20	4.74	-0.54	0.07	-7.38
	ATT	4.21	4.27	-0.06	0.07	-0.84
KS2 Reading 2020/21	Unmatched	100.84	102.80	-1.97	0.26	-7.68
	ATT	100.83	101.16	-0.34	0.22	-1.5
KS2 Maths 2020/21	Unmatched	102.05	103.19	-1.14	0.23	-4.99
	ATT	102.04	102.20	-0.16	0.20	-0.8
KS2 Reading 2019/20	Unmatched	29.51	31.48	-1.96	0.26	-7.48
	ATT	29.53	29.62	-0.09	0.24	-0.37
KS2 Maths 2019/20	Unmatched	68.61	71.85	-3.23	0.56	-5.75
	ATT	68.64	68.93	-0.29	0.53	-0.55
KS2 Reading 2018/19	Unmatched	29.53	31.40	-1.88	0.26	-7.3
	ATT	29.56	29.66	-0.10	0.24	-0.43
KS2 Maths 2018/19	Unmatched	68.48	72.01	-3.54	0.58	-6.1
	ATT	68.53	68.86	-0.32	0.55	-0.59
KS2 Reading 2017/18	Unmatched	30.97	32.98	-2.01	0.29	-6.94
	ATT	31.01	31.07	-0.06	0.29	-0.22
KS2 Maths 2017/18	Unmatched	68.21	71.20	-2.98	0.62	-4.83
	ATT	68.27	68.40	-0.13	0.56	-0.23
KS2 Reading 2016/17	Unmatched	29.66	31.82	-2.16	0.27	-8.01
	ATT	29.69	30.09	-0.39	0.26	-1.53
KS2 Maths 2016/17	Unmatched	67.24	71.05	-3.81	0.58	-6.62
	ATT	67.33	67.78	-0.45	0.53	-0.84

Treatment assignment	Off support	On support	Total
Untreated	0	2056.00	2056.00
Treated	4	155.00	159.00
Total	4	2,211	2,215

ATT, average treatment effect on the treated; KS, Key Stage; S.E., standard error; T-stat, T-statistics

Appendix B

Variables used for matching and included as controls in all regressions.

- Income Deprivation Affecting Children Index (IDACI);
- Achieving Excellence Areas (AEA) score;
- interaction between IDACI and AEA scores;
- Key Stage (KS)1 to KS2 value added attainment, at district level;
- management/school type secondary—Community, Academies, Foundation, Free schools, Sponsored Academies, Voluntary school, Studio schools, and University Technical College;
- region (London, Government Office Region, and regional dummies);
- Free School Meals (FSM)—percentage eligible in previous year; and
- KS4 Maths and English in 2019/20.

Pupil-level controls used in all regressions:

- KS2 maths and English scores;
- female;
- FSM ever in 6 years;
- ethnicity (White British, Asian, Black, Unknown, Other);
- English as an additional language (EAL) and EAL unknown;
- special educational needs (SEN); and
- looked after for 12 months, looked after since 31 March, looked after for 6 months.

School-level controls used in all regressions:

- Ofsted (Office for Standards in Education) ratings and missing Ofsted rating;
- school FSM percentage above median and missing FSM percentage;
- urban/rural area;
- IDACI quintiles and IDACI missing; and
- Tuition Partners participation.

Appendix C: Year 11 checks

Year 11 Teacher Assessed Grades (TAGs) considerations and checks

The checks presented below informed the approach to the analysis and our interpretation of the results. All of the checks were outlined in the study plan, which was published prior to accessing the TAGs data. It is important to note that it was known in advance that the proposed checks would not be able to detect the presence of systematic bias with certainty (i.e., failure to detect systematic bias, does not mean that there is no systematic bias) therefore, the findings will need to be treated with caution.

Ex ante (before analysis) tests:

i) To address consideration 1: that TAGs may be distributed differently compared to previous years (in particular there may be differences around the grade ¾ boundary), we compared the distributions of GCSEs awarded in the years before AM (2018, 2019, and 2020) and the TAGs awarded in 2021 for all pupils and for PP-eligible pupils (as a group in itself) across all schools. If the distribution of grades across the years is significantly different for both groups of pupils, this is a potential concern we will account for in the interpretation of results. If there is an AM effect, we might expect a change too.

ii) To address consideration 2: that schools selected the pupils who undertake AM and that, as a result, teachers may have applied some conscious or unconscious bias in their assessment of these pupils, we used across-subjects (maths and English) variation to help identify if any bias was subject-specific and not pupil-specific in the sample of pupils tutored. As long as any bias is a teacher bias and each teacher teaches a different subject, the cross-subject comparison should reveal the presence of bias: if the bias was across-subjects, then the cross-subject comparison could not reveal any systematic teacher bias. However, bias across-subjects may reveal the presence of pupil-specific bias (negative selection of pupils in the AM programme) or spillover effects across-subjects. Evidence of teacher bias at pupil level would represent a serious concern to the validity of the analysis, as it may point to systematic bias in AM schools versus non-AM schools. Note that: i) it would be difficult to disentangle the impact of AM from the effect of bias as they may both go in the same direction and they both affect the same population; and ii) we could not observe the counterfactual, how these pupils would have performed in the absence of AM. To explore this, we regressed English and maths TAGs separately on English and maths AM hours received. We would expect AM English hours to be correlated with English TAGs and AM maths hours to be correlated with maths TAGs.

Ex post (after analysis) tests:

i) To address considerations 3 and 4: uncertainties around whether the TAGs reflect pupils' performance after the tutoring and whether the assessments are sufficiently sensitive to change, we planned to see whether higher dosage (amount) of tutoring is associated with higher grades on the sample of AM schools only. The assumption is that time of enrolling is exogenous to performance. However, we point out that schools that enrol earlier may be more enthusiastic about the programme and have higher dosages. We expected to see a bigger effect among those with larger dosage. This can be due to: i) dosage matters in improving ability; and ii) larger dosages reduce the dilution in TAGs if they reflect the performance over the entire academic year. To explore this, we regressed English and maths TAGs on dosage of tutoring, and controlling for pupil-level and school-level characteristics.

ii) To address consideration 1: that TAGs may be distributed differently compared to previous years (concern 1), we also tested if the distribution of tests across the years (i.e., exam year 2021 vs. 2018/2019/2020) was different across AM compared to non-AM schools for all pupils and PP-eligible pupils. Evidence of a significant difference in distribution across AM and non-AM schools may suggest the presence of measurement errors. If AM and non-AM schools both allocated TAGs equally, then we would expect AM schools to have slightly more higher grades because of AM (if AM is effective). If AM and non-AM schools allocated TAGs differently, then we would need to investigate whether the difference is related to bias (see checks (ii) in the 'Ex ante tests' section and (ii) and (iii) in the 'Ex post tests' section) or is an indication of a positive (or indeed negative) effect of AM. Note, it is difficult to disentangle the impact of AM from a potentially systematic difference in how TAGs are awarded in AM compared to non-AM schools as they may both go in the same direction and they both affect the same population.

iii) To address the consideration that TAGs may be distributed differently compared to previous years (concern 1): we planned to perform the analysis on Year 11 pupils predicted to do AM. To do this, we

first estimated a pupil-level logistic model for AM participation and checked that the model had good predictive power. If so, we proceeded with predicting pupils participating in AM and pupils not participating in AM. If the effect of AM on pupils predicted to participate is positive, this can be due to the positive effect of AM or to different allocation of TAGs compared to examinations. We also planned to perform the analysis on Year 11 children predicted NOT to participate in AM. If there was an impact also on children not predicted to participate in AM, then it could be interpreted as evidence of TAGs being allocated differently to examinations. However, we caveat for the fact that this could also be due to the presence of spillovers or because of non-random selection of schools into treatment that are not fully controlled for in the methodological approach. If AM is effective, predicted AM should always have higher TAGs than predicted non-AM even if TAGs are allocated differently compared to examinations. The reliability of this test depends on how well we can predict participation to AM. The test cannot disentangle the increase in grade due to AM from a systematic teacher bias towards AM pupils only.

While we already knew (at the time of writing the study plan) that TAGs were likely to be distributed differently compared to previous year (concern 1), it is not indicative of systematic bias between AM and non-AM schools.19 The risk that the assessments are not sensitive enough to change, informed the interpretation of the results in case of no significant effect found.

We are more concerned about the validity of these measures as outcomes if some of the checks outlined above addressing considerations 2, 3, and 4 pointed towards the presence of systematic bias between AM and non-AM schools (specifically: ex ante test (ii); and ex post tests (ii); and (iii), see above). We highlight the fact that there could be more than one interpretation to these checks, that may not allow us to detect bias: i) the possibility of between-subject spillovers; and ii) the heterogeneous effects of the pandemic itself across pupils and subjects. Results are shown below.

Ex ante (before analysis) 1: distribution of teacher assessed GCSE over time

The first ex ante check shows if the distribution of teacher assessed GCSE grades in 2021 is significantly different with respect to the distribution of GCSE grades in the three previous academic years (2017/18, 2018/19, and 2019/20), in both maths and English and across all pupils and PP-eligible pupils only. This check was conducted on the whole population of secondary schools (see 'Ex post 2' for a complementary check taking into account AM status of the school). The Kolmogorov–Smirnov test measures for equality of distribution. The null hypothesis is that two dataset values are from the same continuous distribution. It tests the hypothesis that one value of the distribution for group 1 contains smaller or larger values than for group 2 and combines these differences. Table C1 shows that the null is rejected in all comparisons of 2021 grades with previous years, for maths and English, and for all pupils and for PP-eligible pupils only.

The results of the t-test indicate in Table C2 that, on average, grades are significantly lower in previous years, although the difference with respect to grades in 2019/20 is much smaller than the difference with respect to grades awarded in 2017/18 and 2018/19. In 2019/20, grades were allocated on the basis of schools' best judgement regarding what grade they believed candidates would have achieved if exams had gone ahead. These were referred to as Centre Assessment Grades.

The box plots show the distribution of grades across the years for both maths and English, for all pupils (Figure C1) and PP-eligible pupils only (Figure C2).20 The plots indicate that the distribution of grades was different, and higher, in 2019/20 and 2020/21 compared to 2017/18 and 2018/19, especially for all pupils. This is consistent with data published by Ofqual (2021), which summarised that 'Overall [2021] GCSE results are higher at grade 7 and above compared to 2020 (28.5% in 2021 compared with 25.9% in 2020, and 20.7% in 2019) and relatively stable at grade 4 and above compared to 2020 (76.9% in 2021 compared with 75.9% in 2020, and 67.1% in 2019).'

The results from these tests show evidence of concern 1 that TAGs are distributed differently compared to previous years. This is consistent with what has been widely reported and it is not a serious concern for our analysis as long as AM and non-AM schools do not allocate TAGs differently (we perform this check in ex post 2).21

Table C1: Kolmogorov–Smirnov tests for equality of distribution

[19]Ofqual has published the following note: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1010126/6828-3_Student-level_equalities_analysis_for_GCSE_and_A_level_summer_2021.pdf Among others, it documents an increased gap between FSM candidates relative to prior-attainment-matched non-FSM pupils.

[20] The dots in these figures indicate lower number of observations in correspondence of a given grade. In all cases where the count is lower than 10 it is not reported.

[21] For example, the Education Policy Institute https://epi.org.uk/wp-content/uploads/2021/08/GCSE_analysis_2021_EPI_.pdf.

	Math, all pupils	English, all pupils	Maths, PP-eligible pupils	English, PP-eligible pupils
Combined K-S, 2019/20-2020/21:	0.0152	0.0254	0.0166	0.018
p-value	0.000	0.000	0.000	0.000
Combined K-S, 2018/19-2020/21:	0.0836	0.0952	0.0798	0.117
p-value	0.000	0.000	0.000	0.000
Combined K-S, 2017/18-2020/11:	0.0775	0.0946	0.0782	0.119
p-value	0.000	0.000	0.000	0.000

Source: Year 11 population data, 500K observations per year.
Rows 1, 3, and 5 compare distribution of grades in two academic years. So, 2020 refers to academic year 2019/20 and 2021 to academic year 2020/21. Numbers in columns 2–3 use the sample of all pupils; columns 4 and 5 use the sample of Free School Meals pupils only.
K-S, Kolmogorov–Smirnov; PP, Pupil Premium.

Table C2: Ex ante 1: Estimates of KS4 maths and English grades on year dummies, reference: 2020/21

	KS4 Maths			KS4 English			KS4 Maths – FSM			KS4 English – FSM		
	Coeff	S.E.	p-value	Coeff	S.E.	p-value	Coeff	S.E.	p-value	Coeff	S.E.	p-value
Academic year 2017/18	-0.407***	0.012	0.000	-0.360***	0.012	0.000	-0.335***	0.015	0.000	-0.326***	0.016	0.000
Academic year 2018/19	-0.434***	0.010	0.000	-0.358***	0.010	0.000	-0.367***	0.013	0.000	-0.320***	0.013	0.000
Academic year 2019/20	-0.433***	0.008	0.000	-0.350***	0.008	0.000	-0.371***	0.013	0.000	-0.336***	0.012	0.000
Constant	5.063***	0.018	0.000	4.888***	0.019	0.000	4.435***	0.018	0.000	4.165***	0.019	0.000
N	6806			6806			6806			6806		
R-squared	0.082			0.050			0.054			0.040		

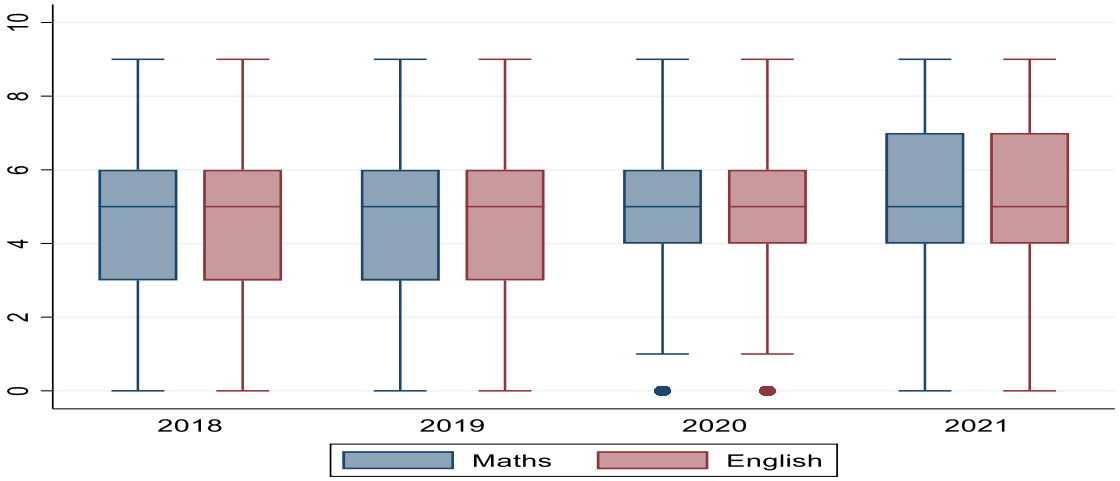
Source: Year 11 population data.

Results in columns 2–7 use the sample of all pupils; columns 8 and 13 use the sample of FSM pupils only. Data collapsed at school and year level. Residuals clustered at school level.

P-values: * <0.1; ** <0.05; *** <0.001.

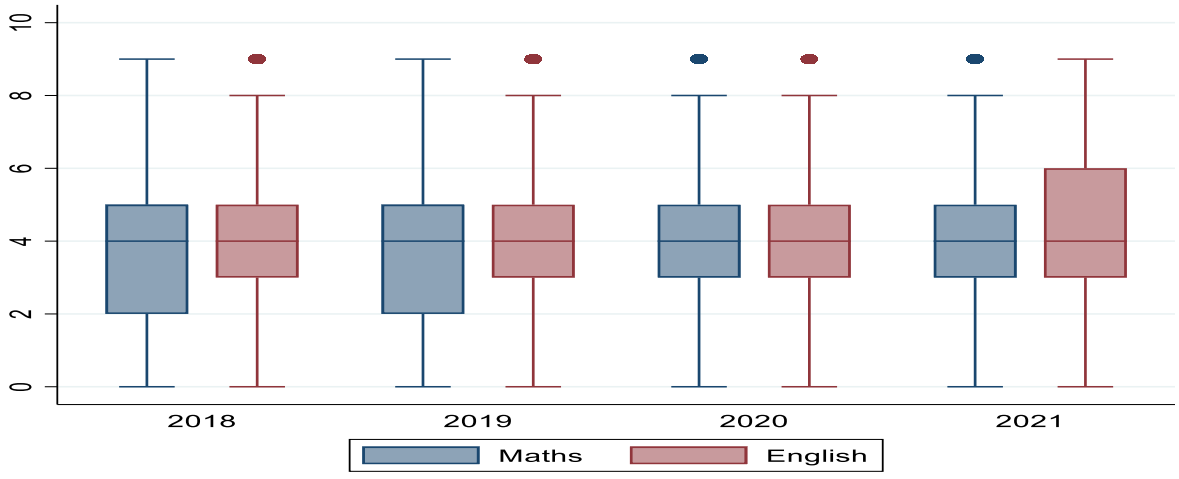
Coef, coefficient; KS, Key Stage; FSM, Free School Meals; N, number; S.E., standard error.

Figure C1: Distributions of grades across the years (i.e., 2021, 2020, 2019, and 2018) for all pupils



Note: years in X-axis refer to the end of the academic year. So, 2018 refers to academic year 2017/18. Y-axis = GCSE point score.

Figure C2: Distributions of grades across the years (i.e., 2021, 2020, 2019, and 2018) for PP-eligible pupils



Note: years in X-axis refer to end of the academic year. So, 2018 refers to academic year 2017/18. Y-axis = GCSE point score.

Ex ante (before analysis) 2: across-subject variation in grades

The second ex ante check investigates across-subjects variation in AM pupils' grades to verify the presence of subject-specific or pupil-specific bias. To do this, we selected the sample of AM pupils in Year 11 and regressed teacher assessed GCSE grades in 2020/21 on dummy variables for pupil-level AM delivery in English and AM delivery in maths, controlling for individual- and pupil-level characteristics, regional dummies, and Income Deprivation Affecting Children Index (IDACI) rank.

If the bias is subject-specific, the coefficient associated with the mentored subject should be significant only when regressed on the same subject grade (i.e., AM maths should be significant on maths grades, not on English grades and vice versa). If the bias is pupil-specific, the coefficients associated with both mentored subjects would be large and significant.

The regression results in Tables C3 and C4 indicate that being tutored in maths and being tutored in English are both associated with a lower English grade and maths grades, for both PP-eligible pupils and all pupils, although the coefficient of being mentored in maths is smaller than the coefficient for English mentoring. Being mentored in maths is not associated with significantly different grades.

The negative sign of the AM subject coefficient of English mentoring could be explained by lower performing pupils being selected for the AM programme in English or it could represent negative bias in the awarding of grades with the knowledge that they were selected for AM. It should also be noted that here we are not comparing AM pupils to similar pupils in non-AM schools. The lack of counterfactual suggests that, if the negative coefficient is driven by negative selection, AM English pupils may have performed worse in the absence of AM anyway. While the results of this check represent a concern on the validity of TAGs as outcome measures, it would be necessary to compare the outcomes of

AM pupils with a comparison group of similar pupils in non-AM schools to further assess whether the negative coefficient is representing negative bias in the awarding of grades or also negative selection into the programme and/or subject. The difficulty of doing so relies on the pupil-level selection mechanism, which we cannot control for. We planned to focus on PP-eligible pupils only to account for pupil-level selection in both AM and non-AM schools, but as the average percentage of Pupil Premium doing AM is 13%, we cannot properly identify AM pupils in comparison schools (the model aiming to predict participation in AM is reported below in ex post 3).

Table C3: AM impact across-subjects on AM pupils, all pupils in analysed sample

	KS4 Maths			KS4 Maths			KS4 Eng			KS4 Eng		
	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values
AM subject Maths	0.032	0.135	0.812	-0.041	0.134	0.760				0.061	0.135	0.652
AM subject English				-0.307*	0.125	0.015	-0.453***	0.126	0.000	-0.436**	0.136	0.002
N	2693			2693			2680			2680		
R-squared	0.520			0.523			0.437			0.437		

Source: Year 11 population data, n=318 schools.

Note: Control for: gender, ethnicity dummies, EAL, FSM, KS2, looked after, SEN; school-level FSM, Ofsted dummies, IDACI, school types, regions dummies. School-level clustered residuals.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; EAL, English as an Additional Language; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; KS, Key Stage; N, number; Ofsted, Office for Standards in Education; S.E., standard error; SEN, special educational needs.

Table C4: AM impact across-subjects on AM pupils, PP-eligible pupils in analysed sample

	KS4 Maths			KS4 Maths			KS4 Eng			KS4 Eng		
	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values
AM subject Maths	0.034	0.155	0.825	-0.057	0.155	0.714				-0.061	0.148	0.681
AM subject English				-0.332*	0.140	0.019	-0.451**	0.137	0.001	-0.469**	0.147	0.002
N	1243			1243			1236			1236		
R-squared	0.483			0.488			0.408			0.408		

Source: Year 11 population data, n=318 schools.

Note: Control for: gender, ethnicity dummies, EAL, FSM, KS2, looked after, SEN; school-level FSM, Ofsted dummies, IDACI, school types, regions dummies. School-level clustered residuals.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; EAL, English as an Additional Language; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; KS, Key Stage; N, number; Ofsted, Office for Standards in Education; PP, Pupil Premium; S.E., standard error; SEN, special educational needs.

Ex post (after analysis) 1: dosage of mentoring

The first ex post check assesses whether the TAGs are sensitive enough to change, in the event that TAGs do reflect pupils' performance. We exploited dosage of mentoring on the sample of AM schools and regressed TAGs on a variable indicating the numbers of AM sessions completed by the AM pupils. For English TAGs the dependent variable is the total number of mentoring sessions taken in English and for the maths TAGs it is the total number of mentoring sessions taken in maths. The data only indicate the number of sessions taken, but not the length of each session, so the dosage is measured imprecisely.

Figures C3 and C4 report the distribution of pupil-level maths and English dosage. Among the AM Year 11 pupils, the average number of sessions was 7.7 for maths and 6.7 for English. More than half of pupils received less than five sessions in maths, and less than four sessions in English, with the average (eight and seven sessions, respectively) being driven by a minority of outliers who received many more sessions. We removed from the histograms pupils who had more than 14 sessions in maths more than 11 sessions in English as less than 10 pupils were associated with each of these categories.

The regression was run on the sample of AM pupils in Year 11 (all pupils and PP-eligible pupils) controlling for individual and pupil-level characteristics, regional dummy variables, and IDACI rank. Results are presented in Table C5. We find that a higher number of mentoring sessions received was not correlated with achieving better English and maths TAGs. The test suggests the TAGs may not have significant sensitivity to capture change induced by the intervention (also see the results in subsection 'Weighting/regression RQ1' in the 'Results' section). Alternative explanations include AM not having an impact or pupils who received more sessions had poorer attainment, hence the AM impact is masked.

Figure C3: Distribution of pupil-level dosage in terms of number of sessions in maths

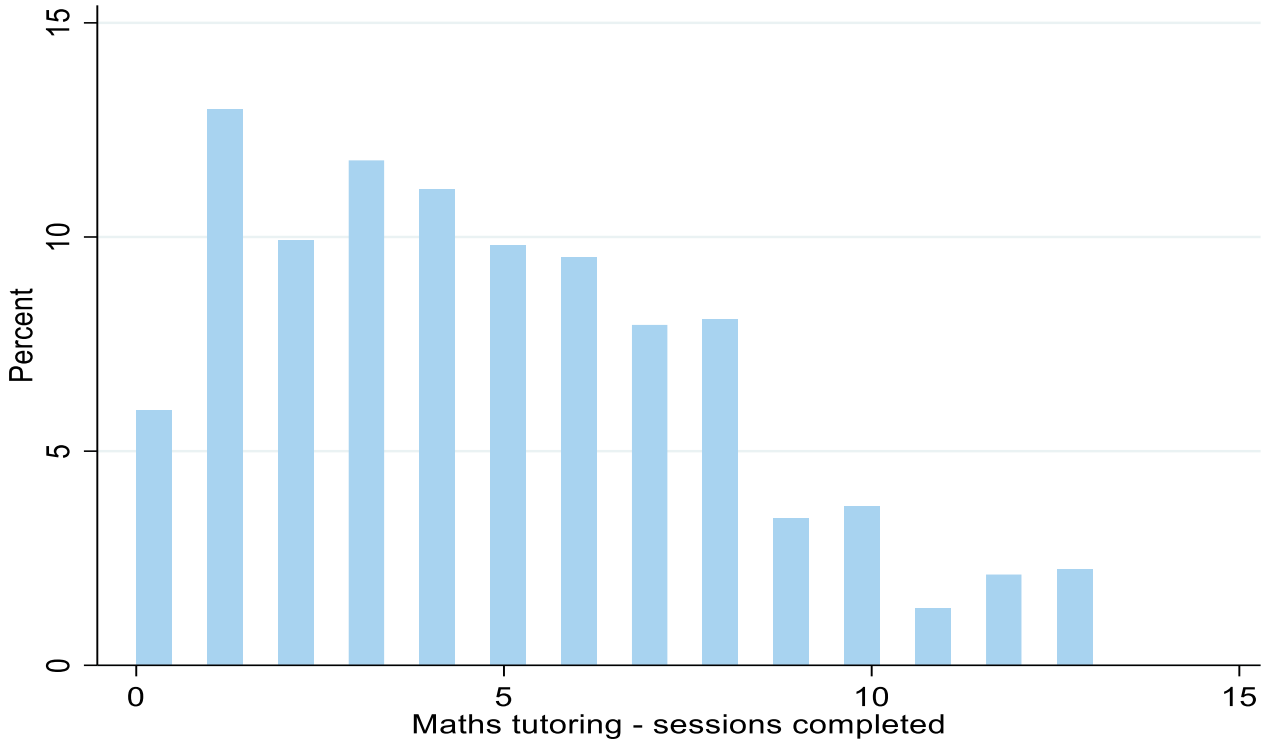


Figure C4: Distribution of pupil-level dosage in terms of number of sessions in English

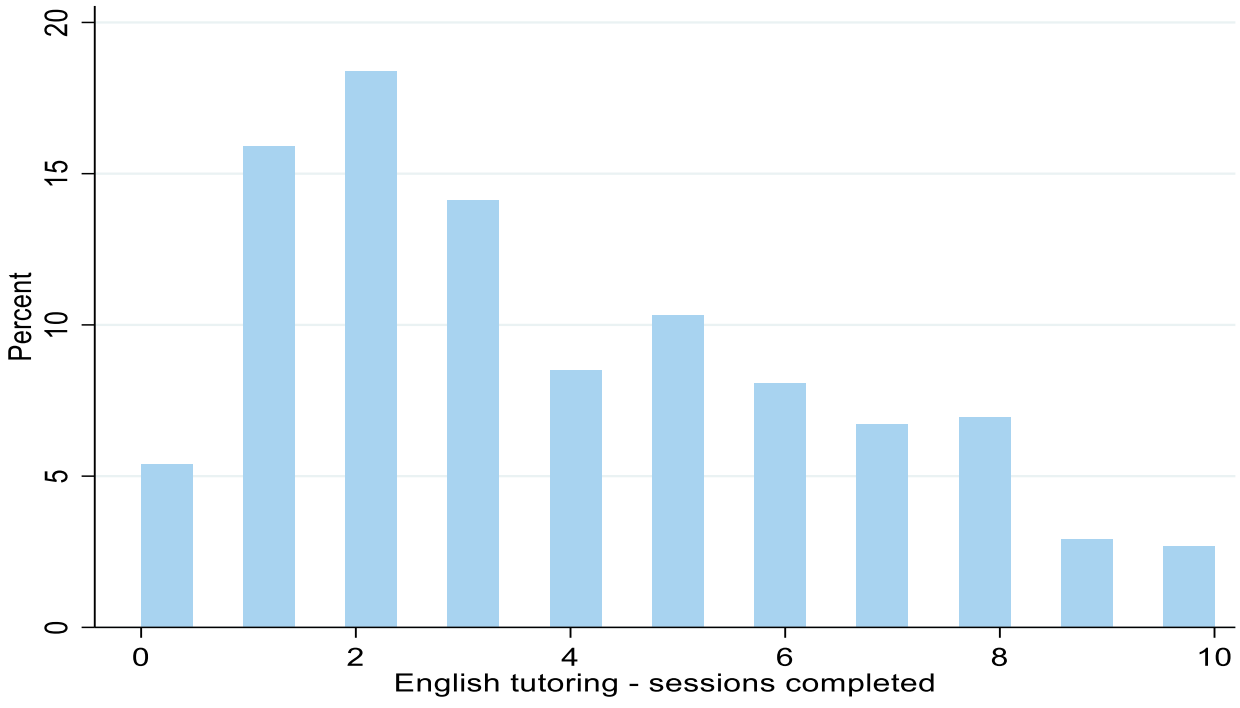


Table C5: Dosage of mentoring

	KS4 Maths						KS4 English					
	All pupils			PP-eligible pupils			All pupils			PP-eligible pupils		
	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values
Total Maths sessions completed	0.010	0.008	0.233	0.019	0.011	0.107						
Total English sessions completed							-0.008	0.011	0.460	-0.015	0.012	0.218
N	842			376			514			256		
R-squared	0.578			0.577			0.342			0.371		

Note: Sample: population of Year 11 AM schools excluding pupils with zero AM sessions (n=159 schools). Results in columns 2–4 and 8–10 use the sample of all pupils; columns 5–7 and 11–13 use the sample of FSM pupils only. Residuals clustered at school-level.

AM, Academic Mentoring; Coef, coefficient; KS, Key Stage; N, number; PP, Pupil Premium; S.E., standard error.

Ex post (after analysis) 2: distribution of TAGs between AM and matched non-AM schools

The second ex post check complements the first ex ante check, by testing if the distribution of tests across the years (i.e., 2021 vs. 2018/2019/2020) is different across AM and matched non-AM schools for all pupils and PP-eligible pupils. This uses the intervention group and matched comparison group in the analysed sample.

When comparing grades across the two samples, the Kolmogorov–Smirnov test for equality of distribution (Table C6) does not reject the null hypothesis of equality of distribution in almost all comparisons of 2021 grades with previous years, for both maths and English, and for all pupils and PP-eligible pupils only. Only the comparison between AM and non-AM schools in 2020/21 and 2019/20 for maths and all pupils is significantly different, at the 7% level.

The histograms in Figures C5 (maths, all pupils), C6 (English, all pupils), C7 (maths, PP-eligible pupils), and C8 (English, PP-eligible pupils) show the distribution of the difference between KS4 grades and teacher assessed GCSEs grades in 2020/21 between AM and non-AM schools for all pupils and PP-eligible pupils. The distribution of differences in maths did not appear to be markedly different across AM and matched non-AM schools over the years. For English, for both PP-eligible pupils and all pupils, it appears the distribution of differences in grades was more on the left, hence closer to zero, for AM schools than for non-AM schools, but not so for the comparison 2019/20 with 2020/21. This indicates that there was less difference in KS4 grades across the last 2 years in AM schools with respect to non-AM schools, which was consistent with the results reported in Table 15 showing higher scores over the last 2 years for AM schools.

Table C6: Kolmogorov–Smirnov tests for equality of distribution, in analysed sample

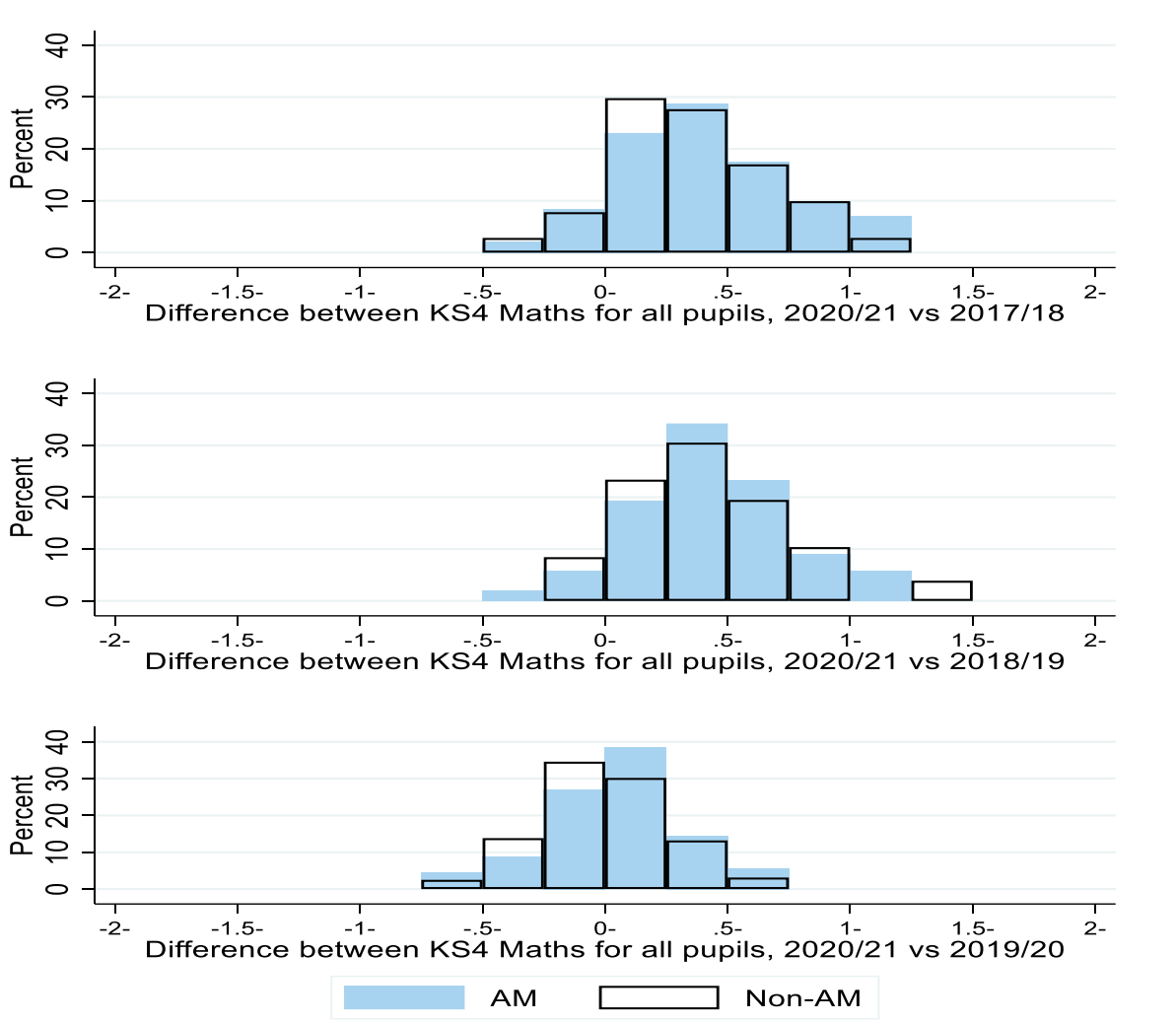
	Maths, all pupils	English, all pupils	Maths, PP-eligible pupils	English, PP-eligible pupils
Combined K-S, 2019/20–2020/21:	0.1447	0.0629	0.0818	0.0629
p-value	0.072	0.912	0.663	0.912
Combined K-S, 2018/19–2020/21:	0.0866	0.0717	0.0996	0.0912
p-value	0.608	0.822	0.428	0.541
Combined K-S, 2017/18–2020/21:	0.1263	0.0978	0.0769	0.0777
p-value	0.204	0.501	0.796	0.785
N	318	318	318	318

Source: Year 11 population data, n=318 schools.

Rows 1, 3, and 5 compare distribution of grades in two academic years. Numbers in columns 1–2 use the sample of all pupils; columns 3 and 4 use the sample of FSM pupils only.

FSM, Free School Meals; K-S, Kolmogorov–Smirnov; n, number; PP, Pupil Premium.

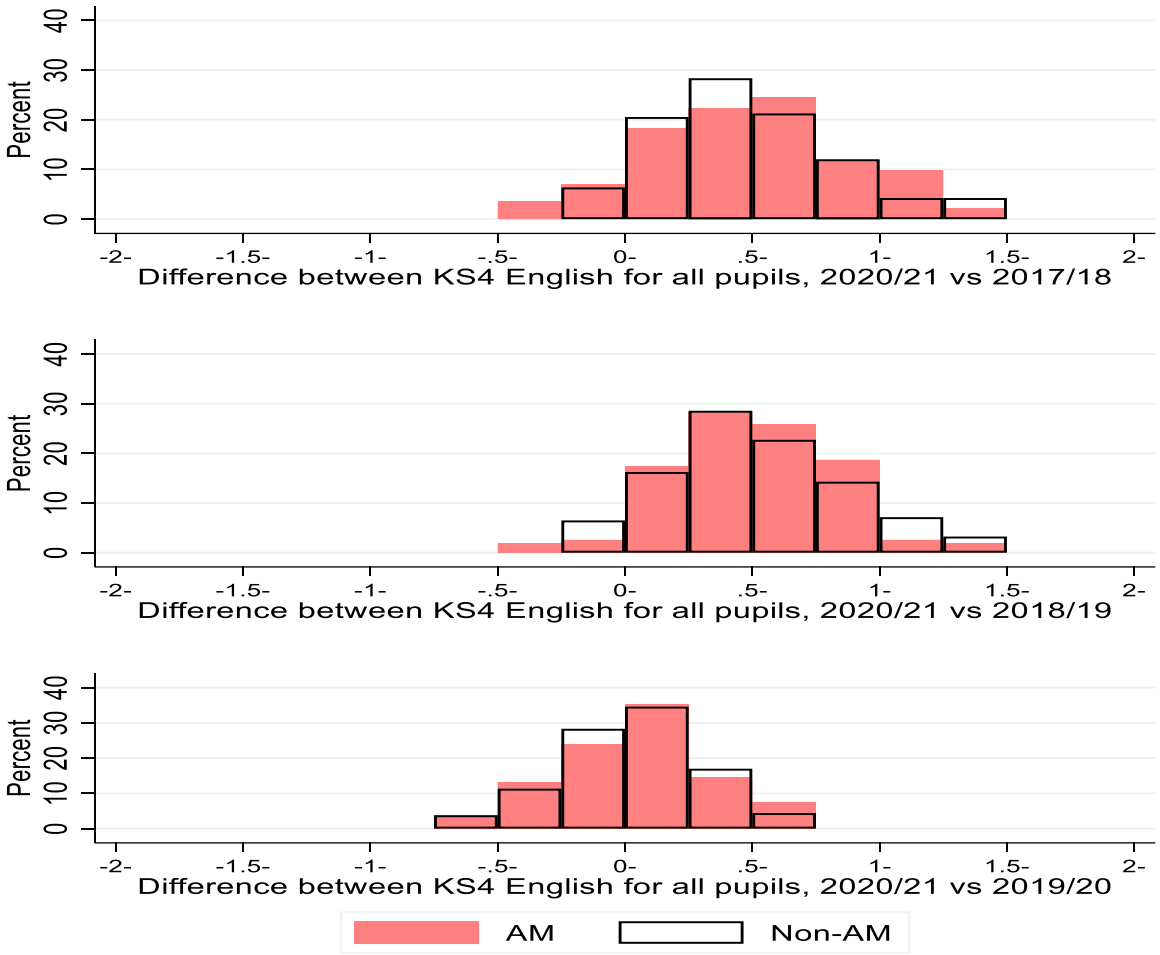
Figure C5: Distribution of difference in KS4 maths scores between AM and matched non-AM schools across the academic years, all pupils



Note: The symbol "–" next to the X-axis numbers indicates that the bands were re-calculated to exclude counts with less than three schools, in adherence with the SRS disclosure rule.

AM, Academic Mentoring; KS, Key Stage; SRS, Secure Research Service.

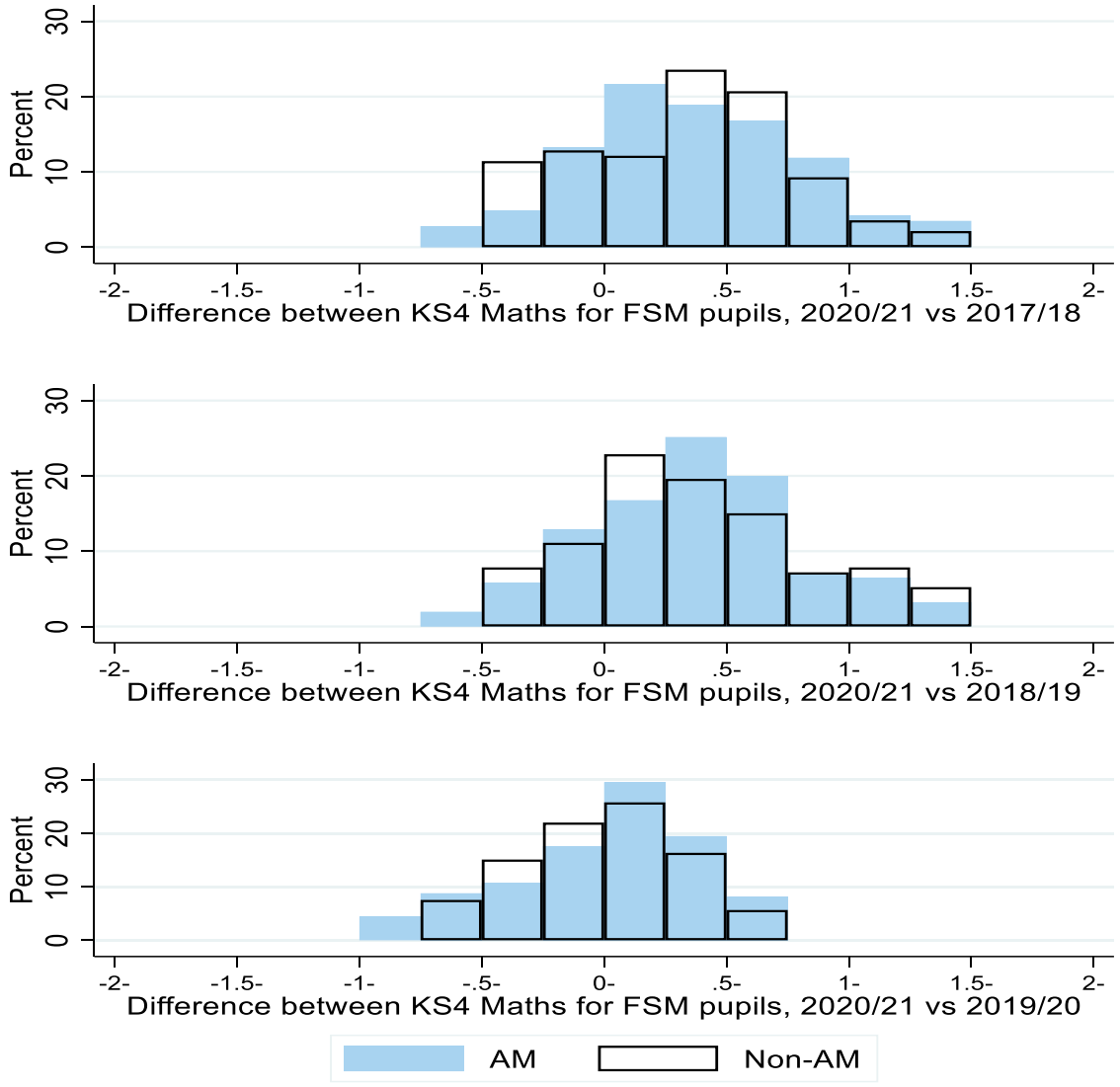
Figure C6: Distribution of difference in KS4 English scores between AM and matched non-AM schools across the academic years, all pupils



Note: The symbol "–" next to the X-axis numbers indicates that the bands were re-calculated to exclude counts with less than three schools, in adherence with the SRS disclosure rule.

AM, Academic Mentoring; KS, Key Stage; SRS, Secure Research Service.

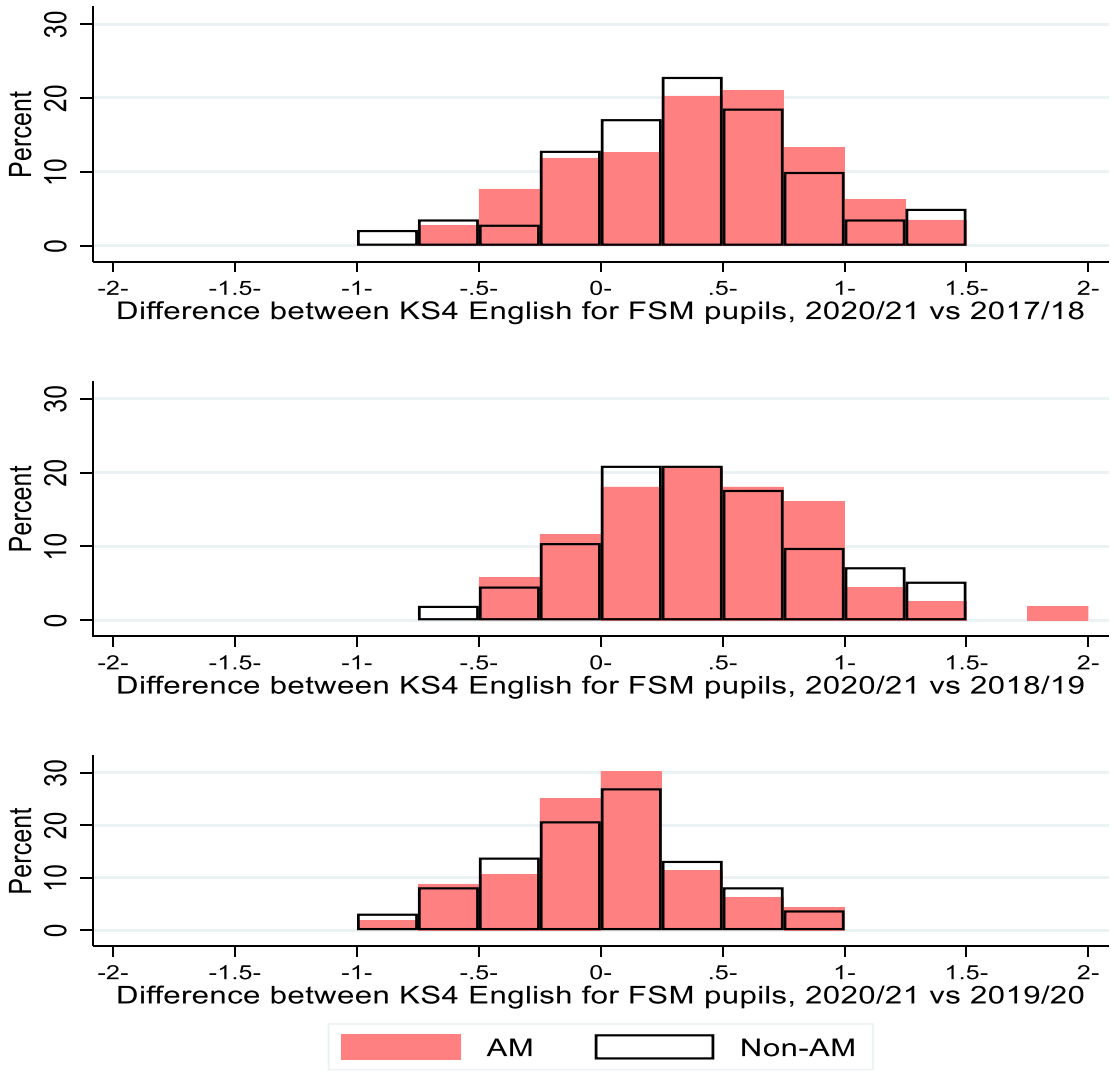
Figure C7: Distribution of difference in KS4 math scores between AM and matched non-AM schools across the academic years, PP-eligible pupils



Note: The symbol "–" next to the X-axis numbers indicates that the bands were re-calculated to exclude counts with less than three schools, in adherence with the SRS disclosure rule.

AM, Academic Mentoring; KS, Key Stage; PP, Pupil Premium; SRS, Secure Research Service.

Figure C8: Distribution of difference in KS4 English scores between AM and matched non-AM schools across the academic years, PP-eligible pupils



Note: The symbol "–" next to the X-axis numbers indicates that the bands were re-calculated to exclude counts with less than three schools, in adherence with the SRS disclosure rule.

AM, Academic Mentoring; KS, Key Stage; PP, Pupil Premium; SRS, Secure Research Service.

Ex post (after analysis) 3: analysis on pupils predicted not to participate

The third ex post check on Year 11 was undertaken to test the presence or absence of the impact of AM on pupils predicted not to participate in AM. To do this, we modelled the probability of pupil participation in AM schools using various markers of disadvantage recorded in the National Pupil Database (NPD) (socio-economic status measured by Free School Meals (FSM)/Pupil Premium, special educational needs, interaction with social service, prior attainment, English as an additional language, and ethnicity), with the plan to use this model to predict participation in both AM and non-AM schools (Table C7). However, our ability to predict AM participation with the observable characteristics available in the NPD was poor. When testing the predictive power of the model, by comparing predicted AM participation with actual AM participation, 87.7% of those predicted to participate in AM did not actually participate in AM and 10.1% of those predicted not to participate in AM actually participated in AM (see Table C8). The predictive power was similar across different thresholds of predicted participation (i.e., predicted participation higher than 0.25, 0.5, and 0.75). It was likely that selection of pupils for mentoring by schools was based on unobservable characteristics, such as their ability to catch up and make good use of mentoring. Hence, we did not proceed with this part of the analysis.

Table C7: Estimation of AM participation using Logit and Linear Probability Model (LPM)

	Logit			LPM		
	Coef	S.E.	P-value	Coef	S.E.	P-value
Female	0.145*	0.059	0.014	0.013*	0.005	0.016
FSM ever in 6 years	0.379***	0.057	0.000	0.036***	0.006	0.000
master only (1)	0.193	0.225	0.391	0.018	0.022	0.402
EAL (no as base):						
EAL: Yes	-0.095	0.077	0.215	-0.009	0.007	0.205
EAL: Missing	0.162	0.417	0.697	0.014	0.039	0.713
Ethnicity (unknown as base):						
White British	0.118	0.209	0.573	0.011	0.019	0.540
Asian	0.271	0.205	0.186	0.025	0.018	0.180
Black	0.186	0.200	0.353	0.017	0.018	0.345
Other ethnicities	0.244	0.194	0.208	0.023	0.017	0.195
SEN (no as base):						
SEN: Yes	0.120	0.094	0.202	0.011	0.009	0.234
Looked after since 31 March	0.439	0.510	0.389	0.057	0.078	0.465
Looked after for 6 months	0.215	0.735	0.770	0.027	0.115	0.813
Looked after for 12 months	-0.148	0.511	0.772	-0.020	0.081	0.807
School %FSM high vs low	0.020	0.162	0.901	0.002	0.014	0.903
Ofsted 2018 (outstanding as base):						
Good	-0.062	0.286	0.828	-0.004	0.028	0.873
Requires improvement/satisfactory	-0.282	0.321	0.381	-0.024	0.029	0.413
Inadequate	-0.229	0.327	0.484	-0.023	0.032	0.468
Ofsted missing	0.026	0.332	0.938	-0.002	0.032	0.956
IDACI rank	0.000	0.000	0.880	0.000	0.000	0.845
School types (Academy as base):						
Community school	-0.113	0.278	0.683	-0.012	0.030	0.701
Voluntary aided school	-0.224	0.282	0.428	-0.023	0.027	0.389
Foundation school	-0.705*	0.279	0.011	-0.061**	0.023	0.010
Free school - Mainstream	-0.605***	0.166	0.000	-0.053***	0.016	0.001
Special free school	-0.890**	0.315	0.005	-0.075**	0.027	0.007
Free school UTC	2.431***	0.289	0.000	0.453***	0.024	0.000
Region (East midlands as base):						
East of England	0.299	0.340	0.380	0.026	0.030	0.392
London	0.093	0.288	0.747	0.008	0.024	0.752
North East	-0.027	0.300	0.928	0.001	0.024	0.960
North West	0.207	0.380	0.587	0.018	0.035	0.597
South East	0.624	0.344	0.070	0.065	0.039	0.095
South West	-0.018	0.371	0.961	0.003	0.028	0.926
West Midlands	0.016	0.303	0.958	0.002	0.024	0.939
Yorkshire and the Humber	0.424	0.315	0.179	0.041	0.032	0.196
KS2 English (80-90 as base):						
KS2 English between 91-100	0.009	0.094	0.924	0.001	0.009	0.940
KS2 English between 101-110	-0.010	0.110	0.926	-0.001	0.010	0.915
KS2 English between 111-120	-0.025	0.137	0.852	-0.002	0.012	0.872
KS2 maths (80-90 as base):						
KS2 maths between 91-100	0.297*	0.116	0.010	0.028*	0.011	0.010
KS2 maths between 101-110	0.184	0.133	0.167	0.017	0.012	0.154
KS2 maths between 111-120	-0.037	0.187	0.844	0.001	0.015	0.969
Constant	-2.782***	0.471	0.000	0.050	0.040	0.210
Observations	24727			24727		
R-squared				0.025		
Pseudo R-squared	0.034					

Note: Population of Year 11 AM schools, all pupils (n=159 schools). School-level clustered residuals.

P-values: * <0.1; ** <0.05; *** <0.001.

AM, Academic Mentoring; Coef, coefficient; EAL, English as an Additional Language; FSM, Free School Meals; IDACI, Income Deprivation Affecting Children Index; KS, Key Stage; Ofsted, Office for Standards in Education; S.E., standard error; SEN, special educational needs; UTC, university technical college.

Table C8: Predictive power of participation model

	Actual AM		
Predicted AM	0	1	Total
0, N	22,185	2,485	24,670
0, %	89.93	10.07	100
1, N	2,446	343	2,789
1, %	87.7	12.3	100
Total, N	24,631	2,828	27,459
Total, %	89.7	10.3	100

Source: Year 11 population data.

Note: y=1 if $\hat{y}$>0.5

AM, Academic Mentoring; N, number.

This document is available for download at https://educationendowmentfoundation.org.uk

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

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn