EVALUATION OF YEAR 1 OF THE TUITION PARTNERS PROGRAMME: IMPACT EVALUATION FOR YEAR 11

Evaluation Report: An exploration of impact in Year 11

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

Report authors

NFER: Helen Poet, Pippa Lord 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

The impact evaluation of the first year of the Tuition Partners (TP) programme (2020/21) was conducted by the National Foundation for Educational Research (NFER) and the University of Westminster.

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

Firstly, we are grateful to all the pupils, school staff, tutors, and Tuition Partners who generously gave their time to take part in the research.

We would like to thank the following members of NFER's operations team who designed the data collection templates and process for the TPs, recruited the schools to the evaluation sample and collated the monitoring and assessment data: Kathryn Hurd, Jishi Jose, Guido Miani, Alison Hale, Shazia Ishaq, Matt Ryan, Amanda Barber, Dan Finn, Tom Shipston, and Chirag Chitroda. Thanks also to Ruth Staunton for her statistical work on the monitoring data, and Emma Hawkins for her efficiency 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 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, Emily Yeomans, and Camilla Nevill.

Notice

This work contains statistical data from the Office for National Statistics (ONS) which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

About the first year of the National Tutoring Programme Tuition Partners

The National Tutoring Programme (NTP) Tuition Partners (TP) programme was designed to offer tutoring support for pupils as a response to the Covid-19 pandemic, and to provide a longer term contribution to closing the attainment gap.[1] The focus was on supporting disadvantaged pupils, including those eligible for Pupil Premium (PP) funding, Free School Meals (FSM) or those identified by schools as having an equivalent need for support.[2] Participating schools were able to identify which of their pupils they felt would benefit from additional support, and decide whether face-to-face or online tuition would be more suitable for them in the current environment.

There was also a second strand to the NTP – Academic Mentoring – which placed trained staff in schools to provide within school tutoring. This part of the NTP was delivered by Teach First. This report focuses specifically on the TP part of the NTP.

The Education Endowment Foundation (EEF) oversaw the delivery of this programme in the academic year 2020/21, starting on the 2nd November 2020 and finishing at the end of August 2021, which included selecting and managing the Tuition Partners (TPs). Thirty-three approved TPs delivered the tutoring, offering a range of tutoring approaches to state-maintained schools throughout England. These approaches included online and face-to-face models, and small-group and 1:1 tuition.

About this study

The EEF commissioned an independent evaluation of the TP programme, led by the National Foundation for Educational Research (NFER) along with Kantar Public and the University of Westminster. The evaluation aimed to quantify the overall impact of year 1 of the TP programme on pupil attainment/learning outcomes, and how this varied by different types of tutoring, pupil, and school characteristics. The study also evaluated the implementation of the programme including the experiences of schools, tutors and pupils, in order to improve the delivery of similar programmes in the future.

About this report volume

This report covers findings from the analysis of the impact evaluation of year 1 of the TP programme (2020/21) for Year 11.

This volume outlines the impact of TP on learning outcomes for Year 11 pupils, through a number of estimators of impact, in both English and maths. The outcome data used in this volume are the Teacher Assessed Grades (TAGs) awarded to Year 11 pupils in the summer of 2021. This report briefly summarises the results of a number of checks carried out to inform the approach to the analysis and interpretation of the results. We also report the findings from the moderator analysis on Year 11 data, which explores how different models of tutoring (e.g. face-to-face vs online; 1:1 vs small group) correlate with TAGs.

[1] Additional information from the EEF: The TP programme was designed to encourage the uptake of tutoring, with the intention of supporting tutoring to become a ‘go to’ choice that schools make to support pupils in the future. In the long-term, and due to the strong evidence around the potential impact of tutoring, it was intended that tutoring would contribute to a closing of the attainment gap. With evidence that the attainment gap has grown over the academic years 2019/20 and 2020/21 and with restricted attendance in schools over both of these years it was not intended that the TP programme would contribute to the closing of the attainment gap in the shorter term, but it was hoped that it would ameliorate some of the negative effects of schools closures in Year 1.

[2] Additional information from the EEF: School freedom around the choice of pupils was an important design feature of year 1 of the TP programme. Due to the unique circumstances of the 2019/20 and 2020/21 academic years it was clear that many families had changing circumstances and pupils would be facing a range of new challenges, including: becoming newly disadvantaged due to socio-economic changes for their families; specific challenges associated with accessing remote learning; missing face-to-face teaching due to both systemic school closures to most pupils, but also their own individual circumstances (e.g. illness, being in a Clinically Extremely Vulnerable category); other changes to family circumstances such as the death or long-term illness of family members. Many of these challenges would not have shown in a change to a pupil’s Pupil Premium status, and even a socio-economic change takes time to be reflected in Pupil Premium status with this information usually taken according to the pupil’s status in January of the prior school year.

How to cite this volume: Oppedisano, V., Poet, H., Zhang, M., Lord P., Styles, B. and Dorsett, R. (2022). *Evaluation of Year 1 of the Tuition Partners Programme: Impact Evaluation Report for Year 11. Evaluation Report: An exploration of impact in Year 11*. NFER and University of Westminster. Part of an overarching evaluation of year 1 of the Tuition Partners programme, conducted by NFER, the University of Westminster and Kantar Public. London: Education Endowment Foundation.

Other volumes in the series

This report is part of a series of volumes on the evaluation of year 1 of the National Tutoring Programme: Tuition Partners. Other volumes in the series are:

- Evaluation of year 1 of the Tuition Partners programme: **Impact evaluation for primary schools**
- Evaluation of year 1 of the Tuition Partners programme: **Implementation and process evaluation**
- Evaluation of year 1 of the Tuition Partners programme: **Summary and interpretation of key findings**

Executive summary

The project

The National Tutoring Programme (NTP) Tuition Partners (TP) programme was designed to provide additional support to schools and teachers to supplement classroom teaching through subsidised, high quality tutoring for pupils from an approved list of tutoring organisations, the Tuition Partners. This evaluation covers the TP programme as delivered in its first year by the Education Endowment Foundation (EEF), from November 2020 to August 2021. Tuition Partners 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 and originally developed by the EEF, Nesta, Impetus, The Sutton Trust, and Teach First, and with the support of the KPMG Foundation.

The EEF appointed 33 approved 'Tuition Partners' that schools could select from to deliver tuition. Schools could access 15 hours of tutoring per selected pupil (with a minimum of 12 hours being considered a completed block of tuition). Tuition was provided online and/or face-to-face; and was 1:1, or in small groups (1:2 or 1:3); and available in English, maths, science, humanities and modern foreign languages. Tuition was expected to be delivered in schools (before, during and after school), in addition to usual teaching; and in certain circumstances, at home. The programme was targeted at disadvantaged pupils attending state-maintained schools in England, including those eligible for Pupil Premium funding (PP-eligible), Free School Meals (FSM), or those identified by schools as having an equivalent need for support. Participating schools had discretion to identify which of their pupils they felt would most benefit from additional tuition support. Pupils in Years 1–11 were eligible (5–16 years old). The programme aimed to reach 215,000 to 265,000 pupils, across 6,000 state-maintained schools in England, and it was expected that approximately 20,000 tutors would be recruited by Tuition Partners.

The TP programme was set up and delivered during the Covid-19 pandemic, requiring continued responsiveness to the challenges faced by schools including restricted attendance, remote teaching, and ongoing widespread staff and pupil absences. During school closures to most pupils from January – March 2021, the EEF approved TPs to deliver online tuition at home, however many schools chose to wait to commence tutoring until schools reopened fully, and therefore started tutoring later than planned. The usual summer exams process for Year 11 pupils could not go ahead as planned in summer 2021, and GCSEs were determined by TAGs instead.

This evaluation report covers the analysis on the impact of the TP programme on the maths and English attainment outcomes for Year 11 pupils only. Separate reports relate to analysis on a sample of primary schools and an implementation and process evaluation (IPE). The evaluation findings for the TP programme are brought together in a summary and interpretation report that is available **here**.

This evaluation uses a quasi-experimental design (QED), involving a group of intervention schools that participated in the TP programme, and a group of comparison schools that did not receive the programme. The evaluation relies on a propensity score matching approach to ensure that the intervention and comparison schools are similar to each other in important, observable regards. As pupils who would have received TP in comparison schools were difficult to identify, the evaluation focused on pupils eligible for Pupil Premium and on all pupils, as these groups can be identified in both TP and non-TP schools. The analysis is based on 1,464 secondary schools with a total of 62,024 pupils eligible for Pupil Premium. The evaluation assessed impact in English and maths using Teacher Assessed Grades (TAGs) from 2021.

Table 1: Summary of findings

Finding
Initial checks on the data indicated that the Teacher Assessed Grades (TAGs) would be suitable as an outcome measure for some exploratory analysis into the impact of TP, 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 TP made similar progress in English and maths compared to pupils eligible for Pupil Premium in comparison schools (there was no evidence of an effect in English or maths). A particular challenge is that, on average, only 12% of pupils eligible for Pupil Premium were selected for tutoring in maths and 9% were selected for tutoring in English, meaning the vast majority of the pupils included in the analysis did not receive tutoring. Therefore, this estimated impact of TP is diluted and it is hard to detect any effect that may (or may not) be present.

When looking at all pupils in Year 11, pupils in schools that received TP made, on average, similar progress in English compared to all Year 11 pupils in comparison schools (there was no evidence of an effect). In maths, Year 11 pupils in schools that received TP made slightly less progress than all Year 11 pupils in comparison schools (though this effect was very small and equivalent to zero months ' additional progress). However, this analysis was subject to even further dilution than the PP-eligible analysis: only 7% of Year 11 pupils were selected for tutoring in maths and 6% in English. Given this context, it is unlikely that any of these differences were due to TP.

Additional analysis restricted the sample of schools to those that targeted higher proportions of pupils eligible for Pupil Premium to receive tutoring, to reduce the issue of dilution and bring the group of analysed pupils closer to those that were selected for the intervention. In schools that selected over 50% of pupils eligible for Pupil Premium for tutoring, pupils eligible for Pupil Premium made similar progress in TP and comparison schools in English and maths. However, when the sample was restricted to schools that selected over 70% of pupils eligible for Pupil Premium for tutoring (and reducing dilution further), the impact of TP on pupils eligible for Pupil Premium is positive. In these schools, pupils eligible for Pupil Premium made, on average, the equivalent of two months additional progress in English and two months additional progress in maths, compared to pupils eligible for Pupil Premium in comparison schools. This analysis was based on a smaller sample of schools that were rematched to a comparison sample. However, different characteristics to the rest of the TP population of schools remained (more 'Outstanding' schools, lower percentage of FSM students), so this finding may not necessarily be generalisable to all TP schools.

Within schools that participated in TP, pupils who received more hours of tutoring in maths obtained higher maths TAGs, and pupils who received more hours of tutoring in English obtained higher English TAGs, than pupils who received fewer hours of tutoring in the respective subjects. These results are associations and are not necessarily causal estimates of impact; there may be other explanations for the higher grades among these pupils.

EEF security rating

The security of findings is usually described through a padlock classification assigned to the primary outcome of the evaluation (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 evaluation was GCSE outcomes. During the course of delivery it became clear that exams in Summer Term 2021 would not go ahead as planned and were replaced with TAGs for the 2021 exam session in response to the ongoing disruption to schools in the 2020/21 academic year. As TAGs have not been used as a research outcome measure before, 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, some 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 TP and non-TP schools over the exam years analysed, although it is not possible to confirm that this is certainly the case. The analysis proceeded on an exploratory basis and therefore the findings need to be treated with caution.

The evaluation also contended with the challenges of the pandemic, meaning not all planned analyses could go ahead. This included plans to measure impact by identifying the characteristics of pupils who participated in TP, so that a matched sample of pupils in comparison schools with similar 'observable' characteristics could be created, and compare the outcomes across both groups of 'predicted' participants. However, it was not possible to accurately predict which pupils participated in TP using available data and this impact analysis did not go ahead. The IPE findings showed that schools used a wide definition of disadvantage when selecting which pupils to receive tutoring which was not narrowly confined to Pupil Premium eligibility. Schools also included 'any pupils whose attainment had suffered' as being disadvantaged, as well as selecting pupils who they perceived as more likely to benefit from and engage with the tutoring. These characteristics cannot be observed or isolated within the available datasets.

In addition to the limitations of the TAGs, the study had several other related limitations that must be taken into account. The inability to randomise and control for unobservable characteristics regarding school and pupil selection into tutoring; the difficulty of identifying the pupil-level counterfactual (pupils that would have participated in TP in non-TP schools) the quality and completeness of the participation data (including information on dosage); and the dilution of any impact in pre-identified groups of pupils (specifically pupils eligible for Pupil Premium, not all of whom received TP). It should be noted that the high dilution is driven by the extent to which pupils eligible for Pupil Premium were selected to participate in TP (or not), as well as by the total number of pupils who participated in TP in the school. With such high dilution, it was unlikely that the analyses focusing on pupils eligible for Pupil Premium and on all pupils would be able to detect an effect. Due to a combination of these factors the estimates are for groups of pupils that do not directly align with the group of pupils that participated in TP. Although the intervention group (TP schools) and comparison group (non-TP schools) were well balanced in terms of observable school-level characteristics, the design was not fully equipped to deal with the way schools actually selected pupils to participate in TP.

In addition, pupils selected for tutoring received on average fewer hours of tutoring by the time of the end-line assessment than had been anticipated (at a pupil-level average, for PP-eligible pupils, 7.6 hours in English and 8.4 hours in maths compared to the expected minimum of 12 hours). This was in part due to delivery shifting to later in the academic year because of restricted attendance at schools in the spring term 2021. The number of hours received was lower than the minimum 12 hours expected, and may mean it was harder to detect an effect of the programme.

When the evaluation tried to address the dilution issue by selecting a (small) sample of schools that targeted over 70% of pupils eligible for Pupil Premium for tutoring, and a matched sample of comparison schools, the result showed a positive impact of TP availability on both maths and English, suggesting that dilution may be preventing the evaluation from detecting significant impact of the TP intervention in the main analyses. This finding, combined with positive associations between the amount of tutoring received and attainment scores, are in line with the evidence on tutoring in the EEF Toolkit that summarises that tutoring has, on average, a positive impact on pupil attainment.

The evaluators recommend that in future years of the TP programme, efforts are made to evaluate different types of tutoring with a pupil-randomised design, for example by varying the number of hours of tuition or how many sessions of tutoring per week are delivered to explore the optimum dosage and pattern of delivery.

Introduction

Background

In response to the Covid-19 pandemic, the government asked all schools in England to close to most pupils in March 2020. 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 2020/21 academic year, 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 and young people who needed the most help (for example, the disadvantaged and vulnerable groups that will have been affected most). In 2020/21, the NTP was made up of two pillars: the Tuition Partners (TP) programme (which provided tutoring support to pupils), and Academic Mentoring (in which mentors were placed in schools to work with small groups of pupils). The EEF was awarded £80,153,065 for delivery of TP during the 2020/21 academic year.

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 five months and four months, respectively) - particularly where they are targeted at pupils' specific needs. Most of the research on small group tuition has been conducted on reading, with an impact on average of + 4 months. The studies in maths show a slightly smaller positive impact (+ 3 months). Impact tends to be greater in primary schools (+ 4 months) than secondary schools, which has fewer studies overall and a lower impact (+ 2 months). 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 (Dietrichson et al., 2017; Torgerson et al., 2018). 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[3]).

Intervention

This evaluation is on year 1 of the TP programme, which is summarised below using the EEF's TIDieR[4] framework:

- **Why:** Research shows that pupils' learning has been affected by school closures to most pupils, and that tutoring is an effective means of support.
- **Who:** The programme was designed to provide additional support to schools to help disadvantaged pupils, including those eligible for Pupil Premium (PP-eligible) funding, Free School Meals (FSM) or those identified by schools as having an equivalent need for support. Schools were able to identify which of their pupils they felt would most benefit from additional tuition support.
- **What (resources):** Tuition was provided to schools at a 75% subsidy, with schools paying 25% of the cost.
- **Who (provider):** The NTP appointed 33 approved Tuition Partners (TPs) who were expected to deliver tutoring via 20,000 tutors. Schools would be able to access high-quality tuition from these approved partners.
- **How (format):** A range of tutoring models were provided, including those suitable for pupils with SEND and in alternative provision. It was provided online and/or face-to-face; and was 1:1, or in small groups (1:2 or 1:3); and available in English, maths, science, humanities and modern foreign languages.
- **Where (location):** Tuition was expected to be delivered in schools (before, during and after school), in addition to usual teaching; and in certain circumstances, at home.
- **When and how much (dosage):** Tutoring took place in the academic year 2020/21. Schools could access 15 hours of tutoring per selected pupil (with a minimum of 12 hours being considered a completed block of tuition).
- **Tailoring:** A range of models were offered, and TPs could adapt their models with capacity building support from Nesta/Impetus throughout the year. To support increased tuition delivery in the shorter time available once

[3] The EEF carried out an online feasibility pilot in preparation: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/online-tuition-pilot?utm_source=/projects-and-evaluation/projects/online-tuition-pilot&utm_medium=search&utm_campaign=site_search&search_term=online
[4] TIDieR stands for Template for Intervention Description and Replication.

schools reopened fully, the EEF introduced more flexibility to the offer, including expanding online at home tuition into weekend and half-term provision, extending the TP programme into the summer holidays. This had implications for the amount of tutoring received by the point of the summer assessment, as discussed later in the report.

The TP programme was set up and delivered during the Covid-19 pandemic, requiring continued responsiveness to the challenges faced by schools including restricted attendance, remote teaching, and ongoing widespread staff and pupil absences. The IPE report found that despite being developed and delivered within a relatively short time frame for a programme of this scale, and in the context of ongoing disruption due to the pandemic, the programme was broadly implemented as intended. However, TPs and schools responded to relatively open aspects of the TP programme by implementing it in different ways – allowing them to adapt delivery to their varying needs and circumstances, while also resulting in variations in reach and perceived quality and impact. Furthermore, during the school closures to most pupils from January – March 2021, the EEF approved TPs to deliver online tuition at home, however many schools chose to wait to commence tutoring until schools reopened fully, and therefore started tutoring later than planned.

Further information about the programme design and its development – including the logic model - is provided in the IPE report.

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

Version 2 of the study plan explains a number of changes that needed to be made to the design of the evaluation, in response to the national lockdown involving school closures to most pupils, which had implications for tuition delivery (Spring 2021).

The evaluation in secondary schools was changed to only evaluate attainment at Year 11 at secondary school level, as it was not possible to recruit enough secondary schools to the evaluation sample. More information can be found in the study plan, and in the ‘Methods’ section below.

Due to the continued disruption to education in the 2020/21 academic year because of the pandemic, the government decided to replace the exam requirement for GCSEs with Teacher Assessed Grades (TAGs) for summer 2021. As this data was somewhat of an unknown quantity, there were a number of potential issues identified relating to using the TAGs as an outcome measure. In version 2 of the study plan we outlined both these issues for consideration and some checks to be run on the TAG data – the results of which are outlined in the report below.

Evaluation objectives

The overarching objective for the impact evaluation was to investigate the impact of TP on learning outcomes for pupils. This was investigated through a number of estimators of impact, in both English and maths, in both primary and secondary schools. This report contains the findings from the analysis on Year 11 pupils. The findings from the impact evaluation for primary schools are published in a **separate report**.

The research questions (RQs) outlined in the study plan which formed the impact evaluation are listed below, with the RQ numbering for this Year 11 report, and the primary school impact evaluation (reported separately).

One of the research questions focuses on all PP-eligible pupils in the year groups involved (RQ4a1) as a way of identifying would-be participants and avoiding selection bias. Any effect of tutoring would be ‘diluted’ amongst all the PP-eligible pupils (as not all would take part in TP), but this was outweighed against 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 a year group, for RQ4a3) would be selected for TP, these research questions therefore look at the impact of the *availability* of TP and not the impact of actual participation. These issues are discussed in further detail in the *Pupil-level selection* section.

Table 2: Summary of research questions (RQs)

	Primary school impact evaluation RQ number	Year 11 impact evaluation RQ number **(This report)**
Outcome analysis:		
What is the impact of TP availability on all PP-eligible pupils' attainment?	RQ1	RQ4a1
What is the impact of TP on the attainment of pupils participating due to encouragement to do so?	n/a	RQ1b
What is the impact of the intensity of TP (dosage) on all PP-eligible pupils' attainment?	RQ1c	RQ4b
Further analysis:		
What is the impact of TP availability on predicted participants' attainment?	RQ2	RQ4a2
What is the impact of TP availability on all pupils' attainment?	RQ3	RQ4a3
Moderator analysis:		
How does the association of TP availability with attainment vary among PP-eligible pupils, by school and pupil characteristics?[5]	RQ5	RQ5
How do outcomes vary among TP pupils, by model of tutoring?	RQ6	RQ6

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.

Schools agreed to take part in the programme by the headteacher signing a Memorandum of Understanding (MoU) (a copy of this can be found in Appendix A.1[6]).

All participants (parents, Key Stage [KS] 4 pupils, tutors, school staff and TP staff) were provided with a privacy notice relevant to processing their (or their child's) personal data. Participants could withdraw from data processing at any time during the evaluation – and instructions of how to do so were provided in the privacy notice and evaluation information sheet (see Appendices A.3 and A.4).

Data protection

All work conducted by the consortium for the impact analysis was compliant with the Data Protection Act 2018 (DPA) and General Data Protection Regulation (GDPR). NFER has ISO27001 and Cyber Essentials Plus certifications and registration with the Information Commissioner's Office.

The EEF, NFER and Kantar 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 University of Westminster (UoW) identified the following legal basis:

[5] In the study plan this RQ was worded as impact rather than association, however the analysis is not causal so the research question wording has been updated.

[6] Appendices numbered A.x can be found in the separate 'Impact Appendices' document

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 TP 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.*

Data controller and processing roles

The Department for Education (DfE), the EEF, and the Evaluator (the consortium of NFER, UoW and Kantar) were joint data controllers for the evaluation. The Evaluator was also a data processor for the evaluation, as were Tuition Partners.

Rights and retention periods

Parents (and KS4 pupils) could withdraw their child from the TP 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 Tuition Partners or the evaluation), the Evaluator would still use the evaluation data that the school provided up to that point and link it to the National Pupil Database (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 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 one year after the publication of the final report.

The Tuition Partner will securely delete any personal data collected for the evaluation alone at the end of the TP programme, when final grants have been paid. The Tuition Partner may keep personal data collected as part of the delivery of their tuition 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 NPD. The UoW 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 UoW team 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 did 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 UoW:

- Richard Dorsett, Professor of Economic Evaluation (UoW)
- Veruska Oppedisano, Senior Lecturer (UoW)
- Helen Poet, Senior Research Manager (NFER)
- Pippa Lord, Trials Director and Consortium Lead (NFER)
- Ben Styles, Head of Classroom Practice and Workforce (NFER)
- Min Zhang, Research Fellow (UoW)
- Greta Morando, Research Fellow (UoW)
- Ruth Staunton, Senior Statistician (NFER)

They were supported by the operations and research team at NFER for the collation of monitoring data about the schools, tutors and pupils taking part. The NFER operations team included:

- Jishi Jose, Project Manager
- Kathryn Hurd, Head of Survey Operations
- Guido Miani, Project Manager
- Alison Hale, Senior Project Manager
- Emma Hawkins, Senior Business Support Manager
- Amanda Barber, Data Management Administrator
- Shazia Ishaq, Senior Data Manager
- Daniel Finn, Data Management Unit Lead
- Tom Shipston, Junior Data Manager
- Chirag Chitroda, Senior Data Manager
- Matthew Ryan, Junior Project Manager

Methods

Evaluation design

Table 3: Evaluation design

Design		Matching Instrumental variables
Unit of treatment		Year 11 Pupils eligible for Pupil Premium (PP, identified as pupils eligible for FSM in the previous six years, NPD variable: EverFSM6)[7]
Stratification variable(s) (if applicable)		n/a
Number of units to be included in analysis (intervention, comparison)		1,464 secondary schools (732 intervention, 732 comparison schools). 62,024 PP-eligible pupils (31,516 from intervention schools and 30,508 from comparison schools)
Primary outcome	Variable	Attainment in English and maths
	Measure (instrument, scale, source)	GCSEs awarded using the TAG process 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 and maths
	Measure (instrument, scale, source)	KS2 SATs from the NPD
Baseline for secondary outcome(s)	Variable	n/a
	Measure (instrument, scale, source)	n/a

The impact evaluation uses a quasi-experimental design (QED) involving a comparison group and a number of estimators of impact rather than a randomised controlled trial (RCT), due to the need to maximise reach to as many schools and pupils as possible.

The challenge, as with any quasi-experimental impact evaluation, is that the selection of schools and pupils into the TP programme is unlikely to be random. We used propensity score matching to control for school selection into TP by constructing a matched comparison group of non-TP schools that was similar in important, observable regards to the TP schools in the population (all TP schools in Year 11) sample (details below in Propensity score matching in the Statistical analysis section). This assumes that sufficient school characteristics can be observed to control for selection (the ‘selection on observables’ or ‘conditional independence’ assumption). It is this type of selection that Weidmann and

[7] Note that the unit of analysis does not directly overlap with the unit of treatment (i.e., not all PP-eligible pupils were selected to receive TP. This is referred to as dilution in this report).

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' (i.e., what schools were doing anyway). However, in the context of Covid-19 recovery, it was likely that pupils who were not selected for TP were provided with other forms of support by schools, and these may have involved 1:1 or small group support.

Participant selection

Selection into the programme and implications for the evaluation

All state-maintained primary, secondary and special schools could access tuition through the TP programme during its first year 2020/21. A total of 6,082 schools signed up to the programme in its first year (according to the monitoring data provided by TPs). Schools could choose which Tuition Partner(s) they wished to work with, and were responsible for identifying pupils for tuition – in which year groups and which subjects. There was no prior information available about which schools would or would not be more likely to use TP.

The programme focused on supporting disadvantaged pupils, including those eligible for Pupil Premium. Participating schools were able to identify which of their pupils they felt would most benefit from additional tuition support, as outlined in the guidance to TPs:

> The focus of the NTP is on supporting disadvantaged pupils aged 5-16. Schools should therefore be asked to focus on disadvantaged pupils, including pupils eligible for Pupil Premium funding, Free School Meals or those identified by schools as having an equivalent need for support. Participating schools will be able to decide which of their pupils will most benefit from additional support.'

Since the group of eligible pupils could not be identified within comparison schools in advance, we attempted to move away from pupil-level selection by focusing the analysis on PP-eligible pupils and all pupils, as these groups could be identified for both TP and comparison schools. PP status was chosen because PP-eligible pupils were expected to represent the core of the eligible group based on the online tutoring pilot[8] and the guidance (above) for schools. However, as noted in more detail in *Pupil-level selection* section (see *Results*), this was a challenge as the PP-eligible group was much less well aligned with treatment status than had been anticipated in advance of the evaluation. We refer to this issue as dilution: the analysis is conducted on a group (PP-eligible pupils) which does not directly align with those participating in the intervention and so the results include pupils who were not selected for TP (as well as some that were). Note, we identified PP-eligible pupils as pupils eligible for FSM in the past 6 years in the NPD.[9]

The guidance stated that pupils selected for tuition could take part in up to 15 hours tuition in one subject through the TP programme. Pupils could be in Year 1 to Year 11. The programme was expected to reach 215,000 to 265,000 pupils in its first year.

Study participants and inclusion criteria

The Year 11 population analysis focuses on secondary schools where at least one pupil in Year 11 was selected for TP.[10] In order to be a 'TP' (intervention) school in the analysis, the tuition 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 of 11 June 2021, which was one week before the deadline for schools to submit their TAGs to the exam boards (assuming some time for schools to complete internal moderation and/or quality assurance processes).

There were 1,679 unique secondary schools accessing TP across all year groups. The data collected from TPs indicated that 47% of secondary schools (787 schools) had at least one pupil in Year 11 that was selected for TP. Of these, we

[8] In the online tutoring pilot in the summer term of 2020 that preceded TP, over 60% of targeted learners were PP-eligible pupils (Marshall *et al.,* 2021).
[9] Eligibility for Pupil Premium is defined on the basis of FSM eligibility, care leavers, and looked after children. We used the FSM, criterion only as we do not have access to data on care leavers.
[10] In any subject - also see '*Participant flow*' section.

dropped from the analysis schools that delivered tuition after the proxy assessment date of 11 June 2021,[11] schools which had missing KS2 and/or TAGs and schools off common support, as described in the '*Participant flow including losses and exclusions*' section (see Figure 1). This resulted in 732 TP (intervention) schools in the analysis.

Selection of the comparison group and identification assumptions

We selected the same number of comparison schools using propensity score matching (based on characteristics listed in the '*Propensity score matching*' subsection under the '*Statistical analysis*' section). The comparison schools were selected among schools which did not participate in TP in Year 11 and did not participate in the Academic Mentoring pillar. The school-level data that we have on the Academic Mentoring pillar included 367 secondary schools. Only 160 of them include detailed pupil-level information on who was selected for the intervention. For 207 secondary schools that participated in Academic Mentoring we do not know if pupils in Year 11 were selected to receive mentoring or not. We therefore removed all 367 schools that delivered Academic Mentoring from the sample of potential comparison schools to account for this uncertainty. We did not remove from the sample TP schools that also delivered Academic Mentoring (13.3% of TP schools, see Table 7a) and we control for participation in Academic Mentoring in all regressions.

In summary, after cleaning the data and performing the matching, the final sample was composed of 732 TP schools that had at least one Year 11 pupil who was selected for TP and 732 comparison schools.

Outcome measures

Baseline measures

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

Primary outcome

The evaluation aims to measure the impact of tutoring on attainment, as the purpose of the TP programme is to support pupils to 'catch up' and reduce the amount of missed learning due to the Covid-19 pandemic and the restrictions on schools in 2020 and 2021. The analysis used teacher assessed GCSE grades (TAGs) for 2021 that were made available in the NPD as the outcome measure.

Originally, GCSE scores were planned to be used as outcome measures for English and maths attainment for Year 11 pupils. However, due to the school closures to most pupils from January 2021 because of Covid-19, the government announced that the usual summer exams process could not go ahead as planned; this was replaced with TAGs. We were concerned that the process for grade determination may mean that it was difficult to detect any potential impact of the TP programme. This was regarding whether we could use the TAGs as an outcome measure for this particular analysis, and was not a reflection on the TAGs themselves. We therefore carried out a number of checks on the Year 11 TAGs prior to proceeding with the analysis in the '*Statistical analysis*' section, as well as identifying caveats to consider when reporting the results. The checks are summarised in the '*Impact evaluation results*' section (with more detail in the Appendices). It is worth noting that they cannot confirm the adequacy of TAGs for our purposes, they can only provide some indication of support. As a result of the change to the outcome measure, the analysis in this report should be considered exploratory.

Pupil-level TP participation data

The lists of pupils who were selected for TP along with a set of intervention, school and pupil characteristics were collected by the TPs who shared this information with NFER.[12] TPs were required to submit this data regularly to the EEF as part of their contractual requirement, with the knowledge that it would be used as part of the evaluation. This dataset included information about participation in TP, including models of tutoring (face-to-face or online, timing of sessions, group size and so on) and the information about sessions booked and completed.

At the end of the evaluation year, NFER checked and cleaned the data before sharing the datasets securely with the ONS. These checks showed that the data was not complete. There were gaps in terms of fields completed and

[11] We did not consider these schools as potential comparison schools, to avoid reassigning TP marker status.

[12] More information about the data collected, and the process of collecting it can be found in the IPE report, see subsection 'Research methods' in section 'IPE methods'.

inconsistencies observed, for example sessions completed did not always align with sessions booked. This has implications for the evaluation as the analysis could only use the data supplied. Due to the size of the dataset and the multiple sources (TPs) of the data it was known from the start that it would not be possible to rectify this in the data and the analysis would need to proceed with the data as supplied.

The data included school and pupil identifiers to allow the NPD team to match them to the NPD data. Once matched, the NPD removed the identifiers and retained a meaningless identifier. It was not possible for the NPD team to match all of the pupil data to the NPD due to missing or incorrect pupil identification details; around a fifth[13] of all pupil records were lost during the match and these pupils are therefore not included in the analysis.

Sample size

We used cluster randomised trial power calculations to provide an indication of the minimum detectable effect size (MDES) for the PP-eligible pupils analysis (RQ4a1). We allowed for clustering of pupils within schools.[14]

During the design phase, the assumed sample was based on a number of 70 PP-eligible pupils per secondary school. Note that the intention-to-treat effect is more diluted the lower the proportion of PP-eligible pupils per school who were selected for the intervention. We assumed an intra-cluster correlation (ICC) of 0.15 and pre–post correlations of approximately 0.7. Note that since our analysis focuses on disadvantaged pupils, we do not produce separate estimates for all pupils.

At the time of calculating the MDES, there were 1,554 unique secondary schools doing TP in 2020/21.[15] We assumed in the study plan that 80% of secondary schools would have usable pupil data and tutoring in Year 11, which would leave us with 1,243 schools. Power calculations suggested that with 1,243 TP secondary schools and 1,243 comparison schools the MDES would be 0.03.

As mentioned in the '*Participant selection'* section, the data collected from TPs indicated that in actual fact only 732 schools had at least one Year 11 pupil that was selected for TP and that were eligible for the analysis. The updated MDES is 0.04 (Table 5).

Statistical analysis

Propensity score matching

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

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 Table 5 of the Study Plan below for reference and in Appendix A:

- KS1 to KS2 value added attainment, at district level in 2018/19[16]
- Management/school type secondary - Community, Academies, Foundation, Free schools, Sponsored Academies, Voluntary school, Studio schools, University Technical College
- School size, total number of students in previous 3 years[17]
- Teacher - student ratio, in 2017/18 and 2018/19[18]
- Ofsted, overall effectiveness, 2017/2018, and 2018/19[19]
- Region (London, Government Office Region, and regional dummies)

[13] It is difficult to be more precise than this, due to the sensitivities of checking the data pre- (outside SRS) and post- (in SRS) match.
[14] MDES are computed also for studies not based on a randomised intervention, as QEDs.
[15] The number of secondary schools that delivered TP was 1,679 by the end of the academic year 2020/21. However, at the time of calculating the MDES, in the autumn of 2021, the data available indicated that the number of secondary schools delivering TP was 1,554.
[16] The most recent data available in the data we ingested in the SRS in November 2021 in the school-level file is from the academic year 2018/19. The data include KS1 to KS2 value added attainment, but not the average attainment at KS1, as stated in the study plan, hence we replaced average attainment at KS1 with KS1 to KS2 value added attainment.
[17] We kept variables for the previous 3 years, rather than only from the previous year, as specified in the study plan, because we performed the placebo analysis in the three pre-intervention years. Hence, it seemed appropriate to control for measures of the matching variables in the past.
[18] See footnote 13.
[19] See footnote 13.

- School in urban/rural area
- Income Deprivation Affecting Children Index (IDACI) quintile, in previous 3 years
- Interaction of IDACI tertiles with average attainment in previous 3 years
- FSM - percentage eligible in previous 4 years
- English as an Additional Language (EAL) - percentage in previous 3 years; and
- Special educational needs (SEN) - percentage in previous 3 years.

Instead of the percentage of students who achieved expected standard in KS2 in the year before, which was not available for all years in our dataset, we used the following measures of baseline and outcome attainment:

- English and maths baseline KS2 attainment of pupils in Year 11 in 2020/21; and
- English and maths KS4 attainment in the previous 3 years.

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 1,679 schools in the TP school-level population file and 367 secondary schools that delivered Academic Mentoring. Thirteen schools that had missing KS2 and/or TAGs were removed before matching as they were not eligible. 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 TP 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 TP 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).[20] Common support is shown in Figure A1 in Appendix A: seven TP schools (denoted in green) have very high propensity scores (i.e. extremely likely to be chosen to be TP schools). Propensity scores between 0.08 and 0.78 were kept. As no comparison schools scored higher than 0.78, these seven schools are dropped due to lack of common support.

Table A3a, A3b, and A3c in Appendix A report different specifications using different calipers. Matching quality is measured by differences in attainment between TP and matched comparison 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 TP and comparison 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 (58 when using 0.05 caliper and 61 when using 0.01 caliper). 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 TP 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 TP schools and comparison schools. This is discussed in the '*Pupil and school characteristics*' section, which shows that the two samples were sufficiently balanced to proceed with the analysis.

Weighting method

We apply inverse probability weights to all regressions to control for differences in school size and to give all schools the same weight.

Preliminary analysis

[20] There is a bias-variance trade-off when deciding between 1:1 matching and many-to-one. There is less bias with pair matching (1:1) because the match is closer, but less efficiency because we are discarding data. With many-to-one, we get a little more bias, but smaller variance. We decided to tolerate less efficiency for less bias.

To test whether TP schools are similar to the comparison schools, we conducted placebo tests using the data from the NPD. If the selection of the comparison group controls adequately for unobserved factors, there should be no significant difference in attainment between TP schools and comparison schools in years prior to the intervention. The placebo testing was done for the preceding years 2016/17, 2017/18, and 2018/19 using GCSE results. We excluded GCSE scores from 2019/20 as they were determined outside of the usual exam process and replaced by centre assessed grades.

The placebo testing, conducted at school level, had two main elements:

1. Identification of target comparison schools. As mentioned above, we assessed the performance of the match by: i) comparing observed characteristics of TP schools and their matched comparators (Table 7a); and ii) comparing baseline assessment (KS2) of TP schools and their matched comparators (Table 7b). This was a school-level analysis but outcomes are considered for PP-eligible pupils.
2. Assessment of the performance of potential estimation approaches. We had planned to compare observed characteristics of TP and matched comparator schools using ordinary least squares (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. We only apply inverse probability weighting to control for school size. We also applied the difference-in-differences estimation approach in pre-TP years as an additional form of placebo test, comparing GCSE results between TP and matched comparison schools in 2016/17 and 2017/18 with GCSE between TP and matched comparison schools in 2017/18 and 2018/19, respectively. Results are reported in the '*Placebo tests*' section.

Analyses

Due to the likely difficulty of identifying the counterfactual (pupils that would have participated in TP in comparison schools), we presented several estimation methods in the study plan, with the intention of assessing different estimates of impact – these are outlined below. Although the sample size calculations are provided for the PP-eligible pupils' analysis (RQ4a1), there was no single primary outcome identified by design.

The first research question was designed to be on PP-eligible pupils because of the specific objective of the programme to help disadvantaged pupils, whose learning had particularly suffered during the course of the pandemic. While schools had discretion over which pupils would receive tutoring, we anticipated that, due to the focus on supporting disadvantaged pupils and the guidance provided to schools, a high proportion of PP-eligible pupils would be selected. Any effect of tutoring would be 'diluted' among all the PP-eligible pupils analysed (as not all would take part), but this was outweighed against being able to identify a majority-type of potential participants (i.e. Pupil Premium; PP) in both intervention and comparison groups. We refer to the issue of dilution throughout the report. As this issue was recognised during the design phase, a number of different estimators were included in order to attempt to account for different selection mechanisms that may have been used by schools. These are described below in the subsequent research questions.

The analysis of the impact of TP availability and TP dosage was designed to be based on two estimators: matched regression and instrumental variable (IV) regression.

Outcome analysis

Regression (RQ4a1): What is the impact of TP availability on all PP-eligible pupils' attainment?

To estimate impacts, we regressed the pupil-level outcome on two measures of TP: i) a 0/1 indicator for TP being available at school level, i.e. if at least one pupil in Year 11 was selected to receive TP; and ii) a categorical variable measuring the school-level average number of blocks of hours completed by the time of the assessment (dosage). Dosage was categorised as the number of blocks of hours of tutoring completed per pupil, averaged at the school level (less than one block of 12 hours completed per pupil on average, or more than one block completed per pupil on average) with respect to TP intervention schools with zero dosage.

All TP sessions taken after 11 June 2021 are not considered in the dosage analysis as we assume they took place too late to influence the work used to inform the TAGs. TP schools that delivered all TP sessions after 11 June 2021 are excluded from the TP sample (see Figure 1 in the '*Participants flow*' section).

We ran an OLS model. We used KS2 scores 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, such as gender, ethnicity, EAL, care status, and SEN. All pupil-level and school-level controls are listed in Appendix B. Residuals were clustered at the school level to account for any common school-specific unobservable component. Regression was based on pupils in the TP schools and their matched comparators. The software used to run the model was Stata.

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

The proportion of Year 11 PP-eligible pupils selected for TP is 25.6%. To address the fact that the population reached in practice was characterised by a wide definition of disadvantage,[21] and that Pupil Premium status was not a majority-characteristic,[22] we provided two additional analyses, proposed in version 2 of the study plan, that restrict the analysis to the sample of TP schools that targeted at least 50% and 70% of PP-eligible pupils for tuition. These are 26% and 8.6% of the population of TP schools in the study. They are described in the '*Additional analyses*' section.

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. The coefficients of the Romano-Wolf correction are slightly different from the coefficients of the OLS specification as Romano-Wolf requires the same set of controls in the two outcomes. Hence, we control for both baseline maths and English scores in the Romano-Wolf specification while only for the subject in the outcome in the OLS specification. We do not present the coefficients associated with the Romano-Wolf multiple hypothesis testing correction for the school-level dosage analysis since the correction can only be applied within a treatment, and not across multiple treatments.

Instrumental Variables (RQ1b): What is the impact of TP on the attainment of pupils participating due to encouragement to do so?

We aimed to use a second technique of IVs to provide estimates of TP that do not rely on the selection on observables assumption. Note that this analysis was not able to proceed, as explained in the '*Impact evaluation results'* section.

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. This approach builds on the 'reach and engagement' RCTs,[23] which aimed to test methods of increasing take-up of tutoring among pupils selected for tutoring. More information on each trial is provided in the study plan.

Two trials randomised tutors to interventions that aimed to improve the relationship with the pupils and to improve pupil attendance at tutoring sessions: the first trial leveraged similarities between pupils and tutors; the second trial aimed to improve tutors' relational self-efficacy. The third trial randomised pupils to an intervention that consisted of weekly motivational messages. In the second and third trials, the interventions were not effective at increasing attendance at tutoring sessions, while in the first trial the intervention had a positive impact on pupil attendance. We then tested whether the trials induced sufficient additional take-up of TP for them to be effective instruments (as opposed to 'weak' instruments). Estimation was based on pupils in TP schools only, and on the subgroup of PP-eligible pupils.

Instrumental Variables (RQ4b): What is the impact of the intensity of TP (dosage) on the attainment of all PP-eligible pupils?

We planned to use a second IV analysis to provide estimates of TP that do not rely on the selection on observables assumption. Note that this analysis was not able to proceed, as explained in the '*Impact evaluation results*' section.

The approach uses only the sample of intervention TP schools. 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

[21] Schools used their discretion to select pupils for tuition, and identified pupils based on how likely they felt they were to engage with, and benefit from, tuition.
[22] Overall, less than half of the pupils selected for tutoring were eligible for Pupil Premium.
[23] The reports are available on the EEF website: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/national-tutoring-programme-nimble-rcts.

the problem of missing or unknown controls by requiring the conditional independence assumption to hold between the instrument and the outcome. In this context, IV methods rely on finding a variable that strongly predicts treatment but does not otherwise directly impact attainment.

This approach exploited the fact that schools signed up to TP at different times so there were some schools that had not yet delivered TP or had delivered it only partially by the time TAGs were submitted (in the case of this analysis, the cut-off date of 11 June 2021, one week prior to the TAG submission date). These schools were in theory similar to TP schools in terms of interest in the programme. We note that the timing of engagement in TP is non-random. We therefore, provided supplementary evidence on this point by checking that prior characteristics of schools were not related to the timing of adoption among participants.

The hypothesis was that date of signing up to TP (via the MoU) may be positively associated with dosage of tutoring and we identified that if so, it could be used as instrument in the IV regression of outcomes on the number of sessions completed. In practice, we used the number of days passed between signing the MoU and our cut-off date for the TAGs (continuous variable) as an instrument for dosage. Dosage was categorised as the number of blocks of hours of tutoring completed per pupil (less than one block of 12 hours completed per pupil on average, or more than one block completed per pupil on average) with respect to TP intervention schools with zero dosage. We implemented a weak instrument test to test the strength of the correlation between the instrument and the treatment.

The treatment–control difference in the number of TP sessions completed between schools that signed the MoU earlier versus later is an estimate of the impact of early sign up to the programme on the intervention delivery. The assumptions for the instrument to be valid are that: i) the instrument (early sign up) is a significant predictor of the treatment (number of TP sessions completed); but ii) it is uncorrelated with the outcome of interest, the TAGs. For this to be the case, higher or lower achieving schools should not systematically be the first ones that sign the MoU. As long as a mix of both high and low achieving schools signed the MoU earlier than others, the assumption could be plausible. However, the second assumption (ii) cannot be fully tested due to the presence of other unobserved school-level characteristics that it is impossible to account for. As a check, we estimated the two-stage least squares regression (2SLS) with a placebo outcome, baseline KS2 Standard Assessment Tests (SATs) of pupils in Year 11 in 2020/21.

We aimed to run two estimates: the first, as specified in the study plan, was estimated using all PP-eligible pupils in year groups doing TP; and the second, in addition to this and not included in the study plan, was estimated on all TP pupils in TP schools, regardless of their PP-eligible status. The coefficient on the dosage is the impact estimate and constitutes a local average treatment effect; the average impact among schools that completed the intervention because of an early MoU sign up. The reliability of this analysis is subject to the quality and completeness of the data received from TPs in relation to the time they delivered the sessions. This information is available for 21% of pupils in the sample of TP schools.

Further analyses

RQ4a2: What is the impact of TP availability on the attainment of pupils predicted to participate?

Our approach to the above outcome analysis (RQ4a1) 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 will receive the intervention. RQ4a2 was therefore designed to involve modelling the probability of a pupil participating in the TP programme in TP intervention schools, using various markers of disadvantage recorded in the NPD (socio-economic status measured by FSM/PP, SEND, interaction with social service, prior attainment, EAL, and ethnicity). We planned to use this model to predict who would have participated in the TP programme in both TP intervention and comparison schools. The ex post 3 Year 11 check reported in Appendix C provides details of the predictive power of the model.

RQ4a3: What is the impact of the availability of TP on all pupils' attainment in the population of schools?

As another means of understanding the overall effect of TP, a fourth analysis focused on the impact of the availability of TP on the attainment of all Year 11 pupils (rather than just PP-eligible pupils or predicted TP pupils) in the full population of secondary schools as observed in the NPD. Similar to RQ4a1, we regressed the pupil-level outcome on two measures of TP—this time on all Year 11 pupils: i) a 0/1 indicator for TP being available at the school level; and ii) a categorical variable measuring the number of blocks of hours completed by the time of the assessment (dosage) again at the school level. As with RQ4a1, dosage was categorised as the number of blocks of hours of tutoring completed per pupil, averaged at the school level (less than one block of 12 hours completed per pupil on average, or more than one block completed per pupil on average) with respect to TP intervention schools with zero dosage.

We anticipated that these estimates were likely to be smaller than RQ4a1 and RQ4a2 estimates, as the TP impact would be more diluted when considered across the entire year group because the percentage of all pupils receiving TP in TP intervention schools would be lower than the percentage of PP-eligible pupils receiving TP. This research question aims to capture the average impact of the intervention on the population of pupils whether they received the treatment or not. This estimator also captured the spillover (peer) effects. The purpose of this is to capture the overall impact of TP, though it should be borne in mind that most pupils in TP schools are not selected to receive the intervention. There would need to be a strong effect of the intervention on a small group and/or important spillover effects to gain insights from this part of the analysis. The regression analysis controls for the same school-level and pupil-level characteristics mentioned in RQ4a1 (PP-eligible pupils analysis).

Moderator analysis

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

RQ5: How does the association of TP availability with attainment 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 rating (high, outstanding, and good vs. low, inadequate, and requiring improvements); proportion of FSM (high vs. low, defined on the basis of the median); type of school (academy vs. maintained); and school size (by quartile).
2. Pupil characteristics: Prior attainment; SEND vs. not; EAL, ethnicity, and gender.[24]
3. Other: Geography (urban/rural; low/high IDACI).

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

RQ6: How do outcomes vary among TP pupils, by model of tutoring?

A descriptive analysis (using the data collected via templates for the above impact analysis) compared outcomes associated with different tutoring models among TP schools.[25] We did not propose any impact analysis within RQ6 since we cannot observe the counterfactual treatment model among comparison schools. Instead, this element of the analysis summarises the mean attainment among participating pupils in TP schools according to the model of tutoring they experienced. Hence, the coefficients may reflect how students were selected by model of tutoring but there are also likely to be other, unmeasured or unobservable factors that influence allocation to tutoring model, which we cannot account for here.

We regressed attainment on the variables listed below for the sample of TP schools and TP participants only to assess heterogeneity. In particular, we looked at the following variables at pupil level:

1. The intervention: Mode of delivery of completed sessions (online vs. face-to-face); timing of the session (during vs. after lessons); tutor:pupil ratio (1:1 vs. 1:2 vs. 1:3); number of blocks schools scheduled on average at the school level and for each pupil taking TP in a specific subject (low/high buy-in schools); school and pupil level number of blocks (high/low dosage); intensity of delivery (determined by sessions attended/number of weeks tutoring is spread over); and completed versus scheduled sessions (determined as high if 80% of the sessions are completed, equivalent to 12 or more sessions out of 15).
2. Tutors: Experience/qualifications; TP tutor training; and shared characteristics with pupil/tutee (gender and ethnicity).

[24] School attendance, listed in the study plan, could not be explored as the data were not available in the NPD in autumn 2021.

[25] The RQ6 analysis was carried out on Year 11 data. This is a change from the study plan, where we planned to conduct it on the evaluation sample of secondary schools. At the time of writing the study plan, it was not clear if it would be possible to conduct the analysis using Year 11 data, which has been since confirmed. Additionally, performing the analysis on Year 11 data and reporting it in the context of all the other Year 11 findings is more coherent than providing the results based on the secondary schools sample for just this one RQ.

3. Other: Early/late delivery.

Missing data analysis

We did not expect to find missing values at school level, as we are using the population of schools in England, and all schools provided TAGs for Year 11 pupils. For missing data on variables used for matching and on covariates, we defined them as missing and controlled for them in the analysis. More information about missing data can be found in the '*Impact evaluation results*' section.

Estimation of effect sizes

Estimates will be presented as effect sizes, calculated using the Hedges' g formula. 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	Consortium
Early November 2020	Tuition Partners (TP) launch. TP evaluation guidance pack launch. TP can start contacting schools	NFER and EEF
November 2020 – July 2021	*Tutoring period (whole programme)*	TPs
End November 2020 – December 2020	Study plan finalisation and publish	Consortium
Early December 2020	Submit National Pupil Database (NPD) request	University of Westminster (UoW)
December 2020	First population data uploads; compilation and checks	NFER
5 January 2021 – 8 March 2021	National lockdown period—many pupils learning from home, schools only open to children of key workers and vulnerable children. TP provision predominantly online during this period.	
End March 2021	Second population data uploads; compilation and checks	NFER
End May 2021	Feedback presentation to TPs	Consortium
18 June 2021	Deadline for schools to submit their Teacher Assessed Grades (TAGs) for GCSEs (Year 11)	
End August 2021	Final population data uploads from TPs	NFER
Mid August 2021 to November 2021	Data cleaning; Send data to NPD to match in	NFER
January 2022	NPD (unamended) data available and matched into dataset	NPD team/UoW
January 2021 – March 2022	Impact analysis	UoW
February 2022 to June 2022	Draft reporting	NPD team/UoW
June 2022 – September 2022	Final reporting and revisions	All
October 2022	Publication	

Impact evaluation results

Participant flow including losses and exclusions

Some TP schools were lost from the analysis due to problems linking the school-level data across the TP and the NPD files.[26] We assumed the school-level ID from the NPD was more reliable than the ID from the TP files. Schools with wrong or inconsistent ID were dropped from the analysis as they could not be merged to the NPD file (22 Academic Mentoring schools). All schools in the population of the TP school-level file (1,679) and in the population of Academic Mentoring school-level file (367) were dropped from the sample of potential comparison schools and are not included in the sample of schools considered for the analysis.[27]

Figure 1 below provides details for the flow of participants through the study. There were 787 schools with at least one Year 11 pupil selected for TP (in any subject) and pupil-level data out of 1,679 secondary schools in the TP school-level population file. About 6% (48 schools) of the 787 TP schools were lost because they did not meet the inclusion criteria: we excluded from the sample pupils who took their first TP session after 11 June 2021, pupils whose first and last date of TP sessions were missing, and pupils whose first date of TP session was after the last date of TP session. This resulted in a loss of 28 schools in the TP sample. Thirteen schools were lost in the TP sample because of missing baseline KS2 test and/or TAGs in one of the two subjects. Seven schools were lost in the TP sample because they had both started delivering TP sessions after 11 June 2021 and have missing KS2 and/or TAGs. No school in the comparison sample was dropped because of missing KS2 and/or TAGs. The school-level dropping due to missing KS2 and TAGs is concentrated only in TP schools because the whole school is dropped from the analysis if all TP pupils in the TP schools have missing tests.

We then lost less than 1% (7 schools) of the remaining 739 TP schools at the matching stage, to improve the quality of the match with comparison schools. In terms of the comparison schools, 596 schools (about 45%) were lost at the matching stage as we matched schools 1:1.

In a small number of cases (32 schools, including TP and comparison schools), pupil-level attainment data for the academic year 2020/21 could not be matched to school-level data. This occurred for relatively new schools, for example. For these cases, we replaced school-level missing information with a dummy for the specific category being missing. In this way, the data for these schools is retained in the analysis.

Our final sample included 732 TP schools and 732 matched comparison schools. TP schools are defined as schools where at least one pupil was selected for TP prior to 11 June 2021 in any subject, as indicated in the study plan. As the main analysis was on the availability of TP in the school (i.e. schools with at least one pupil in Year 11 receiving TP), we did not restrict the definition of TP schools in the maths and English samples to account for the subject where TP was actually received. In the TP sample of 732 schools, 121 schools were selected for TP in maths only, 98 schools were selected for TP in English only, 459 were selected for TP in both subjects and 54 schools were selected for TP in subjects other than maths and English (such as science, modern foreign languages and others). Our analysis focuses on maths and English attainment as outcomes. On average, 43.5% of TP pupils were selected for TP in maths in TP schools. This means that the maths results do not hold for 20% of the sample of schools (152/732) and for 56.5% of pupils in TP schools. The English analysis fails to account for pupils in the 54 and 121 schools who were selected to receive tutoring in subjects other than English. On average, 34% of TP pupils were selected for TP in English in TP schools. This means that the maths results do not hold for 24% of the sample of schools (175/732) and for 66% of pupils in TP schools.

[26] 619 pupils in the TP pupil-level file could not be linked to the NPD TAGs data. As they cannot be linked to the NPD, we cannot identify their true UPN and school ID.

[27] We initially planned to keep schools in the comparison group that were stated to have commenced tutoring or whose MoU date was after 11 June 2021, i.e. those with TP delivery after the point of assessment. However, this information seemed inconsistent across school-level and pupil-level TP files, so we decided to drop all schools in the TP school-level population file from the comparison sample to ensure that as far as possible there were no TP schools in the comparison group.

Figure 1: Participant flow diagram (two arms), for PP-eligible pupils

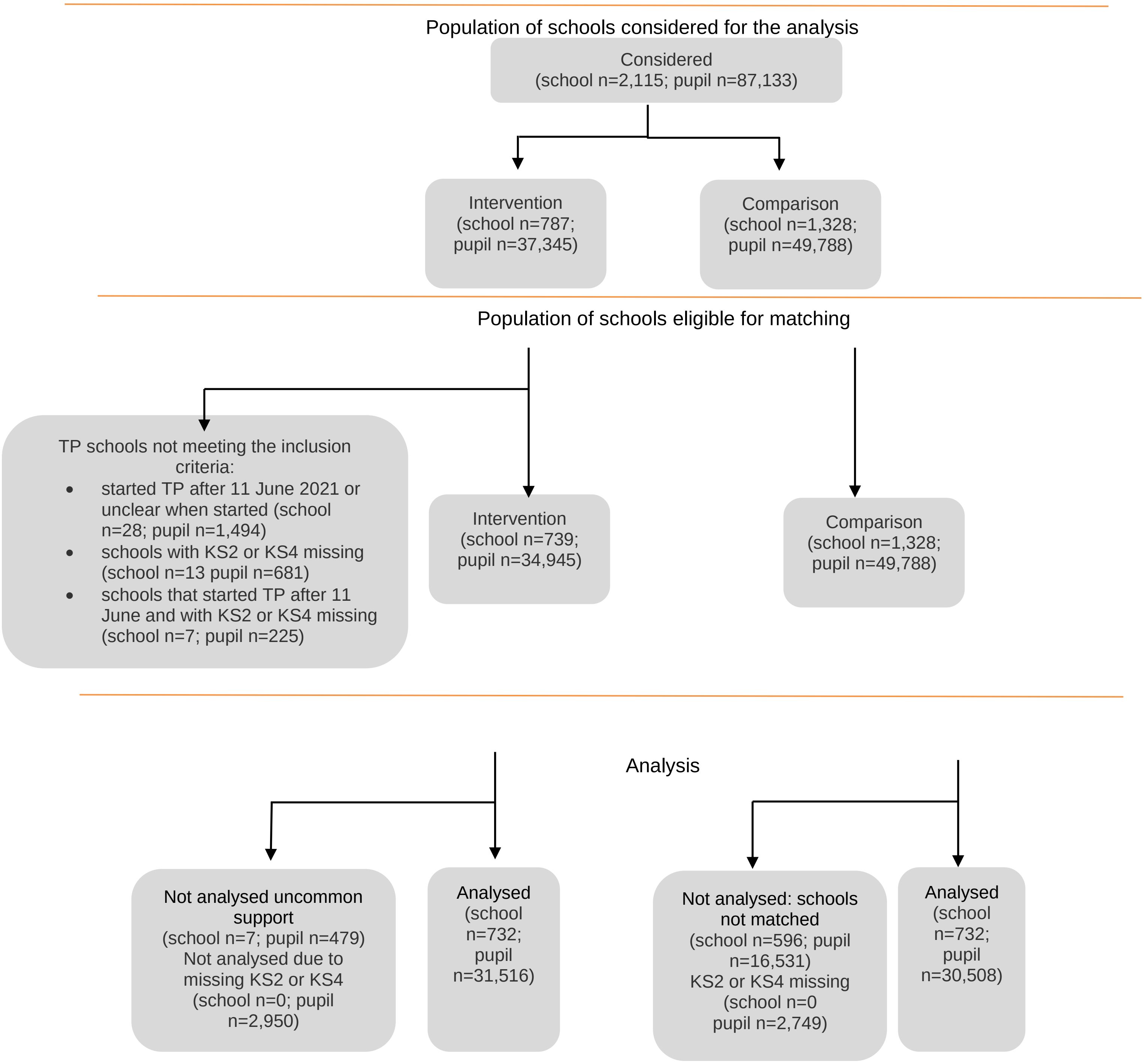

Population of schools considered for the analysis
Considered (school n=2,115; pupil n=87,133)
Intervention (school n=787; pupil n=37,345)
Comparison (school n=1,328; pupil n=49,788)
Population of schools eligible for matching
TP schools not meeting the inclusion criteria:
• started TP after 11 June 2021 or unclear when started (school n=28; pupil n=1,494)
• schools with KS2 or KS4 missing (school n=13 pupil n=681)
• schools that started TP after 11 June and with KS2 or KS4 missing (school n=7; pupil n=225)
Intervention (school n=739; pupil n=34,945)
Comparison (school n=1,328; pupil n=49,788)
Analysis
Not analysed uncommon support (school n=7; pupil n=479) Not analysed due to missing KS2 or KS4 (school n=0; pupil n=2,950)
Analysed (school n=732; pupil n=31,516)
Not analysed: schools not matched (school n=596; pupil n=16,531) KS2 or KS4 missing (school n=0 pupil n=2,749)
Analysed (school n=732; pupil n=30,508)

Table 5: Minimum detectable effect size at different stages, RQ4a1, PP-eligible pupils

		Study plan	Analysis	
		Maths and English	Maths	English
MDES		0.03	0.04	0.04
Pre-test/post-test correlations	Level 1 and 2 (pupil and school)	0.70	0.51	0.37
ICCs	Level (school)	0.15	0.11	0.12
Alpha		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		two-sided	two-sided	two-sided
Average cluster size (PP-eligible pupils per school)		70	42	42
Number of schools	Intervention	1,243	732	732
	Comparison	1,243	732	732
	Total:	2,486	1,460	1,460
Number of pupils	Intervention	87,010	31,516	31,516
	Comparison	87,010	30,508	30,508
	Total:	174,020	62,024	62,024

The difference between the assumed numbers of schools and children and the numbers analysed slightly alters the power of the analysis. Table 5 presents this in the form of the MDES. The MDES for RQ4a1 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 0.03 for both maths and English on the PP-eligible sample. This is set out in Table 5 above, along with the other assumptions used.

The sample loss of TP schools is substantial, and it occurred at the eligibility stage to include only schools that provided pupil-level data on Year 11 pupils selected for the TP intervention. A total of 892 schools out of 1,679 secondary schools in the TP school-level file had not provided any pupil-level information and they were dropped from the sample of schools considered for the analysis.

The top row of Table 5 presents the MDES for the analysis 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 slightly lower than that assumed at the design stage (0.11 for maths and 0.12 for English compared to 0.15), the pupil level and school level correlation is lower (0.51 for maths and 0.37 for English compared to 0.70)[28] and the cluster size is lower (42 for maths and English compared to 70).[29] Together, these have the effect of slightly increasing the MDES to 0.04 for both maths and English. However, the calculations do not take into account the level of dilution so not even that which would have been anticipated at the start of the intervention. If only a percentage X of PP-eligible pupils actually receive the intervention, the MDES for those treated should be multiplied by 1/X.

Pupil-level attrition

Table 6 reports pupil-level attrition. The main reason for attrition is due to sample loss because of missing KS2 and TAGs (8.6% of pupils in the TP matched sample and 8.3% of pupils in the comparison sample).

In particular, there are 2,950 pupils with missing for KS2 and 237 pupils with missing KS4 in TP schools and 2,539 pupils with missing KS2 and 275 pupils with missing KS4 in comparison schools.[30] For KS2 the missing occurred years before the intervention, and hence it should not lead to any bias. Pupils with missing primary baseline and/or outcome were excluded from the analysis.

Table 6: Pupil-level attrition from the trial (RQ1a)

		Intervention group	Comparison group	Total
Number of pupils	Matched	34,466	33,257	67,723
	Analysed	31,516	30,508	62,024
Pupil attrition (from matching to analysis)	N	2,950	2,749	5,699
	%	8.56	8.27	8.42

Source: Year 11 population data

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 TP schools, matched comparison schools and the national averages. As the evidence from the preliminary analysis indicates (see Table 7a and Table 7b and the '*Preliminary analysis*' subsection), the quality of the match is sufficiently good that weights are not needed to restore balance between TP and comparison schools. Hence, we only present results controlling for school size related weights.

The descriptives (Table 7a) indicate that all observable characteristics are well balanced between the TP schools and comparison schools. The only significant difference is the proportion of pupils with White ethnicity: comparison schools have 0.07% higher proportion of pupils with White ethnicity than TP schools.[31] Participation in Academic Mentoring programme is also significantly different between TP and comparison schools as by construction comparison schools have been selected from the sample of schools that did not participate in the Academic Mentoring programme.

[28] The assumptions were based on GCSEs as there was no prior information on the correlation between observable characteristics and the TAGs.

[29] The assumptions we made in the study plan was based, by mistake, on the TP take up rather than on PP-eligible pupils. In addition, it was based on PP-eligible pupils across year groups in secondary schools, not only those in Year 11. Hence the discrepancy in the two numbers.

[30] The numbers do not add up to the total number of pupils with missing KS2 and/or KS4 because of some overlap, i.e. pupils with both KS2 and KS4 missing.

[31] For all the categorical variables we also performed a chi-squared test of the frequency between TP and comparison schools. In all cases, the test rejects the null hypothesis that the variables are independent.

Comparison with the national average indicates that the TP and comparison samples are similar in terms of observable characteristics to the population of secondary schools. Schools in the TP and comparison samples feature a slightly higher proportion of schools with ‘Good’ Ofsted ratings than the population of schools and a slightly lower proportion of schools with ‘Inadequate’ or ‘Missing’ Ofsted ratings than the population of schools.

School types (e.g. Academies, maintained) are similarly distributed between the TP and comparison samples and the population of schools, except for ‘Foundation schools’, slightly higher in the TP and comparison samples than in the population of schools.

There are slightly more schools in urban areas in the TP and comparison samples than in the population of schools (86% in TP and in comparison schools vs. 83% in the population of schools), but fewer missing observations in the urban/rural category than in the population of schools. TP and comparison samples feature a slightly different geographical representation, with fewer schools from East Midlands, North and South East in the TP and comparison sample compared to the national population. The TP and comparison samples have a very similar composition of pupils in terms of ethnic background and measures of disadvantage.

Overall, it appears that the TP and matched comparison schools present similarity among almost all observable characteristics, making us confident to consider school-level selection to be accounted for by the matching procedure.

Table 7a: Baseline characteristics of Year 11 TP schools, matched comparison schools, national proportions, and PP-eligible pupils

Variable	Means: National data	Means: Comparison	SD: Comparison	Means: TP schools	SD: TP schools	Difference TP and comparison
School-level ALL KS4 English 2019/20	5.079	5.032	(0.732)	5.034	(0.647)	0.002
School-level ALL KS4 Maths 2019/20	4.954	4.903	(0.815)	4.892	(0.718)	-0.011
School-level ALL KS4 English 2018/19	4.687	4.642	(0.792)	4.644	(0.661)	0.002
School-level ALL KS4 Maths 2018/19	4.636	4.567	(0.880)	4.565	(0.744)	-0.002
School-level ALL KS4 English 2017/18	4.686	4.653	(0.736)	4.651	(0.661)	-0.002
School-level ALL KS4 Maths 2017/18	4.632	4.578	(0.832)	4.568	(0.718)	-0.011
School-level ALL KS2 Read scores 2019/20	31.197	31.029	(2.986)	31.029	(2.791)	-0.001
School-level ALL KS2 Maths scores 2019/20	71.393	71.167	(6.192)	71.212	(5.625)	0.044
School-level ALL KS2 Read scores 2018/19	31.129	31.005	(2.948)	30.952	(2.765)	-0.053
School-level ALL KS2 Maths scores 2018/19	71.478	71.249	(6.477)	71.168	(5.966)	-0.080
School-level ALL KS2 Read scores 2017/18	32.677	32.570	(3.182)	32.519	(2.956)	-0.051
School-level ALL KS2 Maths scores 2017/18	70.701	70.540	(6.483)	70.495	(5.867)	-0.045
Total pupil counts	925.059	950.387	(346.652)	953.323	(307.166)	2.936
Pupils-to-teacher ratio 2018	16.240	16.240	(2.799)	16.247	(2.667)	0.007
Ofsted 2018: Outstanding	0.207	0.190		0.195		0.005
Ofsted 2018: Good	0.525	0.577		0.563		-0.014
Ofsted 2018: Inadequate	0.047	0.037		0.038		0.001
Ofsted 2018: Requires improvement	0.149	0.148		0.152		0.004
Ofsted 2018: Missing	0.072	0.049		0.052		0.003
School type: Academy-sponsor led	0.233	0.221		0.224		0.003
School type: Community school	0.099	0.094		0.097		0.003
School type: Voluntary aided/controlled school	0.076	0.077		0.081		0.004
School type: Foundation school	0.052	0.061		0.066		0.004
School type: Free school - mainstream	0.480	0.478		0.473		-0.005
School type: Others	0.061	0.068		0.060		-0.008
Urban	0.828	0.859		0.861		0.001
Rural	0.137	0.122		0.119		-0.003
Urban/Rural missing	0.035	0.019		0.020		0.001
Region: East Midlands	0.084	0.074		0.070		-0.004
Region: East of England	0.108	0.107		0.115		0.008
Region: London	0.146	0.189		0.197		0.008
Region: North East	0.041	0.037		0.034		-0.003
Region: North West	0.135	0.124		0.137		0.012
Region: South East	0.150	0.137		0.130		-0.007
Region: South West	0.094	0.119		0.098		-0.020
Region: West Midlands	0.113	0.102		0.107		0.004
Region: Yorkshire & the Humber	0.094	0.093		0.093		0.000
Region: Missing	0.035	0.019		0.020		0.001
AM participation	0.064	0.000		0.133		0.133***
Census school-level % FSM Spring 2021	0.280	0.290		0.295		0.005
% EAL	0.152	0.162		0.176		0.014
% SEN	0.218	0.216		0.218		0.002
% Female	0.498	0.503		0.504		0.001
Average IDACI scores	0.038	0.040	(0.024)	0.040	(0.020)	0.000
% White British	0.117	0.117		0.110		-0.007*
% Asian	0.018	0.018		0.020		0.002
% Black	0.016	0.018		0.020		0.002
% Other ethnic	0.026	0.029		0.030		0.001
% Unknown ethnic	0.004	0.004		0.004		0.000
% Not white	0.060	0.065		0.070		0.005
Observations		732		732		1464

Source: Year 11 population data.

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 for the 2021 cohort of Year 11 pupils. The results are shown in Table 7b for maths and English, respectively.

When looking at school-level attainment, KS2 are well balanced between TP and comparison schools, with no statistically significant differences between the two samples.

Pupil-level KS2 are similarly distributed between TP and comparison schools in the sample of PP-eligible pupils. However when looking at pupil-level KS2 scores in the sample of all pupils, pupils in TP schools have statistically significantly lower KS2 maths and English (reading) scores, by 0.06% and 0.08%, respectively. As the same variables are not significantly different between TP and comparison schools when compared at the school level, school size may drive the difference at pupil level. We weighted each pupil by 1 over the number of pupils in the schools to assign each school the same weight and compared again the KS2 between TP and comparison schools. The significant differences disappeared in KS2 when inverse probability weights were applied (last two columns in Table 7b).

We apply inverse probability weights to all regressions to control for differences in school size and to give all schools the same weight. Results without weights are mentioned in the '*Outcome analysis*' section.

Compared with the national average, TP and comparison schools have similar KS2 to the national population in both the PP-eligible pupils and all pupil specifications.

Table 7b: KS2 scores for 2021 cohort of Year 11s in TP schools, matched comparison schools, and national data, PP-eligible pupils, and all pupils

All pupils Variable	National Averages	Means:Comparison	SD:Comparison	Means:TP schools	SD:TP schools	Difference	Difference TP and non-TP, weighted	Difference TP and non-TP, weighted and std
KS2 Maths score, school level	102.510	102.381	(2.882)	102.326	(2.614)	-0.055		
KS2 Read scores,school level	103.002	102.907	(2.513)	102.886	(2.169)	-0.021		
KS2 Maths score, pupil level	103.188	103.221	(6.716)	103.159	(6.585)	-0.062**	-0.024	-0.007
KS2 Read scores, pupil level	102.742	102.682	(8.080)	102.606	(8.011)	-0.076**	-0.054	-0.007
Observations		119,682		121,821		241,503	241,145	241,145
PP pupils Variable	National Averages	Means:Comparison	SD:Comparison	Means:TP schools	SD:TP schools	Difference		
KS2 Maths score, school level	100.544	101.329	(2.858)	101.357	(2.488)	0.028		
KS2 Read scores,school level	101.196	100.581	(3.106)	100.487	(2.731)	-0.094		
KS2 Maths score	100.966	101.217	(6.827)	101.302	(6.701)	0.085	0.020	0.009
KS2 Read scores	100.027	100.171	(8.023)	100.190	(7.932)	0.019	-0.149	0.002
Observations		30,508		31,516		62,024		

Source: Year 11 population data.

Pupil-level selection: observable characteristics

The TP programme was intended to reach disadvantaged pupils including PP-eligible pupils funding, FSM, or those identified by schools as having an equivalent need for support. As noted in the '*Participant selection*' section (see '*Methods*' section), schools were able to decide which pupils would be selected for tutoring, and while there were no formal targets for who should be reached it was anticipated by the NTP that Pupil Premium eligibility would be one of the key markers of disadvantage that would likely inform selection of pupils into the TP programme. Indeed, this had been the case in the online tutoring pilot during the summer term of 2020 (where over 60% of targeted learners were eligible for Pupil Premium; Marshall *et al.,* 2021). However, as identified in the IPE report, in year 1 of the TP programme, schools considered a much wider range of factors, including those such as motivation or perceived likelihood to make the most of tutoring (which are not observable in the dataset).

The decision to use PP-eligible pupils in the analysis was to avoid the complication of pupil selection as a result of school decision that arises when schools decide which pupils take part in interventions (i.e. pupil-level selection bias). This is because PP-eligible pupils can be identified in the datasets in both TP and comparison schools. This only holds if PP-eligible pupils were actually the ones targeted by the intervention when delivery occurred. However, we know from the IPE findings that overall under half (46%) of the pupils taking part in TP were eligible for PP. Similarly, in the samples analysed here only two-fifths of Year 11 pupils that were selected for TP were PP-eligible pupils (40%; Table 8a).

For context, around a quarter of Year 11 pupils were eligible for PP within the TP schools in our analysed sample (Table 8a: 26% of pupils). So, while the proportion of PP-eligible pupils selected for tuition was higher than the proportion of PP-eligible pupils in the intervention schools, they did not form the majority of the pupils who took part.

Looking at PP-eligible pupils themselves, we see that only around a quarter (26%) of PP-eligible pupils were selected for TP in the intervention sample, which means that pupil-level selection from schools is an issue (Table 8a). Our strategy to focus on PP-eligible pupils in the analysis (identifiable in both TP and comparison schools) is therefore 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 tutoring 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 TP. If the proportion of PP-eligible pupils selected for TP had been higher then our evaluation strategy would have avoided the complication of pupil-level selection bias and would suffer less from dilution. We had also planned to analyse based on predicted participation (RQ4a2), which would have provided an alternative approach to approximating the eligible group however, as reported later on, this analysis could not proceed due to poor predictive power of the model.

In addition to the issue of dilution for the analysis on PP-eligible pupils, it is a consideration for the analysis on all pupils too. Table 8a shows that the proportion of all pupils in the intervention group (Year 11 in TP schools) who were selected for TP is 17%. This suggests that dilution is also likely to be a problem in detecting significant impact of TP in RQ4a3 (analysis on all pupils).

It should be noted that these low proportions are driven by the extent to which PP-eligible/non-PP-eligible pupils were selected, and also by the total number of pupils identified for tutoring in the school.

In addition to pupil-level and school-level selection, pupils' were selected into TP subjects. The numbers at the bottom of the section '*Participant flow including losses and exclusions*' indicate that 20% of schools in the analytical sample did not tutor in maths and 24% of schools did not tutor in English. In the analysis, we avoid trying to account for selection into subject by only selecting TP intervention schools, irrespective of the subject they tutored in. However, we can estimate the level of dilution considering subject. Overall, 11.6% of PP-eligible pupils were selected for TP in maths and 9.1% of PP-eligible pupils were selected for TP in English (see Table 8b).

For an estimate of dilution by subject for the all pupils analysis, 7.3% of all pupils were selected for TP in maths and 5.7% of all were selected for TP in English (see Table 8b).

Table 8a: Percentage of PP-eligible pupils in the sample, of PP-eligible pupils selected for TP and of selected for TP who are PP-eligible

	School counts	Pupil counts, intervention schools	Pupil counts, comparison schools	TP pupils counts, intervention schools	% of pupils who did TP, intervention schools	PP-eligible pupils counts, intervention schools	% of PP-eligible pupils, intervention schools	PP-eligible pupils counts, comparison schools	% of PP-eligible pupils, comparison schools	PP-eligible pupils doing TP counts, intervention schools	Of pupils doing TP, % PP-eligible pupils	Of PP-eligible pupils, % of pupils doing TP
TP Year 11 analysed sample	732 TP, 732 non-TP	121,821	119,682	20,398	16.74	31,516	25.87	30,508	25.49	8,059	39.51	25.57
TP Year 11 population sample	787 TP	141,687	n/a	22,539	15.91	37,345	26.36	n/a	n/a	8,958	39.74	23.99

Source: Year 11 population data.

Table 8b: Percentage of pupils tutored by subject

	PP-eligible pupils	All pupils	PP-eligible pupils counts, intervention schools	Pupil counts, intervention schools	Of PP-eligible pupils, % of PP-eligible pupils doing TP in a specific subject	Of all pupils, % of pupils doing TP in a specific subject
Pupils tutored in maths	3,661	8,870	31,516	121,821	11.62	7.28
Pupils tutored in English	2,873	6,922	31,516	121,821	9.12	5.68

Source: Year 11 population data.

We further explored the distribution of baseline scores (KS2 maths and English) of pupils selected for TP versus pupils not selected for TP in TP schools to assess the ability composition, as measured by quartiles of KS2, of TP pupils. The purpose is to describe the ability composition of TP pupils and assess whether it is different from the ability composition of pupils not selected for TP.

This analysis indicates that TP pupils have lower baseline scores than non-TP pupils, regardless of whether they are PP-eligible pupils or not, pointing towards negative selection of pupils into the TP programme (that is that pupils with lower prior performance were selected to participate in TP by schools). If lower ability pupils were selected into the programme, they would have more room for improvement, showing larger effect of the intervention.

In more detail, the results shown in Table 9, present the regression on TAGs of the interaction between a dummy equal to one for the pupil participating in TP 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. We performed the analysis on the sample of PP-eligible pupils, non-PP-eligible pupils, and all pupils in TP schools.

All three analyses, for English and maths, present negative and statistically significant interaction coefficients. The size of the coefficients is associated with higher quartiles of baseline assessment.

Table 9: Interaction of quartiles of baseline scores (KS2) with TP pupils' status in TP schools

	PP-eligible pupils						Non-PP-eligible pupils						All pupils					
	Maths			English			Maths			English			Maths			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 TP X KS2 Maths quartiles:																		
Math baseline 2nd quartile#TP	-0.299***	0.054	0.000				-0.348***	0.046	0.000				-0.353***	0.035	0.000			
Math baseline 3rd quartile#TP	-0.469***	0.072	0.000				-0.574***	0.055	0.000				-0.590***	0.044	0.000			
Math baseline 4th quartile#TP	-0.670***	0.084	0.000				-0.647***	0.064	0.000				-0.734***	0.054	0.000			
Receiving TP X KS2 English quartiles:																		
English baseline 2nd quartile#TP				-0.299***	0.054	0.000				-0.348***	0.046	0.000				-0.353***	0.035	0.000
English baseline 3rd quartile#TP				-0.469***	0.072	0.000				-0.574***	0.055	0.000				-0.590***	0.044	0.000
English baseline 4th quartile#TP				-0.670***	0.084	0.000				-0.647***	0.064	0.000				-0.734***	0.054	0.000
Constant	31516			31516			90169			90169			121821			121821		0.000
N	0.514			0.403			0.549			0.432			0.541			0.420		

Source: Year 11 TP schools data.

First quartile is the lowest. School-level clustered residuals and inverse probability weighting.

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

Amount of tutoring received

Table 10 presents the pupil-level average of blocks of tutoring pupils received by 11 June 2021. Schools could access 15 hours of tutoring per selected pupil (with a minimum of 12 hours being considered a completed block of tuition). In the sample of PP-eligible pupils, pupils selected for TP received on average 70% of a block of 12 hours in maths and 63% of a block in English, that is 8.4 hours of maths tutoring and 7.6 hours of English by 11 June 2021. In the sample of all pupils, pupils selected for TP received on average 74% of a block of 12 hours in maths and 69% of a block in English, that is 8.9 hours of maths tutoring and 8.3 hours of English by 11 June 2021. One of the reasons pupils received less than 12 hours of tuition prior to the assessment is because part of the tutoring was delivered later in the school year as a result of the early 2021 school closures to most pupils. According to the delivery data provided by TPs, of the sessions where session delivery dates were recorded, 29% of tutoring sessions happened after 11 June 2021 (note, 41% of booking rows in the full dataset across all year groups did not provide detailed dates per session).

Table 10: Pupil-level average of blocks of tutoring received, of pupils selected for TP

	N	Mean	Std. Dev.
* PP-eligible pupils			
Pupil-level maths dosages	3,661	0.697	0.437
Pupil-level English dosages	2,873	0.628	0.449
* All pupils			
Pupil-level maths dosages	8,870	0.736	0.432
Pupil-level English dosages	6,922	0.686	0.447

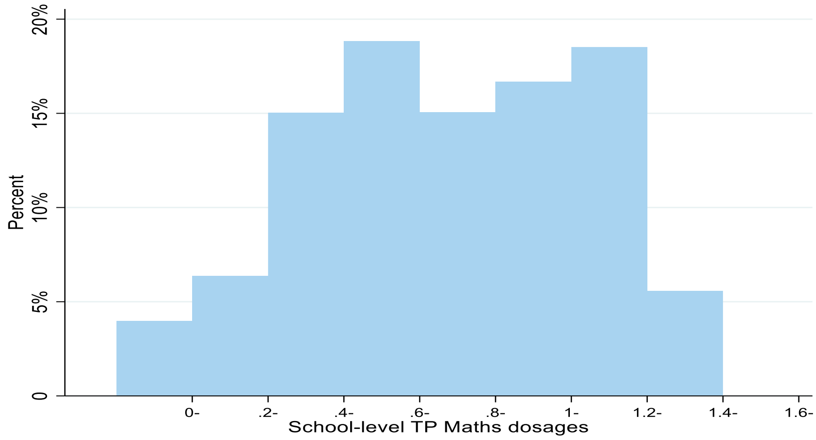
Source: Year 11 TP schools data.

Figure 2a (maths) and 2b (English) show the distribution of school-level dosage in the two samples. Dosage of tutoring is measured in number of blocks of hours completed, where one block includes 12 hours of tutoring. Less than 5% of pupils have zero dosage in one of the two subjects, which is explained by the fact that pupils with missing or blank dosage are recorded as zero.[32] The distribution also indicates that on average pupils had less than 1.25 (15 hours) dosage.

Figure 3 shows the distribution of pupil-level dosage in maths and/or English. The distribution, differently from the school level one, indicates that most pupils had either a low (zero or very few sessions) or high (one block of 12 hours of tutoring or more) level of tutoring.

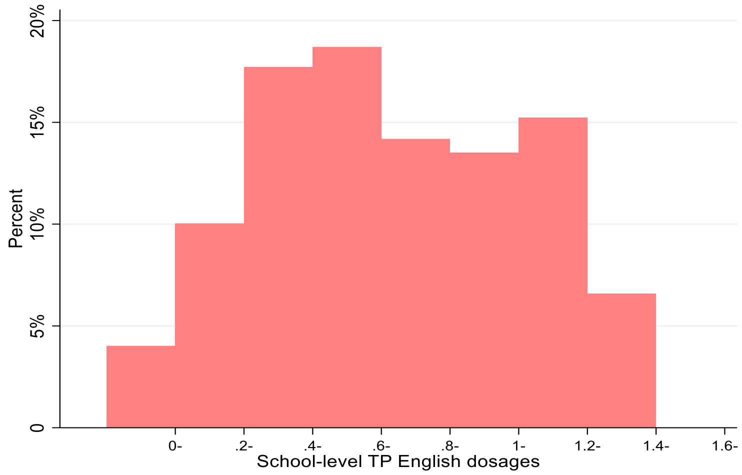
[32] For the maths sample, 26 schools (3.55% of TP schools) have zero school-level dosages. In particular, 20 schools have online dosage ‘0’ and face-to-face dosage blank and six have true zero dosage. For the English sample, 23 schools (3.14% of TP schools) have zero school-level dosages. In particular, 14 schools have online dosage ‘0’ and face-to-face dosage blank and nine have true zero dosage.

Figure 2a: Distribution of school-level dosage in terms of complete blocks for maths (where one block corresponds to 12 hours of tutoring), continuous variable



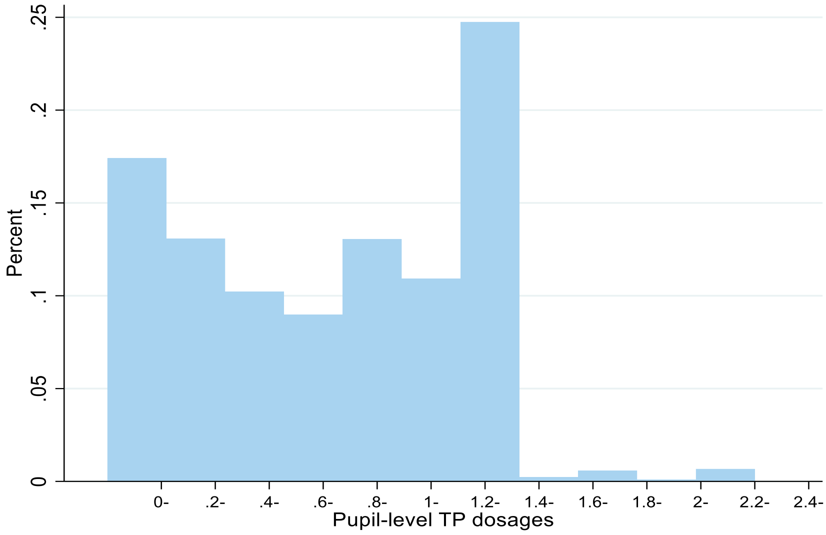
Note: The X-axis shows the percentage of pupils selected for TP. The bar left to zero (-0.2) means the category of true zero; the bar 0.2- means above 0 and lower than 0.2, excluding the zeros.

Figure 2b: Distribution of school-level dosage in terms of complete blocks for English (where one block corresponds to 12 hours of tutoring), continuous variable



Note: The X-axis shows the percentage of pupils selected for TP. The bar left to zero (-0.2) means the category of true zero; the bar 0.2- means above 0 and lower than 0.2, excluding the zeros.

Figure 3: Distribution of pupil-level dosage in terms of complete blocks for English and/or maths (where one block corresponds to 12 hours of tutoring), continuous variable



Note: The X-axis shows the percentage of pupils selected for TP. The bar left to zero (-0.2) means the category of true zero; the bar 0.2- means above 0 and lower than 0.2, excluding the zeros.

The percentage scale is shown as a proportion, so 0.05 = 5%, 0.10 = 10%, etc.

Preliminary analysis: placebo tests

This section describes the results from the preliminary placebo analysis.

Placebo testing showed that TP schools were similar to the matched comparison schools prior to the intervention.

The placebo tests were conducted on the sample of TP and matched comparison schools on KS4 maths and English assessments for the academic years 2016/17 and 2018/19. The results are reported in Table 11a and Table 11b.

In Table 11a, for each subject and year, we present different regressions on KS4 scores: in the first specification we controlled only for TP status; in the second specification we added baseline KS2 assessment; and in the third specification we included the variables listed in the '*Methods*' section and in Appendix B.

Results, across all three specifications, indicate that, before the intervention, TP schools did not show significantly different KS4 scores than matched comparison schools. In all the cases, coefficients are close to zero and not significant. This provides support for the use of the matched comparison group as a means of estimating the counterfactual.

Table 11a: Placebo test on TP schools compared to comparison schools in KS4 maths and English in pre-intervention years, all pupils and PP-eligible pupils

		All pupils											
		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
TP schools	No controls	-0.003	0.045	0.940	1381	-0.015	0.046	0.750	1304	-0.031	0.048	0.516	1225
TP schools	Only KS2	0.001	0.025	0.969	1381	-0.008	0.025	0.751	1304	-0.023	0.026	0.365	1225
TP schools	KS2+controls	-0.013	0.013	0.312	1381	-0.004	0.018	0.825	1304	-0.023	0.020	0.247	1225
		KS4 English 2018				KS4 English 2017				KS4 English 2016			
TP schools	No controls	0.001	0.040	0.990	1381	-0.004	0.041	0.927	1304	0.009	0.044	0.842	1224
TP schools	Only KS2	0.007	0.022	0.752	1381	0.008	0.023	0.718	1304	0.004	0.024	0.883	1224
TP schools	KS2+controls	-0.008	0.013	0.539	1381	-0.002	0.017	0.901	1304	-0.004	0.019	0.850	1224

		PP-eligible pupils											
		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
TP schools	No controls	0.016	0.046	0.725	1380	-0.016	0.047	0.735	1304	0.000	0.050	0.995	1218
TP schools	Only KS2	-0.009	0.027	0.744	1380	-0.041	0.028	0.144	1304	0.001	0.029	0.971	1218
TP schools	KS2+controls	-0.021	0.019	0.258	1380	-0.029	0.022	0.194	1304	0.002	0.023	0.924	1218
		KS4 English 2018				KS4 Eng 2017				KS4 English 2016			
TP schools	No controls	0.017	0.043	0.686	1380	0.007	0.044	0.863	1304	0.001	0.045	0.986	1216
TP schools	Only KS2	0.009	0.029	0.758	1380	-0.030	0.029	0.302	1304	-0.09	0.029	0.751	1216
TP schools	KS2+controls	-0.006	0.02	0.775	1380	-0.026	0.022	0.256	1304	-0.016	0.023	0.485	1216

Source: Year 11 school-level population data.

Controls are listed in Appendix B.

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

In Table 11b, we present the difference-in-differences estimates in pre-TP years, which was run as an additional form of placebo test. The estimates in the first two columns compare KS4 outcomes between TP and matched comparison schools in 2016/17 with KS4 outcomes between TP and matched comparison schools in 2017/18. We also compared outcomes between TP and matched comparison schools in 2017/18 with KS4 outcomes between TP and matched comparison schools in 2018/19.

The relevant coefficients are the interaction between the time dummy and the TP schools dummy. The coefficients of this interaction between the time dummies (2017/18 in the upper section of the table and 2018/19 in the lower section) and the dummy for TP schools are close to zero and non-significant in the sample of all pupils and PP-eligible pupils. This evidence indicates that the difference-in-differences approach is a good design to provide estimate of the counterfactual. We estimate impacts using difference-in-differences in the '*Additional analyses*' section.

Table 11b: Placebo impact of TP schools on KS4 maths and English in pre-intervention years, using difference-in-differences (Diff-in-Diff) specification, PP-eligible pupils and all pupils

	All pupils											
	Maths-Diff-in-Diff 2016vs2017			English-Diff-in-Diff 2016vs2017			Maths-Diff-in-Diff 2017vs2018			English-Diff-in-Diff 2017vs2018		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
TP schools	-0.015	0.014	0.281	0.004	0.014	0.807	-0.010	0.009	0.273	0.005	0.010	0.628
Year (2016/17 as base):												
2017/18	-0.045***	0.011	0.000	-0.157***	0.012	0.000						
TP schools # 2017/18	0.002	0.015	0.889	-0.006	0.016	0.711						
Year (2017/18 as base):												
2018/19							-0.063***	0.010	0.000	0.154***	0.011	0.000
TP schools # 2018/19							0.009	0.015	0.559	-0.013	0.015	0.382
Constant	-0.192	1.174	0.870	0.483	1.367	0.724	-0.837	0.682	0.220	-0.726	0.929	0.435
N	382513			380858			407307			406299		
	PP-eligible pupils											
	Maths-Diff-in-Diff 2016vs2017			English-Diff-in-Diff 2016vs2017			Maths-Diff-in-Diff 2017vs2018			English-Diff-in-Diff 2017vs2018		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
TP schools	0.001	0.021	0.970	-0.009	0.022	0.677	-0.027	0.018	0.130	-0.020	0.020	0.313
Year (2016/17 as base):												
2017/18	-0.029	0.018	0.116	-0.145***	0.020	0.000						
TP schools # 2017/18	-0.027	0.026	0.302	-0.014	0.027	0.596						
Year (2017/18 as base):												
2018/19							-0.085***	0.018	0.000	0.093***	0.020	0.000
TP schools # 2018/19							0.036	0.024	0.139	0.031	0.027	0.265
Constant	6.783**	2.146	0.002	1.581	2.503	0.528	3.694*	1.811	0.042	2.174	2.173	0.317
N	99832			99053			105159			104640		

Source: Year 11 population data.

Note: in the heading, the academic year 2016 versus 2017 refers to the comparison between the academic year 2016/17 and 2017/18; 2017 versus 2018 refers to the comparison between the academic year 2017/18 and 2018/19. Controls are listed in Appendix B.

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

Outcomes and analysis

Results of the Year 11 checks

In January 2021, the government announced, in the context of new national restrictions, that exams during the summer term 2021 could not go ahead as planned.[33] As a result of this, the secondary school analysis planned to use the GCSE results awarded by the TAG process in 2021 instead. Before proceeding with the main analysis reported in the following sections, we reviewed the outcomes of the checks before determining that, on balance, we should continue with the analysis on the key research questions on an exploratory basis.

Further information on the checks is summarised here, and in Appendix C for readers interested in technical details.

While deciding how to proceed, and prior to sight of the data, the evaluation team reflected upon a number of considerations about the appropriateness of using TAGs as an outcome measure. These considerations 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 TP 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 tutoring. This consideration is linked to the three prior considerations, 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 considerations. The method and results of the checks are 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 about the presence of any bias. 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.

Outcome analysis

We emphasise that the subsequent findings reported here are exploratory, and should be considered in the context of the outcome of TAGs, which were determined by teachers following the TAG guidance rather than the usual GCSE standardised assessments. Furthermore, there are a number of other caveats to the analysis that are relevant (here, and in the primary school analysis, i.e. not only related to TAGs): the considerable dilution issue as a result of relatively low proportions of PP-eligible pupils doing TP (see '*Pupil selection*' section); the potential negative selection of pupils into the programme; incomplete participation data; the difficulty in selecting the group of pupils who would have received the intervention in comparison schools; and the fact that estimates are for groups of pupils that do not directly align with the group of pupils that participated in may undermine the ability to detect significant improvements of the TP programme.

Outcomes (TAGs)

We compared the TP and comparison schools in terms of the pupil-level and school-level TAGs for the 2021 cohort of Year 11 pupils, which showed that the samples were similar. The results are shown in Table 12 for maths and English, respectively.

[33] 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.

When looking at school-level attainment, TAGs for the 2020/21 cohort of pupils are similar between TP and comparison schools, with no statistically significant differences between the two samples.

Pupil-level TAGs are similarly distributed between TP and comparison schools in the sample of PP-eligible pupils. Maths and English TAGs are significantly lower in TP schools than in comparison schools in the sample of all pupils, by 0.05% and 0.02%, respectively. As the same variables are not significantly different between TP and comparison schools when compared at the school level, school size may drive the difference at pupil level. We weighted each pupil by 1 over the number of pupils in the schools to assign each school the same weight and compared again TAGs between TP and comparison schools. The significant differences disappeared, except for maths TAGs in the population of all pupils (Table 12).

Compared with the national average, TP and comparison schools have similar TAGs scores to the national population in both the PP-eligible pupils and all pupil specifications.

Sets of histograms provide a graphical inspection of the distributions of the TAGs in the analysed TP and comparison schools. Histograms of these scores for all pupils and PP-eligible pupils can be found in Figure 4 (maths) and Figure 5 (English) for TAGs. The histograms support the finding from the comparison of average values of TAGs between TP and matched comparison schools reported above. Maths and English TAGs are similarly distributed for PP-eligible pupils and all pupils between TP and comparison schools.

Table 12: TAGs for 2021 cohort of Year 11s in TP schools, matched comparison schools, national data, PP-eligible pupils, and all pupils

All pupils								
Variable	National averages	Means: Comparison	SD: Comparison	Means: TP schools	SD: TP schools	Difference	Difference TP and comparison, weighted	Difference TP and comparison, weighted and std
KS4 English, school level	5.173	5.128	(0.785)	5.119	(0.718)	-0.009		
KS4 Maths, school level	4.997	4.949	(0.838)	4.917	(0.734)	-0.032		
KS4 English, pupil level	5.207	5.262	(1.869)	5.245	(1.846)	-0.017**	-0.015	-0.007
KS4 Maths, pupil level	5.038	5.083	(2.037)	5.035	(2.018)	-0.047***	-0.042***	-0.016
Observations		119,682		121,821		241,503	241,145	241,145
PP-eligible pupils								
Variable	National averages	Means:Comparison	SD:Comparison	Means:TP schools	SD:TP schools	Difference		
KS4 English, school level	4.256	4.579	(0.842)	4.579	(0.751)	0.000		
KS4 Maths, school level	4.531	4.298	(0.882)	4.265	(0.778)	-0.032		
KS4 English	4.380	4.503	(1.871)	4.522	(1.831)	0.019	0.004	0.007
KS4 Maths	4.010	4.215	(1.988)	4.219	(1.984)	0.004	-0.026	0.001
Observations		30,508		31,516		62,024		

Source: Year 11 population data.

Figure 4: Histograms of the distributions of Maths and English TAGs, all pupils, in TP schools and matched comparison (non-TP) schools

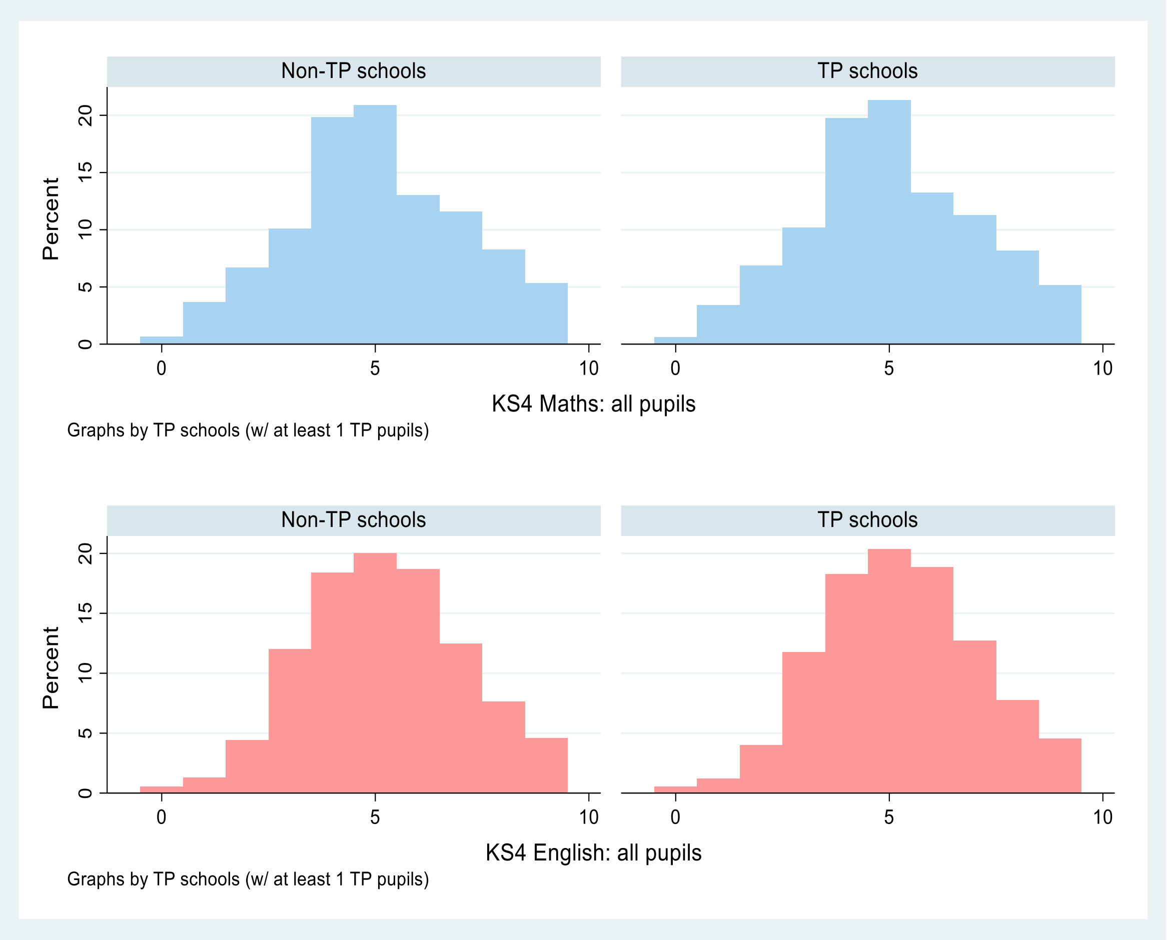
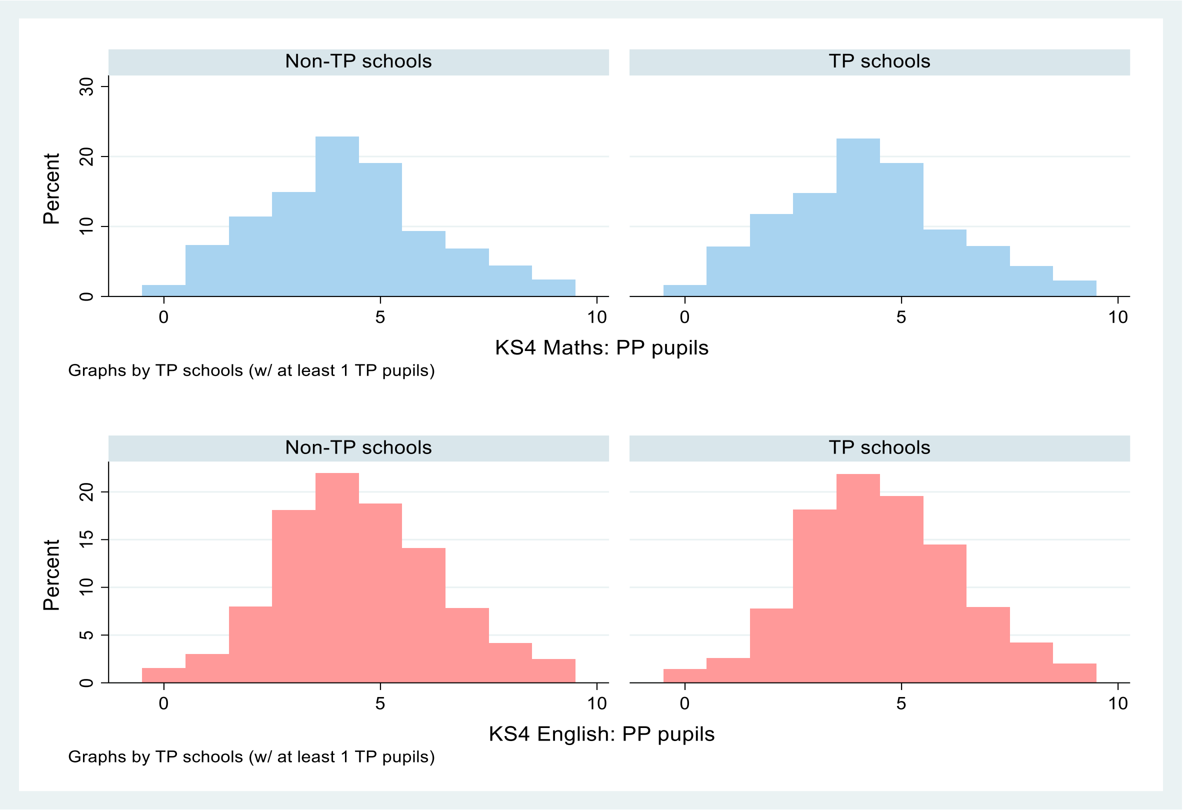


Figure 5: Histograms of the distributions of maths and English TAGs, PP-eligible pupils, in TP schools and matched comparison (non-TP) schools



Outcome analysis

Regression (RQ4a1): What is the impact of TP availability on all PP-eligible pupils' attainment?

We present the results of the two measures of TP on PP-eligible pupils below: i) a 0/1 indicator for TP being available at the school level; and ii) a categorical variable measuring the fraction of hours completed on average at the school level by the time of the assessment (dosage).

Results in Table 13 show the impact of TP tutoring in maths and in English on PP-eligible pupils under the assumption that observable characteristics control for selection into the TP programme. TP is measured with a dummy equal to one if TP is available at school level. For English the coefficients are zero and the effect size of TP is 0.000 (CI: -0.021 to 0.025) (see Table 24 for effect size table),[34] ruling out large effects of the intervention. For maths, the coefficients are negative and significant at the 5% level when we apply the Romano-Wolf correction. The effect size for maths is -0.019 (CI: -0.038 to 0.001) (see Table 24). For both subjects this is the equivalent of zero additional months progress. The specification presenting the Romano-Wolf correction for multiple testing indicates that the coefficient of TP on maths is negative and only just statistically significant at the 5% level.[35] Removing inverse probability weights provides results with the same level of significance.

Table 13: Impact of TP measured with a 0/1 dummy indicating the availability of TP 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
TP schools	-0.037	0.020	0.063	0.004	0.022	0.856	-0.035*	0.020	0.030	-0.001	0.021	0.950
Constant	-2.066	3.576	0.564	-4.410	3.692	0.232	-3.484	3.563	0.328	-5.030	3.703	0.175
N	62024			62024			62024			62024		

Source: Year 11 population data, number of schools: 1,464.

Note: School-level clustered residuals and inverse probability weighting. Controls are listed in Appendix B. Column 2 reports the p-value associated with the Romano-Wolf correction.

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

The dosage analysis in Table 14 presents the impact of TP when measured with dosage, computed at school level and with respect to TP intervention schools with zero dosage. We have not controlled for selection into dosage category and therefore, we cannot view the dosage relationship as causal.

For both subjects, the findings suggest that all levels of school-level dosage are not correlated with TAGs, with the coefficients being close to zero and thus ruling out large effects of the intervention (Table 14). Results are different from those presented in Table 18, where dosage is computed at pupil level and measured as a continuous variable in the sample of TP schools only, and where the coefficient of dosage is positive and significant. The aggregation of dosage at school level, and the inclusion of matched comparison schools in the sample may explain the non-significant effect here. The histograms in Figures 2a, 2b, and 3 show that the distribution of dosage is different at the school level and at the pupil level, which contributes to explaining the different results in the school-level and pupil-level specifications. Findings in Table 19a suggests that higher levels of pupil-level dosage are correlated with better outcomes in the sample of TP schools and that aggregating the dosage data at school level reduces the pupil-level variability that would capture this effect change.

[34] Effect sizes are computed for the specifications without the Romano-Wolf correction as the latter provides only corrected p-values and not corrected CIs.

[35] This correction is considerably more powerful than earlier multiple testing procedures such as the Bonferroni and Holm corrections, given that it takes into account the dependence structure of the test statistics by resampling from the original data (by assuming a 'worst-case' dependence structure among the p-values, which is 'close' to the individual p-values being independent of each other). This feature, together with the stepwise nature of the procedure, results in improved ability to correctly reject false null hypotheses compared to more traditional multiple testing procedures, such as the Bonferroni procedure and the Holm procedure.

Table 14: Impact of TP measured with dosage on the population of PP-eligible pupils

	Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value
Comparison schools	0.054	0.034	0.109			
<=1	0.014	0.035	0.676			
>1	0.048	0.042	0.255			
Comparison schools				0.015	0.034	0.654
<=1				0.033	0.036	0.365
>1				0.005	0.045	0.917
Constant	-2.089	3.592	0.561	-4.472	3.697	0.227
N	62024			62024		

Source: Year 11 population data, number of schools: 1,464.

Note: School-level clustered residuals and inverse probability weighting. Dosage measured as less than one block of 12 hours, or more than one block (more than 12 hours of tutoring). Controls are listed in Appendix B. Baseline category is represented by TP intervention schools with zero dosage. Column 2 reports the p-value associated with the Romano-Wolf correction.

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

Dilution is even problematic in the sample of PP-eligible pupils, as only 26% of PP-eligible pupils was selected for TP on average (see Table 8a), and only around 10% of pupils were selected for TP in the specific subject (English or maths). We present the results of an additional analysis at the end of the '*Outcomes*' section, which restrict the sample of TP schools to those that targeted at least 50% and 70% of PP-eligible pupils (see the '*Additional analysis*' section).

Instrumental Variables (RQ1b): What is the impact of TP on the attainment of pupils participating due to encouragement to do so?

The IV analysis was intended to be conducted on the sample of PP-eligible pupils in TP schools only. However, as outlined below, this analysis could not proceed.

This analysis aimed to use the encouragement to participate in tutoring (or not) as an instrument as we hypothesised that it may be positively associated with dosage (amount) of tutoring. The intervention leveraged similarities between pupils and tutors. Before proceeding with the analysis, we ran the first stage and tested for weak instrument using the Montiel Olea-Pflueger (2013) approach, to check that the encouragement trial induced sufficient change in TP participation. We estimate a regression of TP dosage on a dummy equal to one if the tutor was randomised to receive the encouragement intervention,[36] zero otherwise. Fewer than 20 pupils had been tutored by both tutors who were in the treatment arms and other tutors in the control arms. We removed these pupils from this analysis.

The procedure tests the null hypothesis that the estimator's approximate asymptotic bias exceeds a fraction τ of a 'worst-case' benchmark. The results of the weak instrument test are reported in Table 15. For both English and maths, the F statistic is less than the worst-case benchmark (37.4 with τ = 5%), indicating that the instrument is weak, i.e. not strongly correlated with the amount of dosage. Hence, we did not proceed with this part of the analysis.

[36] Study plan: https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/NTP-RCT1-%E2%80%94-Leveraging-Similarity-to-Improve-Pupil-Attendance-Protocol_SAP_2021-09-06-104614_kdqt.pdf?v=1630925174

Table 15: Weak instrument test for tutor encouragement trial

Maths			English		
Montiel-Pflueger robust			Montiel-Pflueger robust		
Effective F statistic:	0.004		Effective F statistic:	0.01	
Confidence level alpha:	5%		Confidence level alpha:	5%	
Critical values	TSLS	LIML	Critical values	TSLS	LIML
% of worst-case bias			% of worst-case bias		
ʋ =5%	37.418	37.418	ʋ =5%	37.418	37.418
ʋ =10%	23.109	23.109	ʋ =10%	23.109	23.109
ʋ =20%	15.062	15.062	ʋ =20%	15.062	15.062
ʋ =30%	12.039	12.039	ʋ =30%	12.039	12.039

Source: Year 11 population data, schools N=732.

Instrumental Variables (RQ4b): What is the impact of the intensity of TP (dosage) on the attainment of all PP-eligible pupils?

The IV analysis (in TP intervention schools only) aimed to use the date of signing up to the programme as an instrument as we hypothesised that it may be positively associated with dosage (amount) of tutoring by the date of the endpoint assessment. We ran two estimates: the first, as specified in the study plan, was estimated using all PP-eligible pupils in year groups doing TP. In addition to this and not included in the study plan, we also estimated it on all TP pupils in TP schools, regardless of their Pupil Premium status. However, as outlined below, this analysis could not proceed.

Before proceeding with the analysis, we ran the first stage and tested for weak instrument using the Montiel Olea-Pflueger (2013) approach. The procedure tests the null hypothesis that the estimator's approximate asymptotic bias exceeds a fraction τ of a 'worst-case' benchmark. The results of the weak instrument test are reported in Table 16. For both English and maths, and for both PP-eligible pupils and all TP pupils, the F statistic is lower than the worst-case benchmark (37.4 with ʋ = 5%), not rejecting the null hypothesis of a weak instrument, i.e. not strongly correlated with the amount of dosage. Hence, we did not proceed with this part of the analysis.

Table 16: Weak instrument test for dosage analysis: PP-eligible pupils and all TP pupils

	Maths Montiel-Pflueger robust			English Montiel-Pflueger robust		
PP-eligible pupils	Effective F statistic:	0.506		Effective F statistic:	0.508	
PP-eligible pupils	Confidence level alpha:	5%		Confidence level alpha:	5%	
All TP pupils	Effective F statistic:	0.427		Effective F statistic:	0.426	
All TP pupils	Confidence level alpha:	5%		Confidence level alpha:	5%	
	Critical values	TSLS	LIML	Critical values	TSLS	LIML
	% of worst-case bias			% of worst-case bias		
	ʋ =5%	37.418	37.418	ʋ =5%	37.418	37.418
	ʋ =10%	23.109	23.109	ʋ =10%	23.109	23.109
	ʋ =20%	15.062	15.062	ʋ =20%	15.062	15.062
	ʋ =30%	12.039	12.039	ʋ =30%	12.039	12.039

Source: Year 11 TP schools, schools N=732.

Further analyses

RQ4a2: What is the impact of TP availability on the attainment of pupils predicted to participate?

We aimed to predict which pupils would have participated in the TP programme and estimate the impact of TP on these pupils compared with similar pupils in comparison schools. However, as outlined below, this analysis could not proceed due to the poor predictive power of the model.

Before proceeding with RQ4a2, we estimated the participation equation on all pupils in TP schools to assess its predictive power. Results in Table C7 (see the '*Ex post 3 check*' section in Appendix C), show the pupil-level participation equations for TP maths and TP English, estimated with a logit model and a linear probability model. The numbers in Table C7 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 TP.

We note that this reflects the findings in the IPE report, which indicated that schools selected pupils to participate in TP based on a number of variables that are not observable to us in the dataset, for example motivation or their ability to catch up and make good use of tutoring.

RQ4a3: What is the impact of the availability of TP on all pupils' attainment in the population of schools?

Similar to the approach for RQ4a1 (analysis on PP-eligible pupils) we present the results of the two measures of TP - this time on all pupils- below: i) a 0/1 indicator for TP being available at the school level; and ii) a categorical variable measuring the fraction of hours completed by the time of the assessment (dosage) again at the school level.

Results in Table 17 show that the impact of the availability of TP on English on the population of all pupils is not different from zero and not statistically significant. The effect size for English is -0.010 (-0.027 to 0.006) (see Table 24). For maths, the coefficient is negative and significant with the size of the coefficient close to zero (-0.04). The effect size is -0.020 (CI: -0.033 to -0.007) (see Table 24). For both subjects this is the equivalent of zero additional months progress. This result is consistent with the descriptive statistics reported in Table 7b, where maths TAGs were significantly lower in the sample of TP schools, even after adjusting the estimates by school size. Applying the Romano-Wolf multiple hypothesis testing correction does not change the results. Removing inverse probability weights provides results with the same level of significance.

Table 17: Impact of TP measured with a 0/1 dummy indicating the availability of TP 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
TP schools	-0.040**	0.014	0.003	-0.02	0.016	0.209	-0.040**	0.014	0.0099	-0.019	0.016	0.0891
Constant	-9.038***	2.269	0.000	-8.807***	2.605	0.001	-10.165***	2.288	0.000	-9.772***	2.640	0.000
N	241503			241503			241503			241503		

Source: Year 11 population data, schools N=1,464.

Note: School-level clustered residuals and inverse probability weighting. Controls are listed in Appendix B. Column 2 reports the p-value associated with the RW correction.

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

The dosage analysis in Table 18 presents the impact of TP on all pupils when measured with dosage, computed at school level. Dosage was divided into categories indicating the average number of blocks of hours completed (less than one block of 12 hours, or more than one block, more than 12 hours of tutoring) with respect to TP schools with zero level of dosage (see Figures 2a, 2b, and 3 in *RQ4a1* for histograms of the distribution of dosage).

For both maths and English, the findings indicate that levels of dosage are not associated with statistically different attainment when looking at the impact on all pupils, except for comparison schools having slightly better maths scores than TP schools with zero dosage.

The results on the full sample of the population have the disadvantage of considering all pupils, which includes all TP pupils whether they are PP-eligible pupils or not. Dilution is even more problematic in the sample of all pupils, as only 16.7% of pupils were selected for TP on average (see Table 8a). It is therefore, harder to detect impacts of TP when looking at all pupils because the estimated impact on all pupils is seriously diluted by the attainment of the 83% of pupils in the analysed group that were not selected to receive tutoring.

Table 18: Impact of TP measured with dosage on all pupils

	Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value
Comparison schools	0.059**	0.021	0.006			
<=1	0.020	0.023	0.388			
>1	0.044	0.029	0.127			
Comparison schools				0.003	0.024	0.885
<=1				-0.015	0.026	0.554
>1				-0.052	0.035	0.136
Constant	-9.094***	2.272	0.000	-8.807***	2.605	0.001
N	241503			241503		

Source: Year 11 population data, number of schools: 1,464.

Note: School-level clustered residuals and inverse probability weighting. Dosage measured as less than one block of 12 hours, or more than one block (more than 12 hours of tutoring). Controls are listed in Appendix B. Baseline category is represented by TP intervention schools with zero dosage. Column 2 reports the p-value associated with the Romano-Wolf correction.

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

Moderator analysis:

RQ5: How does the association of TP availability with attainment vary among PP-eligible pupils, by school and pupil characteristics?

Tables 19a and 19b present the results of the interaction between TP schools and a set of school level, pupil level, and geographic characteristics on PP-eligible pupils' TAGs. Results are separately produced for maths (Table 19a) and English. (Table 19b). The tables report the margins to facilitate the interpretation of the interactions. These are calculated from predictions of a fit model at fixed values of some covariates and averaging over the remaining covariates.

There are several caveats to this analysis. First, results are not causal. Second, the majority of PP-eligible pupils were not selected to 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 research question.

None of the interactions is significant for English TAGs in the estimation of marginal effects. For maths TAGs, most interactions were also non-significant. There were two exceptions, which are outlined below.

- Students with KS2 below 100 are associated with slightly lower maths TAGs in TP schools compared to comparison schools (-0.09 lower grades for KS2 lower than 90, -0.06 for KS2 lower than 95, and -0.04 for KS2 lower than 100).

- TP schools in urban areas are associated with slightly lower maths TAGs compared to comparison schools in urban areas (-0.05).

All of these findings need to be considered in the context of the dilution issue, meaning that the effects reported here may not be specifically related to the availability of TP in the school but may instead be a feature of other activities or characteristics of the school. As with the findings reported for the other research questions, not all of the PP-eligible pupils in the analysis participated in TP in the TP schools (overall only 25.6% of PP-eligible pupils participated in TP [Table 8a]).

Table 19a: How does the impact of TP availability on maths vary among PP-eligible pupils, by school and pupil characteristics?

	Interact w/ school-level			Interact w/ pupil-level			Interact w/ KS2			Interact w/ geography		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
Ofsted High vs Low:												
Low	-0.044	0.049	0.363									
High	-0.032	0.022	0.150									
Ofsted Missing	-0.045	0.113	0.689									
School %FSM High vs Low:												
School %FSM high vs low=0	-0.045	0.025	0.073									
School %FSM high vs low=1	-0.017	0.034	0.605									
Maintained vs Non-maintained schools:												
Academy/Free school	-0.025	0.023	0.278									
Maintained school	-0.065	0.040	0.103									
Pupil counts Q4:												
Q1	-0.327	0.198	0.099									
Q2	0.004	0.043	0.917									
Q3	-0.057	0.032	0.077									
Q4	-0.017	0.030	0.569									
school size missing	-0.334	0.178	0.061									
Female=0				-0.033	0.026	0.197						
Female=1				-0.036	0.025	0.151						
Non-SEN				-0.037	0.021	0.079						
SEN				-0.022	0.039	0.568						
Non-EAL				-0.035	0.021	0.100						
EAL				-0.032	0.048	0.503						
EAL unknown				-0.062	0.201	0.759						
White British				-0.030	0.025	0.231						
Asian				-0.088	0.054	0.102						
Black				-0.004	0.059	0.948						
Other ethnic				-0.051	0.043	0.235						
Unknown ethnic				0.058	0.105	0.582						
KS2 Maths scores 90-115:												
90							-0.087**	0.031	0.006			
95							-0.064**	0.023	0.006			
100							-0.041*	0.020	0.041			
105							-0.018	0.023	0.436			
110							0.005	0.031	0.876			
115							0.028	0.041	0.497			
Rural vs Urban:												
urban										-0.048*	0.021	0.025
rural										0.055	0.062	0.377
IDACI scores:												
Low										-0.040	0.030	0.187
High										-0.033	0.023	0.164
IDACI Missing										-0.133	0.309	0.666
N	62024			62024			62024			60502		

Source: Year 11 population data, schools N=1,464.

Note: School-level clustered residuals and inverse probability weighting. Controls are listed in Appendix B. The table presents estimated margins, i.e. the marginal effect of each interaction, estimated using counterfactual analysis. The coef is the difference in TAGs compared to pupils with the same characteristics in comparison schools. Female = 0 compares males in both groups.

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

Table 19b: How does the impact of TP availability on English vary among PP-eligible pupils, by school and pupil characteristics?

	Interact w/ school-level			Interact w/ pupil-level			Interact w/ KS2			Interact w/ geography		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
Ofsted High vs Low:												
Low	0.061	0.053	0.249									
High	-0.009	0.024	0.704									
Ofsted Missing	0.029	0.112	0.798									
School %FSM High vs Low:												
School %FSM high vs low=0	0.015	0.028	0.590									
School %FSM high vs low=1	-0.010	0.036	0.779									
Maintained vs Non-maintained schools:												
Academy/Free school	0.025	0.025	0.312									
Maintained school	-0.055	0.040	0.170									
Pupil counts Q4:												
Q1	0.036	0.173	0.834									
Q2	-0.028	0.047	0.550									
Q3	0.010	0.034	0.768									
Q4	0.021	0.033	0.523									
school size missing	0.161	0.137	0.241									
Female=0				0.004	0.027	0.889						
Female=1				0.008	0.026	0.754						
Non-SEN				0.009	0.023	0.693						
SEN				-0.008	0.042	0.857						
Non-EAL				0.005	0.023	0.819						
EAL				0.011	0.050	0.829						
EAL unknown				-0.017	0.240	0.943						
White British				-0.001	0.027	0.975						
Asian				0.027	0.052	0.609						
Black				0.023	0.057	0.683						
Other ethnic				-0.008	0.046	0.866						
Unknown ethnic				0.127	0.123	0.302						
KS2 Maths scores 90-115:												
90							-0.019	0.029	0.515			
95							-0.007	0.023	0.759			
100							0.005	0.021	0.828			
105							0.016	0.025	0.507			
110							0.028	0.031	0.371			
115							0.040	0.040	0.318			
Rural vs Urban:												
urban										-0.009	0.022	0.700
rural										0.099	0.076	0.195
IDACI scores:												
Low										0.003	0.032	0.913
High										0.006	0.025	0.800
IDACI Missing										-0.339	0.301	0.260
N	62024			62024			62024			60502		

Source: Year 11 population data, schools N=1,464.

Note: School-level clustered residuals and inverse probability weighting. The table presents estimated margins, i.e. the marginal effect of each interaction, estimated using counterfactual analysis. The coef is the difference in TAGs compared to pupils with the same characteristics in comparison schools. Controls are listed in Appendix B.

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

RQ6: How do outcomes vary among TP pupils, by model of tutoring?

This analysis is based only on the TP pupils in TP schools, as here we are exploring differences in delivery (in contrast to previous research questions that compare TP schools with the comparison schools). Consequently the results reported here should be considered as purely descriptive.

In this analysis, each observation is identified as pupil-subject session. Hence, pupils who received more than one tutoring session appear in the data for each subject-specific session. Residuals are clustered: at the pupil level in the first and second columns that explore pupil-level TP characteristics on maths and English TAGs; at the school level in the third and fourth where we list school-level TP characteristics; and at the tutor level in the fifth and sixth columns where tutor-level characteristics are explored. We control for the same set of controls used in previous specifications. These results are subject-specific, so maths TAG outcomes for pupils that were selected for TP in maths (and the equivalent for English).

The correlations are shown in Table 20a and 20b. The key findings are summarised below:

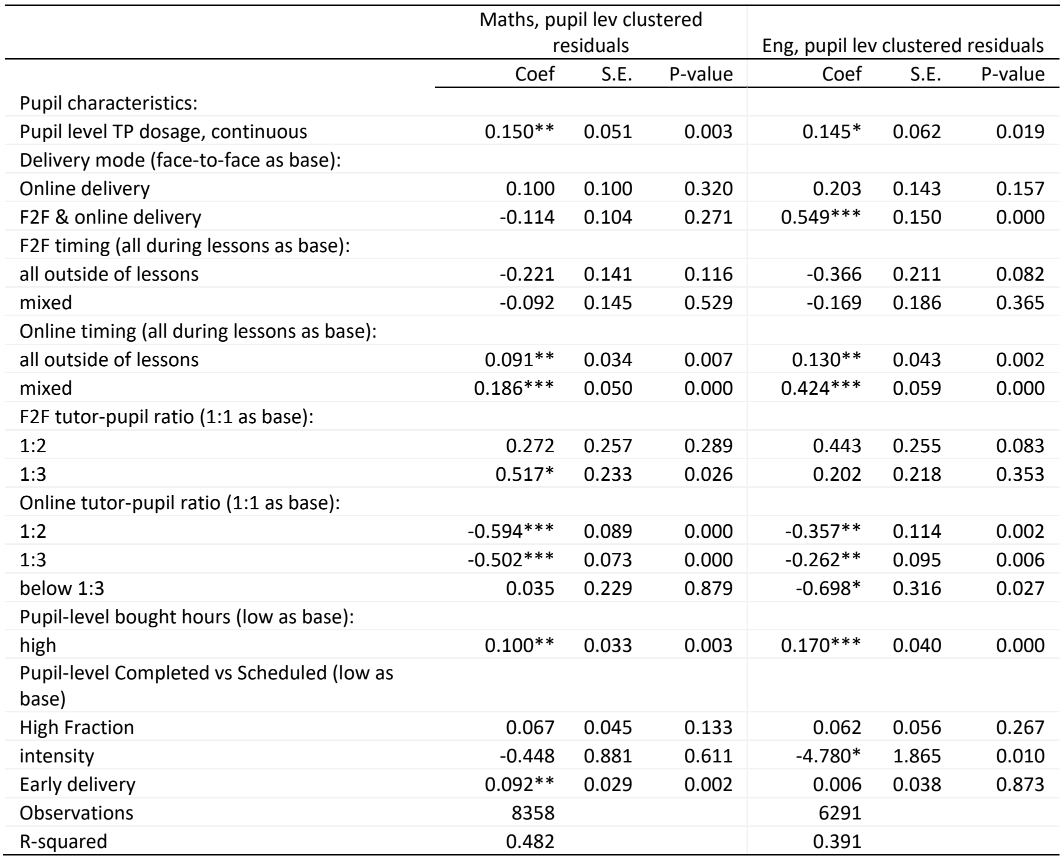
- Higher dosage (higher numbers of sessions completed) is significantly correlated with higher English and maths TAGs. Pupil-level dosage is measured as a continuous variable to avoid capturing non-linearities associated with the definition of dosage in blocks.
- A mix of online and face-to-face sessions are associated with better English TAGs than only face-to-face sessions in English. Similar analysis for maths showed no significant difference.
- The timing of face-to-face sessions, whether during or outside schooling hours, is not associated with significantly different TAGs in either subject.
- Online sessions delivered outside schooling hours or mixed (a combination of sessions taken within schooling hours and outside schooling hours) are associated with better maths and English TAGs than online sessions delivered within schooling hours.
- Attending face-to-face sessions with two other pupils is associated with higher maths TAGs than attending a session alone.
- The association is in the opposite direction for online sessions for both maths and English: online sessions are associated with better TAGs if attended alone. For maths, online 1:1 sessions are associated with better TAGs than ratios of 1:2 or 1:3. For English, there is a similar effect: online sessions in a ratio of 1:1 are associated with better TAGs than online sessions of any larger group size.
- Delivery of tutoring sessions concentrated over a short time frame[37] is negatively correlated with English TAGs suggesting that higher scores in English are associated with tutoring that is more spread out. There is no difference for maths.
- Early delivery (i.e. tutoring sessions delivered earlier in the academic year[38]) are associated with better TAGs in maths. There is no difference for English.
- The number of sessions bought for a pupil, when computed both at pupil and school level, are positively correlated with maths and English TAGs.

When looking at tutor's characteristics and the tutor's highest qualification, it appears that having specialised postgraduate qualifications (rather than undergraduate or Qualified Teacher Status [QTS]) may deliver better tutoring to Year 11 pupils (we selected postgraduate degree as the base as it is the highest possible qualification):

[37] Concentration of delivery was computed as follows. First, we computed dosage as the sum of sessions completed online and/or face-to-face divided by 12, and then we divided it by the time passed between the first and the last date of the sessions. The cut-off point for intensity is 0.3 (i.e. more than 15 hours in 4 days or 28 hours in 7 days), with values above that considered too high and erroneous and replaced with missings.
[38] Early delivery is computed with respect to the median of the time passed between the first and the last session taken at school level. It measures the change in TAGs associated with delivering the sessions earlier in the academic year.

- Having tutors with a postgraduate degree is associated with higher performance in English than having tutors with an undergraduate degree or with QTS. Having tutors with a postgraduate degree is associated with higher performance in maths than having tutors with QTS.

Table 20a: Estimates of KS4 maths and English grades by model of tutoring, among TP pupils, and pupil-level TP characteristics



	Maths, pupil lev clustered residuals			Eng, pupil lev clustered residuals		
	Coef	S.E.	P-value	Coef	S.E.	P-value
Pupil characteristics:						
Pupil level TP dosage, continuous	0.150**	0.051	0.003	0.145*	0.062	0.019
Delivery mode (face-to-face as base):						
Online delivery	0.100	0.100	0.320	0.203	0.143	0.157
F2F & online delivery	-0.114	0.104	0.271	0.549***	0.150	0.000
F2F timing (all during lessons as base):						
all outside of lessons	-0.221	0.141	0.116	-0.366	0.211	0.082
mixed	-0.092	0.145	0.529	-0.169	0.186	0.365
Online timing (all during lessons as base):						
all outside of lessons	0.091**	0.034	0.007	0.130**	0.043	0.002
mixed	0.186***	0.050	0.000	0.424***	0.059	0.000
F2F tutor-pupil ratio (1:1 as base):						
1:2	0.272	0.257	0.289	0.443	0.255	0.083
1:3	0.517*	0.233	0.026	0.202	0.218	0.353
Online tutor-pupil ratio (1:1 as base):						
1:2	-0.594***	0.089	0.000	-0.357**	0.114	0.002
1:3	-0.502***	0.073	0.000	-0.262**	0.095	0.006
below 1:3	0.035	0.229	0.879	-0.698*	0.316	0.027
Pupil-level bought hours (low as base):						
high	0.100**	0.033	0.003	0.170***	0.040	0.000
Pupil-level Completed vs Scheduled (low as base)						
High Fraction	0.067	0.045	0.133	0.062	0.056	0.267
intensity	-0.448	0.881	0.611	-4.780*	1.865	0.010
Early delivery	0.092**	0.029	0.002	0.006	0.038	0.873
Observations	8358			6291		
R-squared	0.482			0.391		

Source: Year 11 population data, schools: N=555 for maths and N=537 for English.

Inverse probability weighting. Controls are listed in Appendix B. Missing categories controlled for but not included in the table.

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

Table 20b: Estimates of KS4 maths and English grades by model of tutoring, among TP pupils, and school-level and tutor-level TP characteristics

	Maths, school lev clustered residuals			Eng, school lev clustered residuals			Maths, tutor lev clustered residuals			Eng, tutor lev clustered residuals		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
School-level dosage (zero as base):												
<=1	0.995	0.647	0.125	1.131	0.772	0.143						
>1	1.096	0.647	0.091	1.144	0.773	0.139						
School-level bought hours (low as base):												
High buy-in	0.121**	0.046	0.009	0.182**	0.055	0.001						
Tutor qualifications (Postgrad as base):												
Undergraduate							-0.034	0.054	0.523	-0.206**	0.064	0.001
QTS							-0.243**	0.092	0.008	-0.244*	0.111	0.029
PGCE							-0.037	0.128	0.771	-0.147	0.134	0.273
A-levels							-0.016	0.053	0.770	-0.123	0.064	0.057
GCSE							0.348	0.692	0.616	-0.085	0.243	0.727
Tutor training (yes as base):												
No							0.071	0.225	0.752	0.061	0.359	0.865
On-going							-0.156	0.289	0.589	-0.383	0.234	0.102
Tutor: same ethnicity (no as base)												
Same							0.111	0.111	0.316	0.053	0.166	0.749
Tutor: same gender (no as base)												
Same							-0.019	0.054	0.725	0.022	0.085	0.795
Observations	8358			6291			7421			5623		
R-squared	0.469			0.376			0.469			0.376		

Source: Year 11 population data, schools: N=555 for maths and N=537 for English in the school-level specification and N=545 and N=525 in the tutor-level regression.

Inverse probability weighting. Controls are listed in Appendix B. Missing categories controlled for but not included in the table.

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

Missing data analysis

The missing data analysis indicates that few pupils are dropped from the analysis because of missing TAGs or missing KS2 maths and English scores. There are 2,789 pupils with missing KS2 and 237 pupils with missing KS4 in TP schools and 2,539 pupils with missing KS2 and 275 pupils with missing KS4 in comparison schools.The majority of missing is concentrated in KS2 assessment, which occurred years before the intervention, and hence it should not lead to any bias. As missingness was low and it is unlikely to lead to bias (coupled with the length of time required to run multiple imputation on the Year 11 data in the SRS), we decided to exclude from the analysis pupils with missing primary baseline and/or outcome.

Additional analyses

In this section we present some additional analyses to further explore the findings reported above.

As reported earlier, in the analysis on PP-eligible pupils (RQ4a1) only 26% of PP-eligible pupils received TP. To explore whether the results reported above hold when the analysed group (PP-eligible pupils) is more aligned with the treated group (TP) we carried out two additional analyses that restrict the sample of TP schools to those that delivered the intervention to least 50% and 70% of Year 11 PP-eligible pupils on RQ4a1. We cannot select comparison schools on the basis of the proportion of PP-eligible pupils who would have received the intervention. It is a level of selection we cannot control for, although it is correlated with the proportion of PP-eligible pupils in the school. We therefore produced new matched samples for the 50% and 70% specifications. Estimates are based on OLS. It is worth noticing that this analysis is not addressing the issue of pupils' selection in the subject.

We first selected TP intervention schools that targeted 50% (N=191 out of 732) and 70% (N=63 out of 732) of PP-eligible pupils in Year 11. Next, we defined a sample of comparison schools for each group using propensity score matching and using the same set of variables listed in Appendix B. Table D1 and D2 in Appendix D show the balance tables before matching (between the TP restricted samples and the sample of eligible comparison schools) and Tables D3 and D4 show the balance tables after matching (between the restricted TP samples and the matched comparison schools). These tables show that matching improves the quality of the match by providing a sample of comparison schools with less significant differences to TP schools with respect to those shown in Tables D1 and D2.

In the matched sample of schools that selected at least 50% of Year 11 PP-eligible pupils to do TP (Table D3), there are no significant differences between TP and matched comparison schools in all listed characteristics, school-level KS2 and TAGs, except for AM participation, by construction, and for 'Good' Ofsted rating, slightly higher in comparison schools. Comparison with the national population indicates that measures of disadvantage and ethnicity are similarly distributed between the TP and comparison samples and the national population. The only notable difference is the fraction of Academy schools, higher in the national population (23% vs. 16% in the TP and comparison samples).

In the matched sample of schools that selected at least 70% of PP-eligible pupils to do TP (Table D4), there are no significant differences between TP and matched comparison schools across all characteristics, school-level KS2 and TAGs, but it has to be noted that some difference had to be suppressed because of small counts. The tables indicate that the matching process provided a balanced group of comparison schools in both samples. Comparison with the national population (see Table 7a) show that measures of disadvantage and ethnicity are similarly distributed between the TP and comparison samples and the national population. The notable differences are: the fraction of 'Outstanding' schools, which is higher in the TP and comparison sample (35% and 33% vs. 21% in the national population), and the regional distribution, with more schools from the East of England and London and fewer schools from the South East in the TP and comparison samples compared to the national population. Finally, the percentage of pupils eligible for FSM is lower in the TP and comparison samples than in the national population (21% in the TP and comparison samples vs. 28% in the national population).

We then estimated RQ4a1 on these selected samples of schools. Results are reported in Table 21 (50% restriction) and Table 22 (70% restriction), for both maths and English and for the sample of PP-eligible pupils.

In the specification selecting 50% of PP-eligible pupils for TP, the impact of TP on maths and English is non-significant. The coefficients shown in Table 18 are negative for both subjects and both subjects have a negative effect size (-0.009 for maths and -0.027 for English: see Table 22), the equivalent of no additional months progress.

Table 21: Regression (RQ4a1): What is the association of TP availability with PP-eligible pupils' attainment using the sample of TP schools that targeted at least 50% of PP-eligible pupils for tuition

	Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p -value
TP schools	-0.018	0.036	0.615	-0.049	0.039	0.214
Constant	-1.821	7.085	0.797	-7.826	6.169	0.205
N	12651			12651		

Source: Year 11 population data, number of schools: 382.

School-level clustered residuals and inverse probability weighting. Controls are listed in Appendix B.

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

In the specification selecting 70% of PP-eligible pupils for TP, the impact of TP on maths and English is positive and significant (Table 22). Schools participating in the TP programme are associated with increases in maths and English TAGs by 0.19 for maths and 0.22 for English. This converts to an effect size of 0.098 for maths and 0.117 for English (Table 24), which equate to 2 months additional progress in both maths and in English. The evidence suggests that, when the sample selection reduces the dilution problem by bringing the analysed group closer to the group that were selected for the intervention, that TP has a positive association with PP-eligible pupils' performance in TAGs compared to PP-eligible pupils in similar schools. Although this reduces the dilution problem, it should be borne in mind that these are a relatively small number of schools participating in TP and different from the TP population in several respects (e.g. more 'Outstanding' schools, lower percentage of FSM pupils). The association can be interpreted as causal if the ‘selection on observables’ assumption is upheld (i.e. if all variables influencing selection of schools to the treatment and the outcome variable is controlled for in the regression).

Table 22: Regression (RQ4a1): What is the impact of TP availability on PP-eligible pupils’ attainment using the sample of TP schools that targeted at least 70% of PP-eligible pupils for tuition?

	Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value
TP schools	0.189*	0.083	0.024	0.218**	0.067	0.001
Constant	-8.154	14.598	0.577	-24.171*	10.808	0.027
N	3657			3657		

Source: Year 11 population data, number of schools: 126.

School-level clustered residuals and inverse probability weighting. Controls are listed in Appendix B.

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

As a third piece of additional analysis, we estimated RQ4a1 (PP-eligible pupils) using a difference-in-difference model and comparing pupil-level outcomes in the academic year 2020/21 in TP schools with the academic outcomes of Year 11 pupils in the same school in the previous academic year (2019/20) (Table 23), for both maths and English and on PP-eligible pupils (first two columns) and all pupils (last two columns). The model uses the same set of controls as in the previous estimates. In all specifications, the interaction between the 2020/21 time dummy and TP schools is close to zero and not significantly different from zero, suggesting that TP did not induce any significant change over time in TP schools.[39] The results are consistent with the ones reported in RQ4a1 and RQ4a3.

It has to be noted that this analysis is performed by using pupil-level data as cross sections rather than as longitudinal, as usually done in this type of analysis. Longitudinal analysis is not possible because different cohorts of pupils are tested in each academic year.

[39] The coefficients of the time dummy are negative and significant, which is at odds with the results of the Year 11 checks reported in Table C2 in Appendix C. The difference is explained by the fact that the results in Table C2 are based on a sample of school-level data, while the ones in Table 20 are based on pupil-level data. In addition, when we control for the time dummy only (and not the other set of controls), the coefficient turns positive and significant in three specifications out of four. It turns non-significant and close to zero in the specification with PP-eligible pupils and maths outcome.

Table 23: Regression: What is the impact of TP availability on PP-eligible pupils' and all pupils' attainment using the Diff-in-Diff with respect to attainment in the previous academic year (2019/20)?

	Maths Diff-in-Diff, PP-eligible pupils			English Diff-in-Diff, PP-eligible pupils			Maths Diff-in-Diff, all			English Diff-in-Diff, all		
	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value	Coef	S.E.	P-value
TP schools # 2020/21	-0.008	0.025	0.752	0.047	0.027	0.086	-0.011	0.015	0.464	0.001	0.017	0.969
TP schools	-0.012	0.015	0.443	-0.035*	0.017	0.037	-0.017**	0.007	0.009	-0.013	0.007	0.071
2020/21	-14.577***	0.128	0.000	-8.200***	0.112	0.000	-16.068***	0.082	0.000	-8.938***	0.058	0.000
Constant	9.471***	2.251	0.000	3.504	2.297	0.127	3.746**	1.335	0.005	-0.891	1.447	0.538
N	117694			116874			458747			456988		

Source: Year 11 population data, schools N=1,464.

School-level clustered residuals and inverse probability weighting. Controls are listed in Appendix B.

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

Estimation of effect sizes

The estimates of RQ4a1 (PP-eligible pupils) are an effect size of -0.019 for maths (95% CI: -0.039 to 0.001) and 0.000 for English (95% CI: -0.023 to 0.025) as shown in Table 24 below. The table also presents effect sizes associated with RQ4a3 and the two additional analyses that selected schools that targeted 50% and 70% of PP-eligible pupils. They can only be interpreted as causal under the conditional independence assumption detailed in the '*Methods*' section. Limitations are discussed in the '*Conclusion*' section. Two of the results reported have an equivalent of 2 months additional progress: English in the specification selecting schools that targeted 70% of PP-eligible pupils, with an effect size of 0.12 and the same specification in maths with an effect size of 0.10.

Table 24: Primary analysis, RQ4a1, RQ4a3, and RQ4a1, on restricted samples

	Intervention group	Comparison group	Effect size		
Outcome	N (missing)	N (missing)	Total n (intervention; comparison)	Hedges g (95% CI)	p-value
Maths TAG (PP-eligible pupils; RQ4a1)	31,516 (4,110)	30,508 (19,280)	62,024 (23,390)	-0.019 (-0.039 to 0.001)	0.063
English TAG (PP-eligible pupils; RQ4a1)	31,516 (4,110)	30,508 (19,280)	62,024 (23,390)	0.000 (-0.021 to 0.025)	0.856
Maths TAG (all pupils; RQ4a3)	121,821 (10,071)	119,682 (10,102)	241,503 (20,173)	-0.020 (-0.033 to -0.007)	0.003
English TAG (all pupils; RQ4a3)	121,821 (10,071)	119,682 (10,102)	241,503 (20,173)	-0.010 (-0.027 to 0.006)	0.209
Maths TAG (RQ4a1, schools that targeted 50% of PP-eligible pupils)	6,419 (588)	6,232 (529)	12,651 (1,117)	-0.009 (-0.044 to 0.026)	0.614
English TAG (RQ4a1, schools that targeted 50% of PP-eligible pupils)	6,419 (588)	6,232 (529)	12,651 (1,117)	-0.027 (-0.068 to 0.015)	0.214
Maths TAG (RQ4a1, schools that targeted 70% of PP-eligible pupils)	1,799 (178)	1,858 (139)	3,657 (317)	0.098 (0.014 to 0.181)	0.023
English TAG (RQ4a1, schools that targeted 70% of PP-eligible pupils)	1,799 (178)	1,858 (139)	3,657 (317)	0.117 (0.046 to 0.187)	0.001

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.

Conclusion

Table 25: Summary of findings

Finding
Initial checks on the data indicated that the Teacher Assessed Grades (TAGs) would be suitable as an outcome measure for some exploratory analysis into the impact of TP, 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 TP made similar progress in English and maths compared to pupils eligible for Pupil Premium in comparison schools (there was no evidence of an effect in English or maths). A particular challenge is that, on average, only 12% of pupils eligible for Pupil Premium were selected for tutoring in maths and 9% were selected for tutoring in English, meaning the vast majority of the pupils included in the analysis did not receive tutoring. Therefore this estimated impact of TP is diluted and it is hard to detect any effect that may (or may not) be present.
When looking at all pupils in Year 11, pupils in schools that received TP made, on average, similar progress in English compared to all Year 11 pupils in comparison schools (there was no evidence of an effect). In maths, Year 11 pupils in schools that received TP made slightly less progress than all Year 11 pupils in comparison schools (though this effect was very small and equivalent to zero months ' additional progress). However, this analysis was subject to even further dilution than the PP-eligible analysis: only 7% of Year 11 pupils were selected for tutoring in maths and 6% in English. Given this context, it is unlikely that any of these differences were due to TP.
Additional analysis restricted the sample of schools to those that targeted higher proportions of pupils eligible for Pupil Premium to receive tutoring, to reduce the issue of dilution and bring the group of analysed pupils closer to those that were selected for the intervention. In schools that selected over 50% of pupils eligible for Pupil Premium for tutoring, pupils eligible for Pupil Premium made similar progress in TP and comparison schools in English and maths. However, when the sample was restricted to schools that selected over 70% of pupils eligible for Pupil Premium for tutoring (and reducing dilution further), the impact of TP on pupils eligible for Pupil Premium is positive. In these schools, pupils eligible for Pupil Premium made, on average, the equivalent of two months additional progress in English and two months additional progress in maths, compared to pupils eligible for Pupil Premium in comparison schools. This analysis was based on a smaller sample of schools that were rematched to a comparison sample. However, different characteristics to the rest of the TP population of schools remained (more 'Outstanding' schools, lower percentage of FSM students), so this finding may not necessarily be generalisable to all TP schools.
Within schools that participated in TP, pupils who received more hours of tutoring in maths obtained higher maths TAGs, and pupils who received more hours of tutoring in English obtained higher English TAGs, than pupils who received fewer hours of tutoring in the respective subjects. These results are associations and are not necessarily causal estimates of impact; there may be other explanations for the higher grades among these pupils.

Interpretation

The findings reported here should be considered exploratory due to the use of the TAGs as an outcome measure; the TAGs were introduced for the 2021 exam session in response to the ongoing disruption to schools in the academic year 2020/21. Some initial 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. The checks indicated that the TAGs were slightly higher in 2020/21 compared to KS4 in previous years. There is some evidence from another of the tests that there may have been negative bias and/or negative selection of pupils into TP; however, this test was conducted without a comparison group. It does not appear that there were systematic differences in grading between TP and comparison 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 analysis proceeded 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 Tuition Partners (TP) programme on pupil attainment, and was designed to do so using a QED design involving several estimators of impact.

None of the school-level impact estimators was able to detect an impact of the availability of TP in its first year on English TAGs. However a very small negative effect, equivalent to zero months additional progress, was detected on maths TAGs (PP-eligible pupils and all pupil specifications). These analyses were subject to high dilution (i.e. they were registering the outcomes of a large proportion of PP-eligible pupils who were not selected for TP). In this Year 11 analysis, there was a bigger pool of schools for the analysis than in the primary school analysis (reported separately). This meant that it was possible to apply some sample restrictions in the form of additional analyses on the PP-eligible pupil analysis. Additional analysis was conducted on subsamples of schools (as set out in the study plan) where a higher proportion (50% and 70% or more) of PP-eligible pupils participated in TP – thus bringing the group of analysed pupils closer to those who participated and therefore reducing dilution. While no evidence of an effect was found in the 50% PP-eligible sample, the analysis of the 70% PP-eligible sample found that the availability of TP had a positive impact on PP-eligible pupils in both maths and English in these schools, with effect sizes of 0.098 for maths and 0.117 for English. This equates to two months additional progress in maths and in English. These analyses were based on small samples of TP schools, and with different characteristics to the rest of the TP population of schools, as they were more likely to be rated 'Outstanding' and featured a lower fraction of FSM pupils compared to the rest of the TP schools. This needs to be borne in mind when interpreting these results. However, the 70% PP-eligible sample much reduces the issue of dilution faced in the other analyses in this study.

The issue of dilution is important to understand. 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 TP from comparison schools despite efforts to do so. For a pupil-level intervention such as TP, this is a major challenge to its evaluation. The analysis was conducted on proxy groups that may give an indication of impact, but as most of the pupils in the analysed group (whether considering the PP-eligible pupils analysis or the all pupils analysis) did not in fact participate in TP, the dilution of any effect may have meant the analyses were underpowered. In addition to the pupil-level dilution, the shift in delivery to later in the academic year meant was one of the reasons that not all of the pupils that received TP had completed a full block of tutoring at the time the TAGs were assumed to have been determined, which is a further challenge to the analysis. On average at pupil level, pupils had received only 63% of a block of 12 hours in English (approx. 7.6 hours) and 70% of a block in maths (approx. 8.4 hours) by the time of the assessment, short of the minimum of 12 hours that was considered a complete block. The dosage data was incomplete (see '*Limitations and lessons learned*' section, below), but there were also delivery reasons for partial completion of blocks. The IPE found a number of reasons for not completing/attending tutoring sessions including: Covid-related absences (of pupils and tutors); lack of engagement from pupils and/or parents; poorer attendance in after school tutoring than in school time; disruption where whole-class bubbles had to isolate; and where tutors failed to establish a good rapport with pupils.

The moderation analysis on tutoring model (RQ6) indicates that higher pupil-level dosage is correlated with better English and maths TAGs. It has to be noted that dosage is endogenous and it is not possible to control for this level of selection as the IV analysis could not proceed in the Year 11 analysis. These results could easily be explained by higher ability children having a tendency to attend more sessions, for example. However, as tutoring is a pupil-level intervention, these pupil-level dosage results are of more interest than any school-level dosage findings. These positive associations between the amount of tutoring received and attainment scores are in line with the evidence on tutoring in the EEF Toolkit although the Toolkit notes more evidence is available on English (reading) tutoring at primary than secondary phase. Given the exploratory nature of this analysis and the limitations reported in more detail below, the evaluators recommend further research on the role of moderators in tuition delivery.

The TP programme was initiated at a time of great pressure on schools, when the education system had been disrupted by school closures to most pupils, and schools were contending with ongoing widespread pupil and staff absences. The TP programme was backed by central investment and support, but it was not the only way schools chose to support their pupils, and it was not possible to account for other initiatives and practices that comparison schools may have been deploying to support their pupils.

Given the unprecedented Covid-related circumstances in which the TP programme was implemented, and the continuing Covid-related disruptions in schools throughout the academic year 2020/21, the findings from the evaluation need to be interpreted in light of this context. The evidence presented here is specific to the implementation of TP in Year 11 during the 2020/21 academic year. Therefore, these results may not be fully generalisable to future years of the programme or to tutoring more widely.

In summary, some of the main analyses were unable to detect if TP had an effect because of the relatively low proportion of PP-eligible pupils receiving tutoring, and because schools selected pupils for tutoring based on characteristics that were unobservable in the available data. The novel use of TAGs as an outcome measure should also be noted. It is therefore, both prudent and important to interpret the evaluation's results in this context and to exercise caution when drawing conclusions. However, despite these challenges, the evaluation found that higher amounts of tutoring at a pupil level seemed to be associated with better TAGs at Year 11 in English and in maths. And where it was possible to analyse outcomes in schools with higher proportions of PP-eligible pupils taking part (70% or more), the results indicated that the programme may have a positive impact on Year 11 TAGs in English and in maths (however, it should be noted that this analysis was based on a smaller sample of schools with some different characteristics to all TP schools).

Limitations and lessons learned

A first limitation of the study concerns the validity of TAGs as an outcome measure for research. There were several potential issues around the appropriateness of using TAGs as an outcome measure that were considered. It must be emphasised that this discussion is no reflection or commentary on the use of TAGs as an assessment outcome. This arose because of the intended use of the TAGs as a research 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 would usually be drawn on to determine the reliability and validity of an outcome measure was not available. Furthermore: teachers awarding the TAGs may have known, which pupils had been selected for tutoring, which has the potential to influence allocation of TAGs (unconsciously – positively or negatively); and the nature of the TAGs meant that teachers could use evidence from earlier in the year and the TAGs may not in all cases be determined on performance after tutoring. However, in the absence of any other outcome data it was felt that some exploratory analysis, as reported here, was better than not trying to estimate impact at all. To mitigate against this, some initial checks on the data were carried out as to its likely suitability as an outcome measure for this evaluation.

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, the checks do not point towards systematic differences across TP and comparison schools in the way they allocated TAGs. Second, there is some evidence that there may have been negative bias and/or negative selection of pupils into TP, in terms of attainment. Third, the checks indicate that the dosage analysis should pick up impact change induced by the TP programme. This is unsurprising given the purpose of TP and is not a concern for the PP-eligible pupils or ‘all pupils’ analysis.

The issue of dilution is an important one for this analysis. The original design introduced a range of research questions designed to complement each other as a counterbalance in the event that schools selected pupils for participation in TP in different ways. It was anticipated that, due to the focus on supporting disadvantaged pupils and the guidance provided to schools, PP-eligibility would be a common characteristic of pupils selected to receive TP. The evaluation also intended to predict which pupils would participate in TP using the data available. It was anticipated that one or other of these would enable the evaluation to identify a good counterfactual in the comparison schools. However, in the event, neither of the strategies were very successful due to the way pupils were selected for TP. First, only one-quarter of PP-eligible pupils in Year 11 in TP schools were identified to take part in the programme. Second, it was not possible to identify the pupils who would have participated in TP in the comparison schools because the participating schools used information to select pupils into the programme that is not observable in the datasets, suggesting that pupil-level selection was driven by unobservable dimensions and thus, could not be accounted for in the analysis (which is supported by the findings from the IPE report). Therefore the analysis reports the impact on attainment of the availability of TP on specific groups of pupils that it was possible to identify in both intervention and comparison schools (specifically PP-eligible pupils and all pupils). Taken together, this means that the estimates are for groups of pupils that do not directly align with the group of pupils that participated in TP; the report refers to this issue as dilution. With such high dilution, it was unlikely that the PP-eligible pupils and all pupils analyses would be able to detect an effect, despite having a suitable MDES. However, when the evaluation aims to address this issue by selecting a (small) sample of schools that targeted 70% of PP-eligible pupils and a matched sample of comparison schools, there appears to be a positive and significant impact of TP availability on both maths and English, suggesting that dilution may be preventing the evaluation from detecting significant impact of the TP intervention. (This additional analysis, while useful, was based on a smaller sample of schools that were rematched to a comparison sample and differences to the rest of the TP population of schools remained (more ‘Outstanding’ schools, lower percentage of FSM students), so this finding may not necessarily be generalisable to all TP schools.)

Another related limitation concerns the study design. Neither schools nor pupils were randomly assigned to treatment and control groups. Given the urgency of the requirement for catch-up support in schools it was not considered ethical to randomise. QEDs are the next best impact evaluation tool, but they have challenges and limitations, chiefly relating to creating a suitable comparison group. In this Year 11 analysis, the schools that signed up to the intervention represented the treatment group and comparison groups were 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, the analysis 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 TP intervention. Given the evaluation design was based on recent research by Weidmann and Miratrix (2020) the evaluation team are reasonably confident that the evaluation design removed school-level selection bias from its comparisons. Weidmann and Miratrix (2020) compared school-level comparison groups matched on observable characteristics with randomised control groups, and they found little trace of unobserved factors that might invalidate conclusions from such a QED. However, the inability of the design reported here to fully address pupil-level selection bias severely limits the conclusions that can be drawn.

The analysis reported here is based on the participation and monitoring data supplied by TPs. This dataset held information about the tutoring itself for this analysis (e.g. model of tutoring, dosage data). TPs were required to submit this data regularly to the EEF as part of their contractual requirement, with the knowledge that it would be used as part of the evaluation. It should be noted that there was no quality checking on the data in the same way that the evaluators were able to check the completeness (for example) of the primary school evaluation sample assessment data that forms the outcome data for the analysis in primary schools. While the quality and completeness of this ‘population’ (participation) data was better than originally expected, there were some gaps and inconsistencies in the dataset, for example related to numbers and dates of sessions of tutoring, as well as pupil details required for matching. This has implications for the data about the intervention, but also in terms of how well the pupil data could be matched to the NPD: if pupil data did not match, it dropped out of the analysis.

Originally the majority of TP delivery was scheduled to take place before the main testing period in the summer term. However, the national restrictions in spring 2021 and the announcement that exams in summer term 2021 could not go ahead as planned (and subsequent introduction of TAGs for 2021) influenced the pattern of delivery. Delivery shifted later in the year, and some of it moved to online delivery rather than face-to-face. Consequently, dosage was added to the data submission request to TPs part way through the programme – originally this was not requested to minimise burden on TPs. The delay to adding this meant that not all TPs were able to supply this information, for example not all TPs recorded this level of detail centrally. Therefore, there are some gaps in the data upon which the dosage analysis is based.

The number of schools in the analysis, while substantial, is lower than anticipated at the study plan stage. Allocation of tuition, which initially was highest for Year 11 at secondary, shifted so that in later months the allocation to Year 10 pupils increased. This may have been related to the change to the usual exams process for summer 2021.

Future research and publications

Given the exploratory nature of this analysis, and the large scale of the intervention, future evaluations might look to explore the presence of more firm evidence on the impact of TP. It may also be worthwhile to examine the effect of increased dosage, in which pupils are offered more tuition, if that is somehow exogenously assigned to pupils.

It would be important to address pupil-level selection, if possible, specifically to ensure the selection of pupils is not endogenously determined.

Flexibility over delivery of the NTP had been built in from the start but challenging circumstances meant that the level of tailoring to delivery was greater than originally planned. Not least, the method of pupil selection for tutoring which was influenced by much more than simply Pupil Premium status than was observed in the tutoring pilot in summer 2020. This caused issues for the evaluation’s ability to identify a suitable counterfactual group of pupils; similar pupils who did not participate in tutoring. 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 unobservables which is so ingrained here. The evaluators recommend that in future years of the TP programme, efforts are made to evaluate different types of tutoring with a randomised design; for example, by varying the number of hours of tuition or how many sessions of tutoring per week are delivered, to explore the optimum dosage and pattern of delivery.

This was an evaluation of not one tutoring provider, but of a tutoring programme comprising 33 different tuition provider organisations. The providers were selected according to specific criteria and required to follow some key delivery principles (for example blocks of up to 15 hours in a single subject) and given guidance and support from the EEF, Nesta and Impetus. However, in practice there was a wide variety in delivery, as reported by the IPE report. Future evaluation work should ideally focus on the myriad of different factors that might influence the effectiveness of tuition, to follow up on the associations suggested by the moderator analysis reported here. Mode (online or in person), location, subject, year group, duration, frequency, qualifications of tutor, integration with the curriculum and extent of tutor/tutee matching are examples of these.

Changes have been made to the NTP programme since the delivery of Year 1, which is evaluated here. In Year 2, a school-led tutoring model was introduced as a third pillar, and the Department for Education has announced plans to simplify the programme for year 3. In year 3 the DfE will provide £358 million of core tutoring funding directly to schools, giving them the freedom to decide how best to provide tutoring for their pupils.

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.

EEF, 2020. *Impact of school closures on the attainment gap: rapid evidence assessment*. Education Endowment Foundation. [online] Available at: https://educationendowmentfoundation.org.uk/public/files/EEF_(2020)_-_Impact_of_School_Closures_on_the_Attainment_Gap.pdf [Accessed 29 July 2022].

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

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

Marshall, L., Bury, J., Wishart, R., Hammelsbeck, R. and Roberts, E., 2021 *Online tuition pilot*. [online] EEF. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/online-tuition-pilot [Accessed 29 July 2022].

Ofqual (Office of Qualifications and Examinations Regulation). 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 6 September 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.

Sharp, C., Nelson, J., Lucas, M., Julius, J., McCrone, T. and Sims, D., 2020*. Schools' responses to Covid: 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].

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 environmen*t. University College London [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 TP participation

	Coef	S.E.	P-value
School types (Academy as base):			
Community school	-0.195	0.123	0.111
Voluntary aided school	-0.007	0.144	0.961
Voluntary controlled school	0.115	0.334	0.731
Foundation school	0.074	0.146	0.610
City technology	1.285	0.984	0.192
Free school – Mainstream	-0.036	0.092	0.697
Special free school	0.161	0.195	0.409
Free school UTC	-0.090	0.278	0.746
Free school – studio school	0.209	0.338	0.537
Ofsted rating 2017 (missing as base):			
Outstanding	0.264	0.252	0.295
Good	0.238	0.200	0.235
Inadequate	-0.345	0.260	0.184
Requires improvement	0.149	0.209	0.474
Ofsted rating 2018 (missing as base):			
Outstanding	-0.116	0.297	0.696
Good	0.045	0.248	0.855
Inadequate	0.110	0.273	0.688
Requires improvement	0.060	0.248	0.809
Missing Ofsted rating	-0.349	0.392	0.373
Region (East midlands as base):			
East of England	0.377**	0.142	0.008
London	0.262	0.167	0.117
North East	-0.090	0.203	0.656
North West	0.257	0.148	0.083
South East	-0.004	0.135	0.978
South West	0.130	0.143	0.362
West Midlands	0.074	0.140	0.598
Yorkshire and the Humber	0.169	0.147	0.249
School-level IDACI 2019 Quintile (missing as base):			
Q1	-0.905	0.847	0.285
Q2	-1.490	0.838	0.075
Q3	-1.798*	0.842	0.033
Q4	-1.935*	0.871	0.026
Q5	-2.225*	0.913	0.015
School-level IDACI 2018 Quintile (missing as base):			
Q1	0.492	0.532	0.355
Q2	0.654	0.491	0.183
Q3	0.438	0.519	0.398
Q4	0.809	0.561	0.149
Q5	1.052	0.660	0.111
School-level IDACI 2017 Quintile (missing as base):			
Q1	0.090	0.465	0.846
Q2	0.303	0.446	0.498
Q3	0.728	0.487	0.135
Q4	0.459	0.529	0.386
Q5	0.267	0.596	0.654
KS1 to KS2 value added 2018 at local district level	0.104	0.097	0.283
Pupil count 2020	0.000	0.000	0.745
Pupil count 2019	0.001	0.001	0.246
Pupil count 2018	-0.000	0.000	0.364
Pupils-to-teacher ratio 2018	0.026	0.015	0.088
Pupils-to-teacher ratio 2017	-0.027	0.017	0.123
IDACI tertiles with average attainment in 2017	0.024	0.038	0.527
IDACI tertiles with average attainment in 2018	-0.049	0.044	0.257
IDACI tertiles with average attainment in 2019	0.052	0.033	0.115
Census school level % FSM Spring 2019	-1.479	1.716	0.389
Census school level % FSM Spring 2018	-1.120	1.433	0.435
Census school level % FSM Spring 2017	1.158	0.726	0.110
Census school level % FSM Spring 2020	2.890*	1.264	0.022
Census school level % EAL Spring 2019	-0.036	0.970	0.971
Census school level % EAL Spring 2018	0.619	0.724	0.392

Census school level % EAL Spring 2020	-0.140	0.700	0.841
Census school level %SEN Spring 2019	-2.519*	1.201	0.036
Census school level %SEN Spring 2018	0.532	0.952	0.576
Census school level %SEN Spring 2020	0.645	0.984	0.512
Read KS2 attainment 2017	0.062	0.035	0.076
Read KS2 attainment 2018	0.030	0.038	0.426
Read KS2 attainment 2019	-0.018	0.040	0.655
Maths KS2 attainment 2017	-0.003	0.016	0.834
Maths KS2 attainment 2018	-0.016	0.018	0.369
Maths KS2 attainment 2019	0.029	0.018	0.094
English KS4 attainment 2017	0.021	0.149	0.887
English KS4 attainment 2018	-0.034	0.154	0.825
English KS4 attainment 2019	-0.039	0.142	0.781
Maths KS4 attainment 2017	-0.311*	0.152	0.040
Maths KS4 attainment 2018	-0.103	0.157	0.512
Maths KS4 attainment 2019	0.013	0.163	0.937
Rural (urban as base):			
rural	-0.059	0.098	0.546
Constant	-10.577	9.810	0.281
Observations	2067		

Note: TP intervention sample. Missing variables of all controls included but not listed.

Table A2: Baseline characteristics of Year 11 TP schools and eligible comparison schools

Variable	Means:Eligible comparison	SD:Eligible comparison	Means:TP schools	SD:TP schools	Difference	Std Difference
School-level PP KS2 Maths scores 2020/21	101.564	(3.719)	101.197	(2.531)	-0.367**	-0.082
School-level PP KS2 Read scores 2020/21	101.021	(3.873)	100.447	(2.736)	-0.573***	-0.121
School-level KS4 English, 2019/20	5.192	(0.825)	5.031	(0.647)	-0.161***	-0.153
School-level KS4 Maths 2019/20	5.094	(0.972)	4.890	(0.718)	-0.204***	-0.169
School-level KS4 English, 2018/19	4.811	(0.884)	4.640	(0.661)	-0.171***	-0.155
School-level KS4 Maths, 2018/19	4.790	(1.019)	4.561	(0.744)	-0.229***	-0.181
School-level KS4 English, 2017/18	4.809	(0.847)	4.648	(0.662)	-0.161***	-0.150
School-level KS4 Maths, 2017/18	4.785	(0.998)	4.571	(0.720)	-0.214***	-0.174
Total pupil counts	883.232	(335.271)	954.779	(307.484)	71.548***	0.157
Pupils-to-teacher ratio 2018	16.298	(2.725)	16.257	(2.663)	-0.041	-0.011
Ofsted 2018: Outstanding	0.248		0.194		-0.055***	-0.094
Ofsted 2018: Good	0.496		0.562		0.065***	0.093
Ofsted 2018: Inadequate	0.045		0.039		-0.006	-0.021
Ofsted 2018: Requires improvement	0.135		0.152		0.017	0.034
Ofsted 2018: Missing	0.075		0.054		-0.021*	-0.061
School type: Academy-sponsor led	0.191		0.229		0.038**	0.066
School type: Community school	0.108		0.096		-0.012	-0.029
School type: Voluntary aided/controlled school	0.070		0.080		0.010	0.026
School type: Foundation school	0.049		0.066		0.017*	0.053
School type: Free school - mainstream	0.515		0.470		-0.046**	-0.064
School type: Others	0.067		0.060		-0.007	-0.022
Urban	0.797		0.861		0.063***	0.119
Rural	0.162		0.119		-0.043***	-0.087
Urban/Rural missing	0.041		0.020		-0.020**	-0.084
Region: East Midlands	0.090		0.070		-0.020	-0.052
Region: East of England	0.087		0.114		0.026*	0.062
Region: London	0.133		0.196		0.063***	0.12
Region: North East	0.037		0.034		-0.003	-0.012
Region: North West	0.106		0.138		0.032**	0.069
Region: South East	0.184		0.130		-0.054***	-0.105
Region: South West	0.111		0.099		-0.012	-0.027
Region: West Midlands	0.120		0.106		-0.014	-0.032
Region: Yorkshire & the Humber	0.091		0.093		0.002	0.006
Region: Missing	0.041		0.020		-0.020**	-0.084
AM participation	0.000		0.134		0.134***	0.393
Census school-level % FSM Spring 2021	0.255		0.296		0.041***	0.196
% EAL	0.143		0.177		0.034***	0.128
% SEN	0.216		0.218		0.003	0.022
% Female	0.494		0.503		0.009	0.034
Average IDACI scores	0.037	(0.023)	0.040	(0.020)	0.003***	0.093
% White British	0.123		0.110		-0.013***	-0.129
% Asian	0.018		0.021		0.003*	0.063
% Black	0.015		0.020		0.005***	0.122
% Other ethnic	0.026		0.030		0.003***	0.09
% Unknown ethnic	0.004		0.004		0.000	0.031
% Not white	0.058		0.070		0.012***	0.127
Observations	1,328		739		2,067	

Source: Year 11 population data.

Figure A1*: Common support*

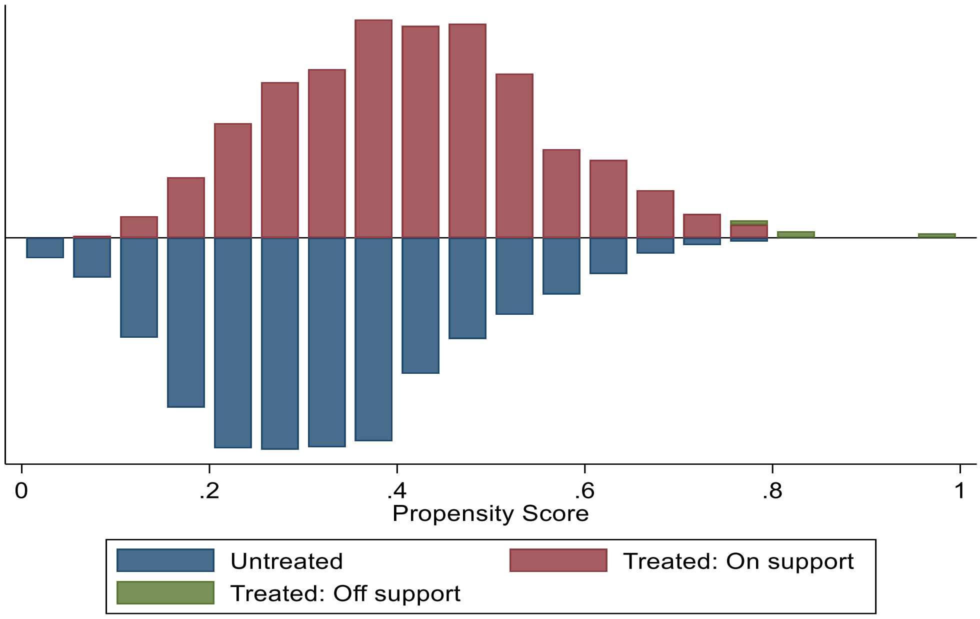


Table A3a: Summary statistics of previous years' GCSEs and previous years' and current year KS2, before and after matching, no caliper

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
KS4 English 2016/17	Unmatched	4.71	4.87	-0.16	0.04	-4.48
	ATT	4.71	4.70	0.01	0.04	0.31
KS4 English 2017/18	Unmatched	4.65	4.81	-0.16	0.04	-4.46
	ATT	4.65	4.66	0.00	0.04	-0.03
KS4 English 2018/19	Unmatched	4.64	4.81	-0.17	0.04	-4.59
	ATT	4.64	4.64	0.00	0.04	0.05
KS4 English 2019/20	Unmatched	5.03	5.19	-0.16	0.04	-4.58
	ATT	5.03	5.03	0.00	0.04	0.06
KS4 Maths 2016/17	Unmatched	4.59	4.82	-0.23	0.04	-5.63
	ATT	4.59	4.61	-0.02	0.04	-0.52
KS4 Maths 2017/18	Unmatched	4.58	4.79	-0.21	0.04	-5.13
	ATT	4.57	4.58	-0.01	0.04	-0.25
KS4 Maths 2018/19	Unmatched	4.56	4.79	-0.23	0.04	-5.36
	ATT	4.57	4.57	0.00	0.04	-0.03
KS4 Maths 2019/20	Unmatched	4.89	5.09	-0.20	0.04	-5.00
	ATT	4.89	4.90	-0.01	0.04	-0.27
KS2 Read 2020/21	Unmatched	102.32	102.99	-0.67	0.15	-4.61
	ATT	102.33	102.38	-0.05	0.14	-0.38
KS2 Maths 2020/21	Unmatched	102.89	103.35	-0.46	0.13	-3.51
	ATT	102.89	102.91	-0.02	0.12	-0.17
KS2 Read 2016/17	Unmatched	31.36	32.07	-0.71	0.15	-4.63
	ATT	31.37	31.33	0.04	0.16	0.27
KS2 Read 2017/18	Unmatched	32.53	33.22	-0.69	0.16	-4.25
	ATT	32.53	32.58	-0.05	0.16	-0.30
KS2 Read 2018/19	Unmatched	30.94	31.63	-0.69	0.15	-4.68
	ATT	30.96	31.01	-0.05	0.15	-0.35
KS2 Reaf 2019/20	Unmatched	31.02	31.64	-0.63	0.15	-4.18
	ATT	31.03	31.03	0.00	0.15	0.00
KS2 Maths 2016/17	Unmatched	70.07	71.55	-1.48	0.32	-4.57
	ATT	70.06	70.10	-0.04	0.32	-0.13
KS2 Maths 2017/18	Unmatched	70.52	71.71	-1.19	0.34	-3.48
	ATT	70.52	70.56	-0.04	0.32	-0.13
KS2 Maths 2018/19	Unmatched	71.15	72.48	-1.33	0.33	-4.04
	ATT	71.17	71.25	-0.08	0.33	-0.24
KS2 Maths 2019/20	Unmatched	71.21	72.19	-0.98	0.32	-3.06
	ATT	71.21	71.17	0.04	0.31	0.14
Total	7	2,060	2,067			

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 2016/17	Unmatched	4.71	4.87	-0.16	0.04	-4.48
	ATT	4.72	4.70	0.02	0.04	0.41
KS4 English 2017/18	Unmatched	4.65	4.81	-0.16	0.04	-4.46
	ATT	4.67	4.66	0.01	0.04	0.19
KS4 English 2018/19	Unmatched	4.64	4.81	-0.17	0.04	-4.59
	ATT	4.65	4.64	0.01	0.04	0.27
KS4 English 2019/20	Unmatched	5.03	5.19	-0.16	0.04	-4.58
	ATT	5.04	5.03	0.01	0.04	0.36
KS4 Maths 2016/17	Unmatched	4.59	4.82	-0.23	0.04	-5.63
	ATT	4.60	4.61	-0.01	0.04	-0.23
KS4 Maths 2017/18	Unmatched	4.58	4.79	-0.21	0.04	-5.13
	ATT	4.59	4.58	0.01	0.04	0.16
KS4 Maths 2018/19	Unmatched	4.56	4.79	-0.23	0.04	-5.36
	ATT	4.58	4.56	0.01	0.04	0.26
KS4 Maths 2019/20	Unmatched	4.89	5.09	-0.20	0.04	-5
	ATT	4.90	4.90	0.00	0.04	-0.05
KS2 Read 2020/21	Unmatched	102.32	102.99	-0.67	0.15	-4.61
	ATT	102.37	102.37	0.00	0.15	0
KS2 Maths 2020/21	Unmatched	102.89	103.35	-0.46	0.13	-3.51
	ATT	102.89	102.90	-0.01	0.13	-0.1
KS2 Read 2016/17	Unmatched	31.36	32.07	-0.71	0.15	-4.63
	ATT	31.43	31.31	0.11	0.17	0.68
KS2 Read 2017/18	Unmatched	32.53	33.22	-0.69	0.16	-4.25
	ATT	32.58	32.55	0.03	0.17	0.18
KS2 Read 2018/19	Unmatched	30.94	31.63	-0.69	0.15	-4.68
	ATT	31.05	31.00	0.05	0.15	0.34
KS2 Reaf 2019/20	Unmatched	31.02	31.64	-0.63	0.15	-4.18
	ATT	31.06	31.00	0.06	0.16	0.39
KS2 Maths 2016/17	Unmatched	70.07	71.55	-1.48	0.32	-4.57
	ATT	70.11	70.09	0.02	0.34	0.05
KS2 Maths 2017/18	Unmatched	70.52	71.71	-1.19	0.34	-3.48
	ATT	70.52	70.52	0.00	0.34	-0.01
KS2 Maths 2018/19	Unmatched	71.15	72.48	-1.33	0.33	-4.04
	ATT	71.26	71.24	0.02	0.34	0.06
KS2 Maths 2019/20	Unmatched	71.21	72.19	-0.98	0.32	-3.06
	ATT	71.16	71.12	0.04	0.33	0.11
Total	58	2,009	2,067			

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 2016/17	Unmatched	4.71	4.87	-0.16	0.04	-4.48
	ATT	4.72	4.70	0.02	0.04	0.41
KS4 English 2017/18	Unmatched	4.65	4.81	-0.16	0.04	-4.46
	ATT	4.66	4.66	0.00	0.04	0.09
KS4 English 2018/19	Unmatched	4.64	4.81	-0.17	0.04	-4.59
	ATT	4.66	4.64	0.01	0.04	0.25
KS4 English 2019/20	Unmatched	5.03	5.19	-0.16	0.04	-4.58
	ATT	5.05	5.03	0.02	0.04	0.44
KS4 Maths 2016/17	Unmatched	4.59	4.82	-0.23	0.04	-5.63
	ATT	4.60	4.61	-0.01	0.04	-0.34
KS4 Maths 2017/18	Unmatched	4.58	4.79	-0.21	0.04	-5.13
	ATT	4.58	4.58	0.00	0.04	0.08
KS4 Maths 2018/19	Unmatched	4.56	4.79	-0.23	0.04	-5.36
	ATT	4.58	4.57	0.01	0.05	0.27
KS4 Maths 2019/20	Unmatched	4.89	5.09	-0.20	0.04	-5
	ATT	4.90	4.90	0.00	0.04	0.02
KS2 Read 2020/21	Unmatched	102.32	102.99	-0.67	0.15	-4.61
	ATT	102.39	102.37	0.01	0.15	0.09
KS2 Maths 2020/21	Unmatched	102.89	103.35	-0.46	0.13	-3.51
	ATT	102.88	102.90	-0.02	0.13	-0.15
KS2 Read 2016/17	Unmatched	31.36	32.07	-0.71	0.15	-4.63
	ATT	31.44	31.31	0.13	0.17	0.76
KS2 Read 2017/18	Unmatched	32.53	33.22	-0.69	0.16	-4.25
	ATT	32.59	32.55	0.04	0.17	0.21
KS2 Read 2018/19	Unmatched	30.94	31.63	-0.69	0.15	-4.68
	ATT	31.06	31.00	0.05	0.15	0.35
KS2 Reaf 2019/20	Unmatched	31.02	31.64	-0.63	0.15	-4.18
	ATT	31.10	31.00	0.09	0.16	0.58
KS2 Maths 2016/17	Unmatched	70.07	71.55	-1.48	0.32	-4.57
	ATT	70.12	70.10	0.02	0.34	0.05
KS2 Maths 2017/18	Unmatched	70.52	71.71	-1.19	0.34	-3.48
	ATT	70.53	70.53	0.00	0.34	0
KS2 Maths 2018/19	Unmatched	71.15	72.48	-1.33	0.33	-4.04
	ATT	71.29	71.25	0.04	0.34	0.11
KS2 Maths 2019/20	Unmatched	71.21	72.19	-0.98	0.32	-3.06
	ATT	71.25	71.13	0.12	0.33	0.36
Total	61	2006.00	2067.00			

Appendix B

Variables listed in Table 5 of the study plan, used for matching and included as controls in all regressions. All variables from the Census are included for the previous 3 years, while the variables from the dataset we imported in the Secure Research Service in November 2021 and including school-level characteristics can only be included up to 2017/18.

- Key Stage (KS) 1 to KS2 value added attainment, at district level in 2017/18.
- Management/school type secondary—Community, Academies, Foundation, Free schools, Sponsored Academies, Voluntary school, Studio schools, University Technical College.
- School size, total number of students in previous 3 years.
- Teacher–student ratio, in 2017/18 and 2018/19.
- Ofsted, overall effectiveness, 2017/18 and 2018/19.
- Region (London, Government Office Region, and regional dummies).
- School in urban/rural area.
- Income Deprivation Affecting Children Index (IDACI) quintile, in previous 3 years.
- Interaction of IDACI tertiles with average attainment in previous 3 years.
- Free School Meals (FSM)—percentage eligible in previous 4 years.
- English as an Additional Language (EAL)—percentage in previous 3 years.
- Special educational needs (SEN)—percentage in previous 3 years.
- English and maths KS2 attainment, in the previous 3 years and in the current one.
- English and maths KS4 attainment in the previous 3 years.
- Missing dummies for: Pupils-to-teacher ratio, 2018/19; percentage of FSM, SEN, and EAL (all were included in the model but all dropped due to colinearity); KS2 and KS4 read and maths (included in the model but was dropped due to collinearity).

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).
- EAL and EAL unknown.
- SEN.
- Looked after for 12 months, looked after since 31 March, and looked after for 6 months.

School-level controls used in all regressions:

- Ofsted ratings and missing Ofsted rating.
- School FSM percentage above median and missing FSM percentage.
- IDACI quintiles and IDACI missing; and
- AM participation.

Appendix C: Year 11 checks

Year 11 TAGs considerations and checks

The four considerations are outlined in the report: see the section 'Results of the Year 11 checks' and are repeated here for ease of reference:

- Consideration 1: That Teacher Assessed Grades (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 TP 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. This may lead to grades not reflecting a pupil's latest performance.

- Consideration 4: That the assessments may not be sensitive enough to change as a result of pupils having received tutoring. This consideration is linked to the three prior considerations, with all of these potentially affecting the measure's sensitivity to change.

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.

Methods used for the checks

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 3/4 boundary), we compared the distributions of GCSEs awarded in the years before TP (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 issue we will account for in the interpretation of results. If there is a TP effect, we might expect a change too.

ii) To address consideration 2, that schools selected the pupils who undertake TP and that, as a result, teachers may have applied some conscious or unconscious bias in their assessment of these pupils, we used across-subjects 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 TP 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 TP schools versus comparison schools. Note that: i) it would be difficult to disentangle the impact of TP 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 TP. To explore this, we regressed English and maths TAGs separately on English and maths TP hours received. We would expect TP English hours to be correlated with English TAGs and TP maths hours to be correlated with maths TAGs.

Ex post (after the 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 TP 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 dosage. 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, controlling for pupil-level and school-level characteristics.

ii) To address consideration 1, that TAGs may be distributed differently compared to previous years (consideration 1), we also tested if the distribution of tests across the years (i.e. exam year 2021 vs. 2018/2019/2020) was different across TP compared to comparison schools for all pupils and PP-eligible pupils. Evidence of a significant difference in distribution across TP and comparison schools may suggest the presence of measurement errors. If TP and comparison schools both allocated TAGs equally, then we would expect TP schools to have slightly more higher grades because of TP (if TP is effective). If TP and comparison schools allocated TAGs differently, then we would need to investigate whether the difference is related to bias (see checks (ii) in the 'Ex ante' section and (ii) and (iii) in the 'Ex post' section) or is an indication of a positive (or indeed negative) effect of TP. Note that it is difficult to disentangle the impact of TP from a potentially systematic difference in how TAGs are awarded in TP compared to comparison 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 (consideration 1), we planned to perform the analysis on Year 11 pupils predicted to do TP. To do this, we first estimated a pupil-level logistic model for TP participation and checked that the model had good predictive power. If so, we proceeded with predicting pupils participating in TP and pupils not participating in TP. If the effect of TP on pupils predicted to participate is positive, this can be due to the positive effect of TP or to different allocation of TAGs compared to exams. We also planned to perform the analysis on Year 11 children predicted NOT to participate in TP. If there was an impact also on children not predicted to participate in TP, then it could be interpreted as evidence of TAGs being allocated differently to exams. 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 TP is effective, predicted TP should always have higher TAGs than predicted non-TP even if TAGs are allocated differently compared to exams. The reliability of this test depends on how well we can predict participation to TP. The test cannot disentangle the increase in grade due to TP from a systematic teacher bias towards TP pupils only.

While we already knew (at the time of writing the study plan) that TAGs were likely to be distributed differently compared to the previous year (consideration 1), it is not indicative of systematic bias between TP and comparison schools.[40] 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 would be more concerned about the validity of TAGs as an outcome measure if some of the checks outlined above addressing considerations 2, 3, and 4 pointed towards the presence of systematic bias between TP and comparison schools (specifically: ex ante test (ii); 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.

[40] 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.

Results of the checks

Ex ante (before analysis) 1: Distribution of Key Stage (KS) 4 grades over time

The first ex ante check shows that the distribution of teacher assessed GCSE grades in 2021 is significantly different with respect to the distribution of GCSE grades than 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 TP 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 hypothesis 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.

Table C1: Kolmogorov–Smirnov tests for equality of distribution

	Maths, all pupils	English, all pupils	Maths, PP-eligible pupils	English, PP-eligible pupils
Combined K-S, 2019/20-2010/21:	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.

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 1–2 use the sample of all pupils; columns 3 and 4 use the sample of Free School Meals pupils only.

The results of the t-test are presented in Table C2. They indicate 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 on 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.

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	SE	P-value	Coeff	SE	P-value	Coeff	SE	P-value	Coeff	SE	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.

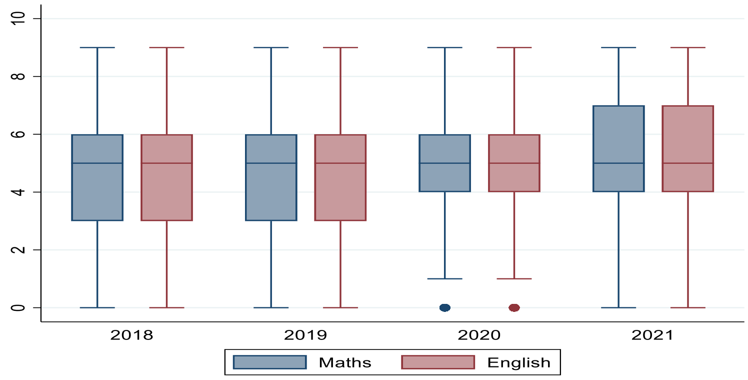
The box plots show the distribution of grades across the years for both maths and English, for all pupils (Figure C1) and PP-eligible pupils only (Figure C1).[41] The plots indicate that the distribution of grades was different, and higher, in 2019/20 and 2020/21 compared to 2017/18 and 2018/19, especially 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).'

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

For PP-eligible pupils, it appears that the distribution of maths grades is tighter in 2019/20 and 2020/21, while English is less tightly bunched in 2020/21 than in 2019/20 and in the previous year.

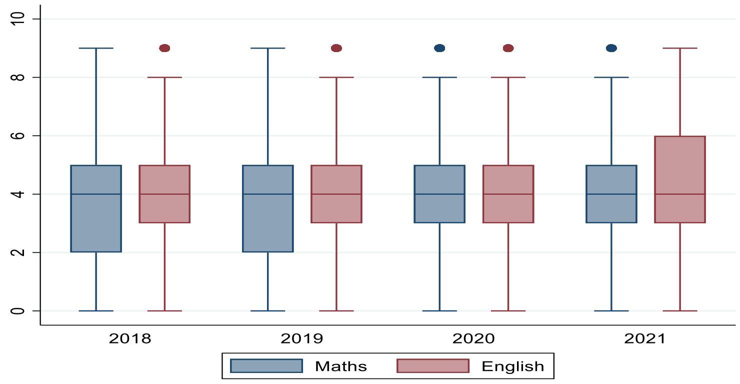
The results from these tests show evidence of consideration 1, i.e. that TAGs are distributed differently compared to previous years. This is consistent with what has been widely reported and it is not a serious concern for our analysis as long as TP and comparison schools do not allocate TAGs differently (we perform this check in ex post *2*).[42]

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 the 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 TP pupils' grades to verify the presence of subject-specific or pupil-specific bias. To do this, we selected the sample of TP pupils in Year 11 and regressed teacher assessed GCSE grades in 2020/21 on dummies for pupil level TP delivery in English and TP delivery in maths, controlling for individual characteristics, regional dummies, and IDACI index.

If the bias is subject-specific, the coefficient associated with the tutored subject should be significant only when regressed on the same subject grade (i.e. TP 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 tutored subjects would be large and significant. If it is not pupil-specific, only the coefficient of the tutored subject is expected to be significant.

The regression results in Table C3 and C4 indicate that for maths, for PP-eligible pupils and all pupils, being tutored in maths and being tutored in English are both associated with a significantly lower grade, although the coefficient of maths

[42] For example, EPI https://epi.org.uk/wp-content/uploads/2021/08/GCSE_analysis_2021_EPI_.pdf**.**

tutoring is larger than the coefficient for English tutoring. For English, for both PP-eligible pupils and all pupils, only being tutored in English is associated with a significantly lower grade in English.[43]

The negative sign of the TP subject coefficient could be explained either by negative teacher bias in the awarding of TAGs towards pupils selected in the TP programme and/or by lower performing pupils being selected for the TP programme (see Table 9). It should also be noted that here we are not comparing TP pupils to similar pupils in comparison schools.

The lack of counterfactual suggests that, if the negative coefficient is driven by negative selection, TP pupils may have performed worse in the absence of TP anyway. While the results of this check represent a serious concern on the validity of TAGs as an outcome measure for this evaluation, it would be necessary to compare the outcomes of TP pupils with a comparison group of similar pupils in comparison 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. We cannot do this because we cannot account for the mechanism through which schools selected pupils for tutoring via observable characteristics. We attempt to move away from pupil-level selection by focusing on PP-eligible pupils and all pupils, as these groups can be identified for both TP and comparison schools, but as the average fraction of PP-eligible pupils doing TP is 25.6%, we cannot properly identify TP pupils in comparison schools (the model aiming to predict participation in TP is reported below in ex post 3).

Table C3: TP impact across-subjects on TP pupils, all pupils

	KS4 Maths			KS4 Maths			KS4 English			KS4 English		
	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values
TP subject English				-0.085*	0.035	0.016	-0.352***	0.029	0.000	-0.376***	0.037	0.000
TP subject Maths	-0.201***	0.029	0.000	-0.242***	0.037	0.000				-0.045	0.036	0.209
N	20387			20387			20374			20374		
R-squared	0.514			0.515			0.405			0.406		

Source: Year 11 population data.

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

Table C4: TP impact across-subjects on TP pupils, PP-eligible pupils

	KS4 Maths			KS4 Maths			KS4 English			KS4 English		
	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values	Coeff	S.E.	p-values
TP subject English				-0.117*	0.051	0.022	-0.287***	0.039	0.000	-0.317***	0.048	0.000
TP subject Maths	-0.155***	0.039	0.000	-0.215***	0.050	0.000				-0.054	0.046	0.248
N	8056			8056			8048			8048		
R-squared	0.474			0.475			0.372			0.373		

Source: Year 11 population data.

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

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

Ex post (after analysis) 1: Dosage of tutoring

The first ex post check assesses whether TAGs are sensitive enough to change, in the event that TAGs do reflect pupils' performance. We exploited dosage of tutoring on the sample of TP schools and regressed TAGs on a variable indicating the blocks of TP hours completed by the TP pupils. The value of the dependent variable is the ratio between the number of hours completed and 12, the total number of hours in a block of tutoring.[44] The regression is run on the sample of TP pupils in Year 11 controlling for individual characteristics, regional dummies, and IDACI index. Results are presented in Table C5. We find that a higher number of hours of tutoring received was correlated with achieving better English and

[43] In both cases, adding an interaction term between being tutored in maths and English does not change the significance and the sign of the coefficients of maths and English tutoring.

[44] The IPE refers to 12 hours rather than 12 sessions, as a session is not always an hour although it often is. We measured the number of sessions in terms of hours as in the IPE report.

maths TAGs. This might point to TAGs being sensitive enough to capture changes in pupil performance (if there is an impact of TP) but could also indicate that pupils who received more hours of tutoring would have achieved higher TAGs even without TP.

Table C5: Correlation between dosage of tutoring and maths / English TAGs among TP pupils in TP schools

	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
TP dosage	0.196***	0.038	0.000	0.150***	0.037	0.000	0.160**	0.049	0.001	0.124**	0.044	0.005
N	17958			17948			7013			7008		
R-squared	0.513			0.399			0.469			0.372		

Note: Sample: population of Year 11 TP schools excluding pupils with zero TP hours. Results in columns 1–3 and 7–9 use the sample of all pupils; columns 4–6 and 10–12 use the sample of FSM pupils only. Residuals clustered at school level.

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

Ex post (after analysis) 2: Distribution of TAGs between TP and matched comparison 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 TP and matched comparison schools for all pupils and PP-eligible pupils. This uses the intervention group and matched comparison group described in the ‘Methods’ section.

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 all comparisons of 2021 grades with previous years, for both maths and English and for all pupils and PP-eligible pupils only.

The histograms in Figure C3 (maths, all pupils), C3 (English, all pupils), C5 (maths, PP-eligible pupils), C6 (English, PP-eligible pupils) show the distribution of the difference between KS4 and teacher assessed GCSEs grades in 2020/21 between TP and comparison schools for all pupils and PP-eligible pupils. The distributions of differences in maths and English grades do not appear to be markedly different across TP and matched comparison schools over the years and, consistent with the other tests, they indicate that the difference in grades between the academic year 2020/21 and 2019/20 is closer to zero than the difference in grades between the academic year 2020/21 and 2018/19 or 2017/18, indicating that TAGs for the academic years 2019/20 and 2020/21 are significantly higher than in previous years for both TP and comparison schools (in line with the result of ex ante 1).

Table C6: Kolmogorov–Smirnov tests for equality of distribution

	Maths, all pupils	English, all pupils	Maths, PP-eligible pupils	English, PP-eligible pupils
Combined K-S, 2019/20-2020/21:	0.0516	0.0227	0.0373	0.0526
p-value	0.291	0.992	0.694	0.271
Combined K-S, 2018/19-2020/21:	0.0364	0.0567	0.062	0.0451
p-value	0.749	0.216	0.142	0.488
Combined K-S, 2017/18-2020/21:	0.0414	0.0342	0.0399	0.0502
p-value	0.635	0.843	0.68	0.388
N	1464	1464	1464	1464

Source: Year 11 population data.

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.

Figure C3: Distribution of difference in KS4 maths scores between TP and matched comparison (non-TP) schools across the academic years, all pupils

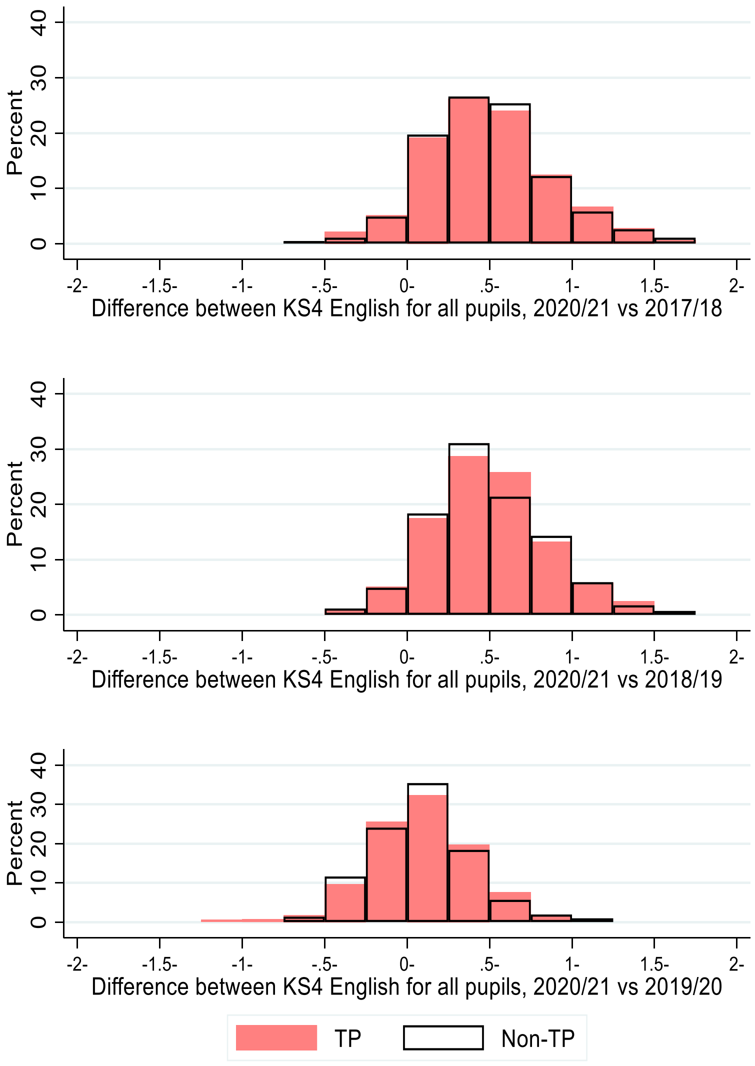


Figure C4: Distribution of difference in KS4 English scores between TP and matched comparison (non-TP) schools across the academic years, all pupils

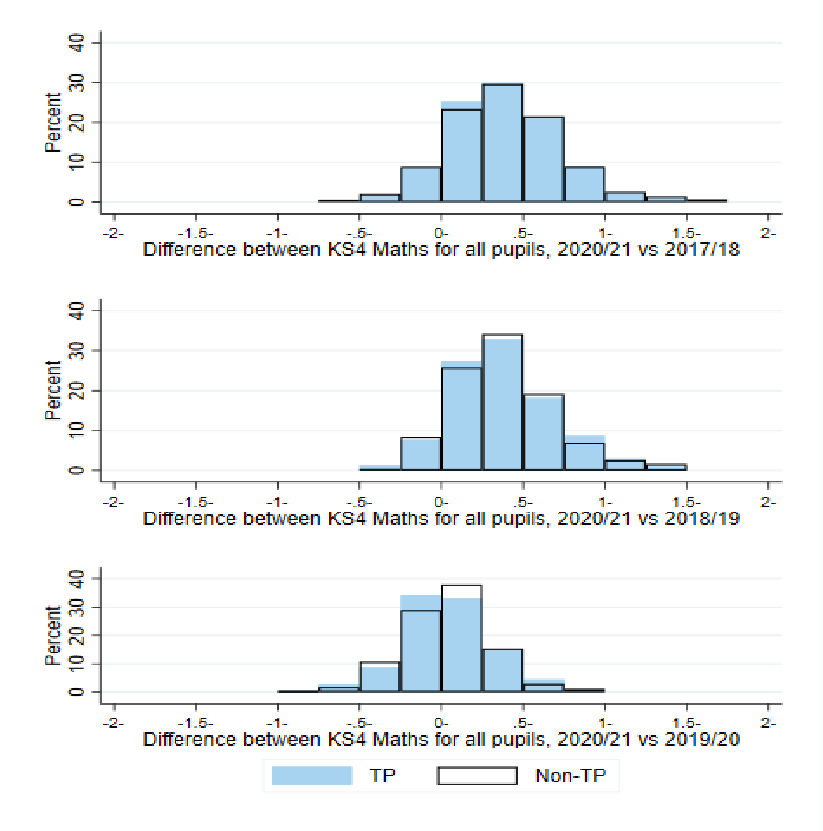


Figure C5: Distribution of difference in KS4 maths scores between TP and matched comparison (non-TP) schools across the academic years, PP-eligible pupils

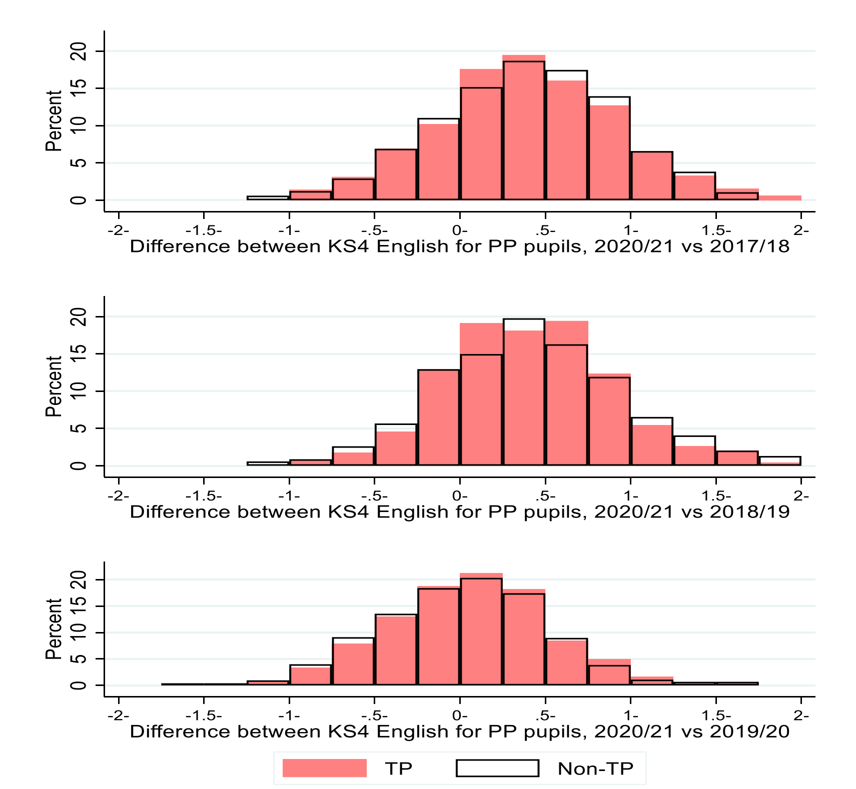
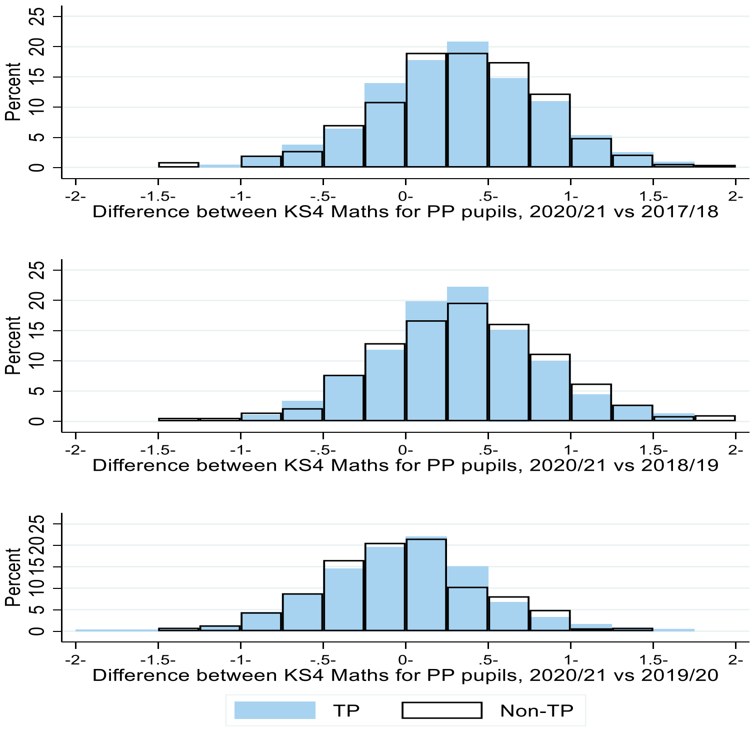


Figure C6: Distribution of difference in KS4 English scores between TP and matched comparison (non-TP) schools across the academic years, PP-eligible pupils



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

The third ex post check on Year 11 was planned to test the presence or absence of impact of TP on pupils predicted not to participate in TP. To do this, we modelled the probability of pupil participation in TP schools using various markers of disadvantage recorded in the National Pupil Database (NPD) (socio-economic status measured by FSM/PP, SEND, interaction with social service, prior attainment, EAL, and ethnicity), with the plan to use this model to predict participation in both TP and comparison schools. KS2 below 110, being eligible for FSM, being in care for 6 months or more and having SEN are positively and significantly correlated with participation (see Table C7). However, our ability to predict TP participation with the observable characteristics available in the NPD was poor. When testing the predictive power of the model, by comparing predicted TP participation with actual TP participation, 86.54% of those predicted to participate in TP did not actually participate in TP and 16.11% of those predicted not to participate in TP actually participated in TP (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). As indicated in the IPE report, it was likely that selection of pupils for tuition by schools was based on unobservable characteristics, such as their ability to catch up and make good use of tutoring. Hence, we did not proceed with this part of the analysis.

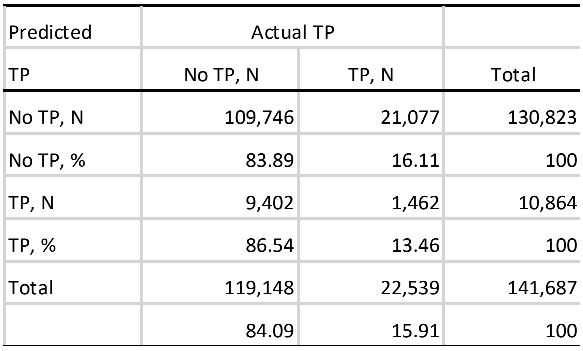
Table C7: Estimation of TP participation using Linear Probability Model and Logit

	Logit			LPM		
	Coef	S.E.	P-value	Coef	S.E.	P-value
Female	0.143***	0.033	0.000	0.019***	0.004	0.000
FSM ever in 6 years	0.737***	0.049	0.000	0.113***	0.008	0.000
master only (1)	0.354**	0.117	0.002	0.053**	0.019	0.004
EAL (no as base):						
EAL: Yes	-0.083	0.048	0.082	-0.012	0.007	0.082
EAL: Missing	-0.145	0.171	0.395	-0.019	0.022	0.376
Ethnicity (unknown as base):	ref.					
White British	0.078	0.090	0.383	0.012	0.012	0.351
Asian	0.112	0.105	0.285	0.016	0.014	0.278
Black	0.118	0.100	0.237	0.018	0.014	0.212
Other ethnicities	0.101	0.091	0.268	0.015	0.013	0.240
SEN (no as base):						
SEN: Yes	0.092*	0.040	0.021	0.013*	0.006	0.021
Looked after since 31 March	0.457*	0.207	0.028	0.088*	0.044	0.046
Looked after for 6 months	-0.943**	0.349	0.007	-0.163**	0.057	0.005
Looked after for 12 months	0.969**	0.308	0.002	0.160***	0.043	0.000
School %FSM high vs low	-0.165	0.098	0.092	-0.023	0.014	0.094
Ofsted 2018 (outstanding as base):						
Good	-0.065	0.100	0.515	-0.009	0.014	0.509
Requires improvement/satisfactory	-0.240	0.132	0.069	-0.032	0.017	0.062
Inadequate	-0.162	0.216	0.452	-0.021	0.030	0.477
Ofsted missing	-0.058	0.274	0.831	-0.008	0.037	0.838
IDACI rank	-0.000	0.000	0.546	-0.000	0.000	0.537
School types (Academy as base):	ref.					
Community school	-0.100	0.157	0.524	-0.014	0.022	0.534
Voluntary aided school	-0.005	0.171	0.978	-0.000	0.026	0.993
Voluntary controlled school	-0.609	0.358	0.089	-0.068*	0.035	0.049
Foundation school	-0.051	0.161	0.751	-0.007	0.023	0.752
City technology	-4.215***	0.319	0.000	-0.226***	0.055	0.000
Free school - Mainstream	-0.069	0.110	0.532	-0.009	0.016	0.548
Special free school	0.325	0.210	0.121	0.050	0.035	0.149
Free school UTC	0.395	0.315	0.209	0.059	0.051	0.242
Free school - studio school	0.516*	0.223	0.021	0.085*	0.041	0.039
Region (East midlands as base):						
East of England	-0.208	0.197	0.289	-0.027	0.026	0.290
London	-0.061	0.168	0.716	-0.008	0.024	0.728
North East	-0.034	0.259	0.895	-0.005	0.036	0.891
North West	0.100	0.168	0.551	0.015	0.024	0.533
South East	-0.207	0.175	0.235	-0.027	0.023	0.249
South West	-0.053	0.181	0.770	-0.007	0.025	0.769
West Midlands	-0.179	0.190	0.347	-0.025	0.026	0.342
Yorkshire and the Humber	-0.250	0.185	0.177	-0.033	0.024	0.180
Region missing	0.315	0.423	0.457	0.049	0.067	0.467
KS2 English (80-90 as base):						
KS2 English between 91-100	0.181***	0.034	0.000	0.026***	0.005	0.000
KS2 English between 101-110	0.195***	0.041	0.000	0.027***	0.006	0.000
KS2 English between 111-120	0.019	0.055	0.737	0.006	0.007	0.398
KS2 maths (80-90 as base):						
KS2 maths between 91-100	0.243***	0.053	0.000	0.035***	0.007	0.000
KS2 maths between 101-110	0.156**	0.060	0.009	0.022**	0.008	0.006
KS2 maths between 111-120	-0.054	0.076	0.478	-0.003	0.009	0.785
Constant	-2.057***	0.228	0.000	0.105***	0.031	0.001
Observations	121860			121860		
R-squared				0.029		
Pseudo R-squared	0.030					

Note: Population of Year 11 TP schools, all pupils. School-level clustered residuals.

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

Table C8: Predictive power of participation model



Predicted	Actual TP		
TP	No TP, N	TP, N	Total
No TP, N	109,746	21,077	130,823
No TP, %	83.89	16.11	100
TP, N	9,402	1,462	10,864
TP, %	86.54	13.46	100
Total	119,148	22,539	141,687
	84.09	15.91	100

Source: Year 11 population data.

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

N, number; TP, Tuition Partners.

Appendix D: Balance tables for additional analysis samples (50% and 70% restrictions)

Table D1: Baseline characteristics of Year 11 TP schools that targeted 50% of PP-eligible pupils and eligible comparison schools before matching

Variable	Means:Eligible comparison	SD:Eligible comparison	Means:TP schools, 50%	SD:TP schools, 50%	Difference	Std Difference
School-level PP KS2 Maths scores 2020/21	101.564	(3.719)	101.479	(2.960)	-0.085	-0.018
School-level PP KS2 Read scores 2020/21	101.021	(3.873)	101.014	(3.211)	-0.007	-0.001
School-level KS4 English, 2019/20	5.192	(0.825)	5.226	(0.728)	0.034	0.031
School-level KS4 Maths 2019/20	5.094	(0.972)	5.107	(0.806)	0.014	0.011
School-level KS4 English, 2018/19	4.811	(0.884)	4.834	(0.749)	0.023	0.020
School-level KS4 Maths, 2018/19	4.790	(1.019)	4.785	(0.817)	-0.005	-0.004
School-level KS4 English, 2017/18	4.809	(0.847)	4.866	(0.743)	0.057	0.050
School-level KS4 Maths, 2017/18	4.785	(0.998)	4.797	(0.797)	0.012	0.009
Total pupil counts	882.278	(335.389)	910.724	(304.753)	28.446	0.063
Pupils-to-teacher ratio 2018	16.298	(2.725)	16.672	(2.723)	0.374*	0.097
Ofsted 2018: Outstanding	0.248		0.272		0.024	0.038
Ofsted 2018: Good	0.496		0.524		0.027	0.039
Ofsted 2018: Inadequate	0.045		0.042		-0.003	-0.011
Ofsted 2018: Requires improvement	0.135		0.110		-0.025	-0.054
Ofsted 2018: Missing	0.075		0.052		-0.023	-0.066
School type: Academy-sponsor led	0.191		0.168		-0.023	-0.042
School type: Community school	0.108		0.099		-0.009	-0.021
School type: Voluntary aided/controlled school	0.070		0.073		0.003	0.009
School type: Foundation school	0.049		0.063		0.014	0.043
School type: Free school - mainstream	0.515		0.529		0.014	0.019
School type: Others	0.067		0.068		0.001	0.003
Urban	0.797		0.812		0.014	0.025
Rural	0.162		0.168		0.006	0.011
Urban/Rural missing	0.041		0.021		-0.020	-0.081
Region: East Midlands	0.090		0.068		-0.022	-0.058
Region: East of England	0.087		0.120		0.033	0.077
Region: London	0.133		0.152		0.019	0.037
Region: North East	0.037		0.042		0.005	0.018
Region: North West	0.106		0.136		0.030	0.065
Region: South East	0.184		0.141		-0.042	-0.081
Region: South West	0.111		0.131		0.020	0.044
Region: West Midlands	0.120		0.084		-0.036	-0.084
Region: Yorkshire & the Humber	0.091		0.105		0.014	0.032
Region: Missing	0.041		0.021		-0.020	-0.081
AM participation	0.000		0.126		0.126***	0.378
Census school-level % FSM Spring 2021	0.255		0.237		-0.018*	-0.093
% EAL	0.143		0.148		0.006	0.022
% SEN	0.216		0.224		0.009	0.066
% Female	0.494		0.526		0.032**	0.113
Average IDACI scores	0.037	(0.023)	0.036	(0.023)	-0.001	-0.039
% White British	0.123		0.119		-0.004	-0.040
% Asian	0.018		0.020		0.002	0.047
% Black	0.015		0.018		0.003	0.071
% Other ethnic	0.026		0.027		0.001	0.016
% Unknown ethnic	0.004		0.004		0.000	0.011
% Not white	0.058		0.064		0.006	0.063
Observations	1,328		191		1,519	

Source: Year 11 population data.

Note: values with less than three schools in a category have been suppressed for SRS clearance.

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

Table D2: Baseline characteristics of Year 11 TP schools that targeted 70% of PP-eligible pupils and eligible comparison schools before matching

Variable	Means:Eligible comparison	SD:Eligible comparison	Means:TP schools, 70%	SD:TP schools, 70%	Difference	Std Difference
School-level PP KS2 Maths scores 2020/21	101.564	(3.719)	101.695	0.130	0.027	0.027
School-level PP KS2 Read scores 2020/21	101.021	(3.873)	101.496	0.475	0.093	0.093
School-level KS4 English, 2019/20	5.192	(0.825)	5.386	(0.643)	0.195*	0.186
School-level KS4 Maths 2019/20	5.094	(0.972)	5.273	(0.723)	0.180	0.148
School-level KS4 English, 2018/19	4.811	(0.884)	4.956	(0.731)	0.144	0.126
School-level KS4 Maths, 2018/19	4.790	(1.019)	4.907	(0.732)	0.117	0.093
School-level KS4 English, 2017/18	4.809	(0.847)	4.947	(0.727)	0.139	0.124
School-level KS4 Maths, 2017/18	4.785	(0.998)	4.873	(0.775)	0.087	0.069
Total pupil counts	882.278	(335.389)	919.284	37.006	0.083	0.083
Pupils-to-teacher ratio 2018	16.298	(2.725)	17.030	0.732**	0.204**	0.204
Ofsted 2018: Outstanding	0.248		0.328		0.125	0.125
Ofsted 2018: Good	0.496		0.522		0.037	0.037
Ofsted 2018: Inadequate	0.045					
Ofsted 2018: Requires improvement	0.135		0.104		-0.066	-0.066
Ofsted 2018: Missing	0.075		0.030		-0.144	-0.144
School type: Academy-sponsor led	0.191		0.134		-0.108	-0.108
School type: Community school	0.108		0.119		0.024	0.024
School type: Voluntary aided/controlled schoo	0.070		0.090		0.051	0.051
School type: Foundation school	0.049		0.045		-0.014	-0.014
School type: Free school - mainstream	0.515		0.567		0.074	0.074
School type: Others	0.067		0.045		-0.068	-0.068
Urban	0.797		0.821		0.042	0.042
Rural	0.162		0.164		0.004	0.004
Urban/Rural missing	0.041					
Region: East Midlands	0.090					
Region: East of England	0.087		0.149		0.136	0.136
Region: London	0.133		0.119		-0.029	-0.029
Region: North East	0.037		0.060		0.075	0.075
Region: North West	0.106		0.149		0.091	0.091
Region: South East	0.184		0.119		-0.127	-0.127
Region: South West	0.111		0.134		0.051	0.051
Region: West Midlands	0.120		0.090		-0.070	-0.070
Region: Yorkshire & the Humber	0.091		0.134		0.096	0.096
Region: Missing	0.041					
AM participation			0.119		0.365	0.365
Census school-level % FSM Spring 2021	0.255		0.205		-0.277	-0.277
% EAL	0.143		0.132		-0.043	-0.043
% SEN	0.216		0.219		0.024	0.024
% Female	0.494		0.548		0.193	0.193
Average IDACI scores	0.037	(0.023)	0.033	-0.003	-0.110	-0.110
% White British	0.123		0.127		0.039	0.039
% Asian	0.018		0.015		-0.052	-0.052
% Black	0.015		0.015		0.002	0.002
% Other ethnic	0.026		0.025		-0.021	-0.021
% Unknown ethnic	0.004		0.004		-0.016	-0.016
% Not white	0.058		0.056		-0.033	-0.033
Observations	1,328		67			

Source: Year 11 population data.

Note: values with less than three schools in a category have been suppressed for SRS clearance.

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

Table D3: Baseline characteristics of Year 11 TP schools that targeted 50% of PP-eligible pupils and matched comparison schools after matching

Variable	Means: Matched comparison	SD: Matched comparison	Means:TP schools, 50% PP	SD:TP schools, 50% PP	Difference	Std Difference
School-level PP KS2 Maths scores 2020/21	101.204	(3.268)	101.479	(2.960)	0.275	0.062
School-level PP KS2 Read scores 2020/21	100.879	(3.283)	101.014	(3.211)	0.135	0.029
School-level KS4 English, 2019/20	5.204	(0.740)	5.226	(0.728)	0.022	0.022
School-level KS4 Maths 2019/20	5.047	(0.794)	5.107	(0.806)	0.061	0.054
School-level KS4 English, 2018/19	4.785	(0.820)	4.834	(0.749)	0.049	0.044
School-level KS4 Maths, 2018/19	4.699	(0.891)	4.785	(0.817)	0.086	0.071
School-level KS4 English, 2017/18	4.811	(0.729)	4.866	(0.743)	0.055	0.053
School-level KS4 Maths, 2017/18	4.727	(0.798)	4.797	(0.797)	0.070	0.062
Total pupil counts	881.827	(327.656)	910.724	(304.753)	28.896	0.065
Pupils-to-teacher ratio 2018	16.555	(2.936)	16.672	(2.723)	0.117	0.029
Ofsted 2018: Outstanding	0.209		0.272		0.063	0.104
Ofsted 2018: Good	0.613		0.524		-0.089*	-0.127
Ofsted 2018: Inadequate	0.031		0.042		0.010	0.039
Ofsted 2018: Requires improvement	0.105		0.110		0.005	0.012
Ofsted 2018: Missing	0.042		0.052		0.010	0.035
School type: Academy-sponsor led	0.157		0.168		0.010	0.020
School type: Community school	0.099		0.099		0.000	0.000
School type: Voluntary aided/controlled school	0.099		0.073		-0.026	-0.066
School type: Foundation school	0.068		0.063		-0.005	-0.015
School type: Free school - mainstream	0.518		0.529		0.010	0.015
School type: Others	0.058		0.068		0.010	0.030
Urban	0.812		0.812		-0.000	0.000
Rural	0.183		0.168		-0.016	-0.029
Urban/Rural missing			0.021		0.016	0.098
Region: East Midlands	0.058		0.068		0.010	0.030
Region: East of England	0.099		0.120		0.021	0.047
Region: London	0.215		0.152		-0.063	-0.115
Region: North East	0.026		0.042		0.016	0.061
Region: North West	0.131		0.136		0.005	0.011
Region: South East	0.131		0.141		0.010	0.022
Region: South West	0.131		0.131		-0.000	0.000
Region: West Midlands	0.105		0.084		-0.021	-0.051
Region: Yorkshire & the Humber	0.099		0.105		0.005	0.012
Region: Missing			0.021		0.016	0.098
AM participation			0.126		0.126***	0.378
Census school-level % FSM Spring 2021	0.247		0.237		-0.010	-0.052
% EAL	0.151		0.148		-0.003	-0.012
% SEN	0.216		0.224		0.008	0.064
% Female	0.502		0.526		0.023	0.084
Average IDACI scores	0.034	(0.020)	0.036	(0.023)	0.002	0.060
% White British	0.112		0.119		0.007	0.067
% Asian	0.016		0.020		0.004	0.077
% Black	0.018		0.018		-0.001	-0.019
% Other ethnic	0.029		0.027		-0.003	-0.072
% Unknown ethnic	0.004		0.004		0.000	0.033
% Not white	0.064		0.064		0.000	0.004
Observations	191		191		382	

Source: Year 11 population data.

Note: values with less than three schools in a category have been suppressed for SRS clearance.

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

Table D4: Baseline characteristics of Year 11 TP schools that targeted 70% of PP-eligible pupils and matched comparison schools after matching

Variable	Means: Matched	SD: Matched comparison	Means:TP schools, 70%	SD:TP schools, 70% PP	Difference	Std Difference
School-level PP KS2 Maths scores 2020/21	101.494	(3.460)	101.715	(3.101)	0.221	0.048
School-level PP KS2 Read scores 2020/21	101.322	(4.471)	101.615	(3.394)	0.293	0.052
School-level KS4 English, 2019/20	5.337	(0.847)	5.393	(0.653)	0.056	0.052
School-level KS4 Maths 2019/20	5.283	(0.927)	5.278	(0.742)	-0.004	-0.004
School-level KS4 English, 2018/19	4.940	(0.904)	4.970	(0.730)	0.030	0.026
School-level KS4 Maths, 2018/19	4.908	(1.021)	4.916	(0.735)	0.008	0.006
School-level KS4 English, 2017/18	4.933	(0.902)	4.985	(0.700)	0.051	0.045
School-level KS4 Maths, 2017/18	4.896	(0.999)	4.901	(0.775)	0.006	0.004
Total pupil counts	846.286	(270.123)	915.460	(272.538)	69.175	0.180
Pupils-to-teacher ratio 2018	17.172	(3.292)	16.961	(2.133)	-0.211	-0.054
Ofsted 2018: Outstanding	0.349		0.333		-0.016	-0.023
Ofsted 2018: Good	0.524		0.524		-0.000	0.000
Ofsted 2018: Inadequate						
Ofsted 2018: Requires improvement	0.111		0.095		-0.016	-0.037
Ofsted 2018: Missing						
School type: Academy-sponsor led	0.143		0.143		-0.000	0.000
School type: Community school	0.111		0.127		0.016	0.034
School type: Voluntary aided/controlled school			0.095		0.079*	0.247
School type: Foundation school					-0.000	0.000
School type: Free school – mainstream	0.651		0.571		-0.079	-0.115
School type: Others						
Urban	0.810		0.810		0.000	0.000
Rural	0.175		0.175		0.000	0.000
Urban/Rural missing					-0.000	0.000
Region: East Midlands	0.063				-0.032	-0.105
Region: East of England	0.190		0.159		-0.032	-0.059
Region: London	0.206		0.127		-0.079	-0.150
Region: North East	0.048		0.048		-0.000	0.000
Region: North West	0.127		0.143		0.016	0.033
Region: South East	0.063		0.127		0.063	0.153
Region: South West	0.095		0.143		0.048	0.103
Region: West Midlands	0.095		0.095		-0.000	0.000
Region: Yorkshire & the Humber	0.095		0.111		0.016	0.037
Region: Missing						
AM participation			0.127		0.127***	0.378
Census school-level % FSM Spring 2021	0.214		0.206		-0.008	-0.046
% EAL	0.130		0.138		0.008	0.035
% SEN	0.219		0.216		-0.003	-0.023
% Female	0.513		0.553		0.040	0.141
Average IDACI scores	0.035	(0.021)	0.032	(0.019)	-0.003	-0.108
% White British	0.129		0.119		-0.009	-0.092
% Asian	0.015		0.016		0.001	0.025
% Black	0.015		0.015		0.000	0.003
% Other ethnic	0.027		0.027		-0.000	-0.010
% Unknown ethnic	0.002		0.003		0.001	0.108
% Not white	0.057		0.058		0.001	0.009
Observations	63		63		126	

Source: Year 11 population data.

Note: values with less than three schools in a category have been suppressed for SRS clearance.

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

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit https://nationalarchives.gov.uk/doc/open-government-licence/version/3 or email: psi@nationalarchives.gsi.gov.uk

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

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn