

EVALUATION OF YEAR 1 OF THE TUITION PARTNERS PROGRAMME: IMPACT EVALUATION FOR PRIMARY SCHOOLS

Evaluation Report

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 (updated in December 2022)

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

Contents

About the evaluator	3
Acknowledgements	3
About the first year of the National Tutoring Programme Tuition Partners	5
About this study	5
About this report	5
Other volumes in the series	6
Executive summary	7
Introduction	10
Methods	15
Impact evaluation results	33
Conclusions	76
References	80
Appendix A	81
Appendix B	86
Appendix C	88

About the evaluator

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 like to thank Greta Morando (formerly at University of Westminster) who carried out the statistical work in the early stages of the project including the initial matching for the primary school analysis.

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.

Update (December 2022)

This version provides an update with respect to the report published in October 2022. The previously published tables (listed below) presented the descriptions relating to all schools that had at least one pupil with a valid end-point assessment, but these same pupils may have had missing baseline data. The tables we now present provide descriptions relating to all schools that had at least one pupil with a valid baseline **and** a valid end-point assessment (that is, schools with at least one pupil with non-missing baseline **and** end-point assessment), which is the sample upon which the analysis had been conducted. For English, the analysed sample of schools is 117 TP (intervention)

schools (previously we had described 119 intervention schools) and 48 comparison schools. For maths, the analysed sample of schools was 84 TP schools and 42 comparison schools (previously we had described 43 comparison schools). All tables now reflect the analysed sample, and are provided in place of:

- Figure 1 (participant flow English, number of analysed schools updated)
- Table 8 (MDES table: only change made to school sample sizes)
- Tables 11a1 and 11a2 (maths unweighted and weighted, respectively)
- Tables 11b1 and 11b2 (English unweighted and weighted, respectively)
- Table 11c (baseline characteristics)
- Tables 12a and 12b (maths and English respectively)
- Tables 13a and 13b (across year groups and by year group, for maths and English respectively)
- Tables 16a and 16b (maths and English respectively).

Updates relate to descriptive information only (rather than outcomes). The main difference in the updated tables is that in the maths sample, the proportion of pupils in Year 2 and the mean number of pupils is significantly different between TP (intervention) and comparison schools in the weighted balance table (Table 11a2). The analysis controls for the number of pupils in the regression, which is the main mitigation when significant differences are observed in the pre-analysis stage.

All the other descriptives are similar to those previously published in the report.

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-eligible) 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 (AM) – 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) at the primary school phase.

This volume outlines the impact of TP on learning outcomes for primary school pupils (Years 1–6), through a number of estimators of impact, in both English and maths. It also reports the findings from the moderator analysis on primary school data, which explores how different models of tutoring (e.g., face-to-face vs online; 1:1 vs small group) correlate with assessment score.

How to cite this volume: Poet, H., Oppedisano, V., Zhang, M., Lord, P., Styles, B. and Dorsett, R. (2022). *Evaluation of Year 1 of the Tuition Partners Programme: Impact Evaluation for Primary Schools. Evaluation Report*. 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.

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 school 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 both to 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 PP 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 previous school year.

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: **Year 11 impact evaluation**
- 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 6000 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 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.

This evaluation report covers the analysis on the impact of the TP programme on the maths and English attainment outcomes for primary school pupils (Years 1–6) using standardised classroom assessments. Separate reports relate to analysis on Year 11 pupils 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 and re-weighting 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 comparison schools. For English, the analysis is based on 165 primary schools with 7073 pupils eligible for Pupil Premium and for maths, 126 primary schools with 5102 pupils eligible for Pupil Premium³. An additional instrumental variable (IV) analysis, based on the sample of TP schools only, looked at the impact of TP in schools that signed up to the TP programme earlier (and that delivered more tutoring) compared to schools that signed up later.

Table 1: Summary of findings

Finding
On average, 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 (no evidence of an effect in English or in maths). This result has a low security rating. A particular challenge is that, on average, only approximately 20% of pupils eligible for Pupil Premium were selected

³ Note Dec 2022: these figures have been updated to describe the number of schools analysed (i.e. schools with at least one pupil with non-missing baseline and end-point assessment data).

for tutoring, meaning a large proportion of pupils eligible for Pupil Premium were included in the analysis who 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.

Similar analysis on all pupils found that pupils in schools that received TP made, on average, similar progress in English compared to all pupils in comparison schools (no evidence of an effect), and an additional one month's progress in maths compared to pupils in comparison schools. However, there is uncertainty around these estimates, with the positive maths result being consistent with a null (0 months) or slightly larger positive effect (2 months) and the English result being consistent with small positive (1 month) or small negative effect (-1 months). Furthermore, this analysis was subject to even further dilution: on average, only 12% (for maths) and 14% (for English) of pupils in the analysed schools were selected for tutoring. Given this context, it is unlikely that any of these differences were due to TP.

In the sample of TP schools, completing a 12-hour block of tutoring (compared to zero hours) was related to higher English scores amongst pupils eligible for Pupil Premium that received more tutoring due to the early sign-up of the school. An equivalent analysis for maths was not able to proceed.

A different analysis within TP schools showed that pupils who received more hours of tutoring were associated with higher English scores on average than pupils who received fewer hours of tutoring. However, this was not the case for maths, where receiving more hours of tutoring was not associated with higher maths scores. These results are associations and are not necessarily causal estimates of impact; there may be other explanations for the results.

EEF security rating

These findings have a low security rating. This study was well powered but, by necessity, did not have a randomised trial design; given the urgency of the requirement for catch-up support in schools it was not considered ethical to randomise. There was high attrition, with a proportion of schools not conducting, submitting, or having access to English and maths assessment data for all pupils – but after weighting TP and comparison schools were well balanced on observable characteristics. It was also harder for the evaluation to detect whether there was an impact of the programme because almost four-fifths of the pupils included in the analysis did not receive tutoring, and those pupils who received tutoring received fewer hours on average than expected.

Additional findings

When looking at how outcomes varied for pupils who received TP by model of tutoring, this indicated that, for English, scheduling TP sessions to all take place within school hours is associated with better scores than sessions delivered in a combination of both during and outside schooling hours, and also that sessions attended with at least one other pupil were associated with better scores. For maths, group size and timing of delivery were not associated with scores. However, delivery of tutoring sessions concentrated over a short timeframe was positively correlated with higher maths scores. At a tutor level, tutors who received on-going tutor training were associated with higher scores in maths. The results also suggested that primary school pupils responded better to tutors with undergraduate qualifications in maths, and with PGCE/QTS in English, rather than other postgraduate qualifications. These results are associations and are not necessarily causal.

The evaluation also contended with the challenges of the pandemic, meaning not all planned analyses could go ahead. The Year 6 analysis using all pupils in the year group was not possible due to the cancellation of the Key Stage 2 statutory tests for summer 2021. The evaluation also aimed 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. By doing this, the outcomes across both groups of 'predicted' participants could be compared. 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.

This study had several related limitations: 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 comparison 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 who did not all receive 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 main

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 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-point assessment than had been anticipated (at a pupil-level average, for PP-eligible pupils, 8.8 hours in English and 8.9 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.

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.

Impact

Table 2: Summary of impact on RQ1a and RQ3 (impact of TP availability on PP-eligible/all pupils' attainment)

Outcome/ Group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No of pupils	P Value
maths assessment (PP-eligible pupils: RQa1)	-0.026 (-0.030 to 0.079)	0	1	5,102	0.631
English assessment (PP-eligible pupils: RQa1)	-0.024 (-0.100 to 0.052)	0	1	7,073	0.533
maths assessment (all pupils: RQ3)	0.050 (-0.03 to 0.133)	1	N/A	15,393	0.236
English assessment (all pupils: RQ3)	0.001 (-0.063 to 0.065)	0	N/A	21,879	0.982

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 (Cullinane and Montacute, 2020; EEF, 2020; Sharp *et al.*, 2020; UCL, 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 Partner (TP) programme (which provided tutoring support to pupils), and Academic Mentoring (AM), 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 in 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 access⁴).

Intervention

This evaluation is on year 1 of the TP programme, which is summarised below using the EEF's TIDieR⁵ framework:

- **Why:** Research shows that pupils' learning has been affected by the 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.

⁴ 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

⁵ TIDieR stands for Template for Intervention Description and Replication.

- **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 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 timeframe 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).

In this primary school evaluation, the biggest change to the design was the removal of analyses on Year 6 tests because the statutory testing was cancelled due to Covid-19 for summer 2021. As a result, this report outlines the results of the analysis on the recruited sample of primary schools. The analysis reported here focuses only on pupils in Years 1–6, using standardised assessments as outcomes. A challenging recruitment period meant that the comparison group is formed of different types of comparison school to originally intended, and the sample sizes were adjusted down – more information can be found later in this report and in the study plan.

We had also originally planned to conduct analysis on an evaluation sample of secondary (Years 7–11) schools using a similar design to that reported here for primary schools. Unfortunately, the sample of secondary schools we managed to recruit was small and it was decided not to proceed with this option. The evaluation of tuition in Year 11 was able to proceed and those results are reported separately in an **Y11 evaluation report**.

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, and in both primary and secondary schools (the latter, in Year 11 only). This report contains the findings from the analysis on attainment outcomes for primary school pupils.

The research questions (RQ) outlined in the study plan which formed the impact evaluation are listed below in Table 3, with the RQ numbering for this primary school impact evaluation report, and for the Year 11 impact evaluation (reported separately).

One of the research questions (RQ1a) focuses on all PP-eligible pupils in the year groups involved, 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 RQ3) 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 3: Summary of research questions (RQs) by volume

	Primary school impact evaluation RQ number (This report)	Year 11 impact evaluation RQ number
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? ⁶	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.1⁷). Schools could sign up to become a Research Champion by completing an additional form online to confirm their eligibility (see Appendix A.2).

All pupils in the participating primary schools took part in the evaluation activities with informed opt-out consent via a letter to parents.

All participants (parents, tutors, school staff and TP staff) were provided with a privacy notice relevant to processing their (or their child's) 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.

⁶ 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.

⁷ Appendices numbered A.x can be found in the separate 'Impact Appendices' document

The University of Westminster (UoW) identified the following legal basis:

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

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

A separate legal basis is identified for processing special data. The legal basis for processing special data for the evaluation of 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 TPs.

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 TPs 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 TP 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 TP 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 team from the UoW did not have access to any identifiable data.

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

- **Safe people:** all researchers accessing the project's data via the SRS are Accredited Researchers and hold a 'basic disclosure' certificate that is no more than 2 years old.
- **Safe projects:** the project meets the conditions for accessing personal-level data. A full request to the NPD team was submitted, outlining the appropriate and ethical use of the data, and the public benefit of the research (to contribute to the evidence base on tutoring, and inform future tutoring programmes). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence about the most effective ways of providing catch-up tuition. The evaluation could be done without processing personal data, but processing does not override the data subject's interests.
- The research team and the EEF were committed to publishing the results of the study.

- Safe settings: all researchers working on the NPD data only accessed the data via the SRS secure online system. Our organisations obtained safe room connectivity/homeworking agreements to have SRS remote connectivity access.
- Safe outputs: All outputs were checked by the ONS team to ensure that the outputs 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 the 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 recruitment of schools to the evaluation sample and 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 overview

Table 4: Evaluation overview

Design	Matching/weighting Instrumental variables	
Unit of analysis	Primary school pupils eligible for Pupil Premium (PP, identified as pupils eligible for FSM in the previous six years, NPD variable: EverFSM6) in Years 1–6 ⁸	
Number of units to be included in analysis (intervention, comparison)	165 English primary schools (117 intervention, 48 comparison schools); 126 maths primary schools (84 intervention, 42 comparison schools) For primary English: 7073 PP-eligible pupils (4274 from intervention schools and 2799 from comparison schools) For primary maths: 5102 PP-eligible pupils (3032 from intervention schools and 2070 from comparison schools)	
Outcome (all RQs)	Variable	Attainment in English and maths
	Measure (instrument, scale, source)	Standardised assessments in English and maths in summer term 2021
Secondary outcome(s)	Variable(s)	n/a
	Measure(s) (instrument, scale, source)	n/a
Baseline for outcome (all RQs)	Variable	Attainment in English and maths
	Measure (instrument, scale, source)	Standardised assessments in English and maths in autumn term 2020
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 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 recruited to the TP intervention 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 Miratrix (2020) consider, providing evidence that simple matching approaches may work well for this purpose. The counterfactual is

⁸ 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).

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. We checked if the schools in the evaluation sample received tutoring from an Academic Mentor, but none did. No other forms of tutoring could be controlled for. Moreover, challenges to the recruitment meant that we also had to recruit non-matched comparison schools. Hence our ability to control for school-level selection was fully reliant on subsequent re-weighting of the achieved sample rather than on re-matching.

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 6082 schools signed up to the programme in its first year (according to the monitoring data provided by TPs). Schools could choose which TP(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.

TP focused on supporting disadvantaged pupils, including those eligible for PP. Participating schools were able to use their discretion 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 PP 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 pilot⁹ and the guidance (above) for schools. However, as noted in more detail in the *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. Note, we identified PP-eligible pupils as pupils eligible for FSM in the past six years in the NPD.¹⁰

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

Study participants and inclusion criteria

In order for a school to be eligible for the TP intervention evaluation sample (and receive the incentives listed in Appendix A of the study plan), the school must have had at least one school year in which pupils were selected for tutoring *and* in which standardised assessments were being conducted in the tutoring subject (maths/English); and agreed to provide the pupil data needed for the evaluation.

Schools needed to have conducted a standardised assessment at baseline, prior to tutoring¹¹ and plan to conduct another assessment in the summer 2021 term, after tutoring was expected to be completed (see the *Outcome measures* section).

Decisions around selection and recruitment were informed by practical considerations, as well as the methodological design. We recognised that this study was conducted in atypical circumstances (the ongoing Covid-19 pandemic) and we were conscious of the pressures and challenges that schools were facing. This was one of the main drivers of our decision to use standardised test data that was already being administered by schools, so as not to place additional burden on pupils and schools.

It was not possible to randomise, withhold or restrict access to the tutoring for schools and pupils that wished to receive it due to the need to maximise reach to as many schools and pupils as possible.

⁹ In the online tutoring pilot in the summer term of 2020 that preceded TP, over 60% of targeted learners were PP-eligible (Marshall *et al.*, 2021).

¹⁰ PP-eligibility for PP 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.

¹¹ Originally this was only tests in autumn 2020; however, as the timeline shifted, we were also able to accept baseline tests that were completed in the spring 2021, so long as it was prior to the tutoring starting.

Recruitment to the evaluation sample intervention group

The TP programme was launched on 2nd November 2020, after which date TPs could sign schools up to receive the subsidised tutoring. In order to take part in TP, schools had to sign a memorandum of understanding (MoU). Once signed up, schools were contacted by the Evaluator to invite them to take part in the impact evaluation as a sampled school. Initially, only schools that had signed an MoU up to 22nd January 2021 were invited to sign up to participate in the impact evaluation. However, in response to the national lockdown period and changes to delivery, we subsequently expanded the recruitment and contacted all schools participating in TP to take part in the evaluation sample, regardless of the date they signed up to the programme.

Schools in the evaluation sample were asked to agree to provide access to their standardised assessment data and provide additional pupil details and tutoring attendance data. To do this, schools completed an online form (see Appendix A.2) which collected information about which standardised tests (provider, subject and timing/version) the school was using in the 2020/21 academic year and for which year groups, if any.

Recruitment of TP schools to the intervention group was closely monitored to ensure that the schools signed up to the impact evaluation would include a spread of key characteristics, reflecting the TP population signed up at that point.¹² However there were a number of challenges to recruitment including the ongoing disruption caused by Covid-19 which affected schools' willingness to participate, and disruptions to the testing and tutoring schedules which affected the eligibility for the evaluation and reduced the potential recruitment pool. Consequently, our recruitment methodology evolved into a more pragmatic and sensitive approach in response to the challenges and additional burden faced by schools during the spring 2021 term. Furthermore, we needed to be responsive to the rate of school sign up to the TP programme itself. These challenges meant that if schools met the key testing and delivery criteria, they were accepted into the evaluation sample.

The key characteristics monitored for high/low coverage were:

- Phase (primary/secondary)
- Subject of tutoring (maths/English (reading))
- Assessment provider/use of standardised assessments in 2020/21
- Pattern of year groups/subjects tested

Schools signing up to the evaluation sample were able to call themselves TP School Research Champions, to emphasise their additional role. In the report we refer to these schools as the TP intervention group/sample.

Inclusion criteria for comparison schools

The eligibility requirements around testing patterns for the comparison schools mirrored those of the TP intervention group: only primary schools with a testing regime in place were recruited in the comparison school sample. This meant that comparison primary schools needed to have done the baseline assessment at the time of recruitment and planned to conduct a standardised assessment in the summer 2021 term.

Selection of the comparison group and identification assumptions

Initial recruitment of matched comparison sample

We aimed to recruit a sample of comparison (i.e., non-TP) schools that shared similar characteristics to the TP intervention schools recruited to the evaluation sample. To do this, we selected all primary schools in England that were either Academies, Colleges, Free Schools, or Local Authority maintained schools, and excluded independent schools and Pupil Referral Units. For each TP school signed up to the evaluation by 29th March 2021, we identified eight similar-looking comparison schools using propensity score matching (based on characteristics listed in the *Propensity score matching* section in the *Statistical analysis* section), with the aim of recruiting one of them as a comparison school.

¹² Recruitment to the impact evaluation took place while schools were still signing up to TP programme to receive tuition. Therefore, it was not possible to ensure representativeness of the TP school population as the population was not yet defined.

Schools not participating in TP by end-March 2021 were viewed as non-TP schools and included in the pool for the match. Note that checks on actual participation in TP were conducted for each comparison school recruited.

We contacted the pool of potential matched comparison schools and invited them to participate in the evaluation with details of the incentives available (see Appendix A of the study plan). Schools were contacted by email and telephone, with a number of attempts to follow up. Interested schools were asked to complete an adapted version of the online form used for the intervention sample, but with fewer questions about participation in TP (Appendix A.2). However, the challenges and additional burden faced by schools during the spring 2021 term affected the recruitment of comparison schools, making it difficult to recruit one matched comparison school for each TP intervention school. Comparison schools had lower vested interest in the intervention and therefore seemed less prone to engage with the research. We adjusted our recruitment approach and opted for a more pragmatic one by looking for schools that may act as a good comparison in other ways; that is, in terms of their interest in the programme.

Expanded eligibility for the comparison group (EOI and MOU schools)

In addition to the matched comparison sample described above, we also approached two further groups, similar to TP schools in terms of motivation to participate in the TP programme:

- schools that completed an Expression of Interest (Eoi) form on the NTP website but that did not go on to sign up (via an MoU) to the programme.
- schools that signed an MoU but that did not go on to start tuition or started tuition after the end-point assessment.

These schools could act as control as they showed a similar interest in TP to schools in the intervention. However, these schools may be a different group of schools in terms of characteristics and motivation to the matched comparison group. Schools in these groups were approached in the same way as that described for the matched comparison group. The same data was collected from these schools. Schools from these 'TP interested' groups were included in the final analysis in the comparison group.

We asked all recruited comparison schools to provide information about their PP-eligible pupils,¹³ at a pupil level, regarding whether they received any formal tutoring during the 2020/21 academic year (arranged by the school). Six schools in the comparison sample said their school participated in TP. These schools had signed an MoU. However, all tutoring information (start date, sessions bought and completed, tutor specific information) for these schools were blank and the assessment scores not valid. Hence, they are not included in the analytic sample.

¹³ For PP-eligible pupils in year groups that have completed the standardised assessments.

Summary of the composition of the comparison group

Table 5 below shows the composition of the comparison sample according to the routes through which schools were recruited: slightly more than 50% of pupils in the maths and English sample are in schools from the matched comparison list, slightly less than 50% signed an EoI or signed an MoU.¹⁴

Table 5: Composition of comparison maths and English samples in terms of recruitment routes

Primary schools Evaluation sample for maths, of pupils in non-TP schools	N pupils	N schools	% N of schools
Matched schools	1329	24	57.14
EoI, or MoU not delivered TP	741	18	42.86
Total	2070	42	100.00
Primary schools Evaluation sample for English of pupils in non-TP schools	N pupils	N schools	% N of schools
Matched schools	1683	26	54.17
EoI, not delivered TP	1116	22	45.83
Total	2799	48	100.00

Sample: Comparison primary schools.

¹⁴ School level numbers are not separately disclosed for schools that signed an MoU and schools that signed an EoI because in one of the two categories they fall below the SRS school-level count threshold.

Outcome measures

For the analysis reported here, we used standardised assessment data for both the baseline and outcome measures. This was selected for a number of reasons. Mindful of the challenges faced by schools, we aimed to reduce the burden of their participation in the evaluation by using assessment data that primary schools were collecting anyway in 2020/21. We also knew that many primary schools already have a preference over which test(s) to use with their pupils, and we wanted to avoid disrupting their existing plans for testing. Other factors that influenced this choice was that it was intended that tutoring would be delivered to any year group (except Reception) and not only those expected to take national curriculum tests. A third reason we planned to utilise existing standardised test data was because Key Stage (KS) 2 tests (and GCSEs) were at risk of disruption due to the pandemic which, as described above, did indeed happen. Of specific relevance to the primary school evaluation, the KS2 tests were cancelled.

The tests used in this analysis were in maths and English (reading), and, as noted above, were the tests that schools had already planned to administer as part of their assessment and monitoring processes. The tests were therefore administered in schools by their own school staff. As we use test data that schools were collecting anyway for their own purposes of monitoring progress and identifying learning gaps, we have no reason to suspect that the tests were administered in a biased way.

The test providers, the names of the assessments and their published means and standard deviations are shown in Table 6 (for more information about the tests used, please see Appendix A.5). All of the providers of standardised assessments publish the means and standard deviations from when the tests were standardised with a nationally representative sample. In each case, the following formula was applied to bring each test onto the same scale:

$$\frac{x - \mu}{\delta}$$

where x is the standardised score, μ is the population mean (or standardisation mean) and δ is the population standard deviation (or standardisation standard deviation).

We aligned the scales using the tests' national standardisation parameters so that we could analyse the standardised scores from different tests together, thus allowing impact to be measured across all year groups in each phase simultaneously.

In addition to the four main test providers, we also accepted data from schools that had used Year 6 SATs (end of KS2 assessments) from previous academic years. As the relevant standardisation parameters were not available for these tests (STA does not publish descriptive statistics for the population (e.g., means and standard deviations)), we calculated the mean and standard deviation for the data collected (Table 6) and included a flag for test type in the models.

Table 6: Standardised tests¹⁵ used as part of the evaluation

Assessment provider	Assessment name	Score type	Expected mean	Expected SD
GL Assessment	Progress Test in English	Standardised score	100	15
	Progress Test in Maths	Standardised score	100	15
NFER	NFER Reading	Standardised score	100	15
	NFER Maths	Standardised score	100	15
Renaissance Learning	STAR Maths	Normed referenced standardised score (NRSS)	100	15
		Standardised score	100	15
		Normal curve equivalent (NCE)	50	21.06
	STAR Reading	Normed referenced standardised score (NRSS)	100	15
		Standardised score	100	15
		Normal curve equivalent (NCE)	50	21.06
Rising Stars	New PIRA	Standardised score	100	15
	New PUMA	Standardised score	100	15
	NTS:Maths	Standardised score	100	15
	NTS:Reading	Standardised score	100	15
	PIRA	Standardised score	100	15
	PUMA	Standardised score	100	15
Old SATs (past papers)*	Old SATs (past papers)	Scaled score ¹⁶	–	–

* As noted above, the expected mean and SD of KS2 assessments are not published by STA. The actual mean of the SATs test in the data collected was 102 and the SD was 9.

Baseline measures

The baseline measures are the standardised tests in maths and English (reading) taken by pupils before the start of TP tuition in the school.

Primary outcome

The outcome measures are the standardised tests in maths and English (reading) taken by pupils in the summer term 2021. We requested that, where possible, schools administer the tests after tutoring had been completed, but given the complexity of delivering normal lessons and tutoring after the lockdown in spring term 2021, flexibility was needed, and we accepted what schools could provide. Schools with pupils tutored before the end-point assessment date were eligible to be TP intervention schools in the analysed sample. We therefore checked the timing of the tuition in relation to the timing of the assessments in recruited schools in order to determine which group they were assigned to (intervention or comparison).

Collection of the baseline and outcome (assessment) data

Once permissions were given by the school for the assessment results to be shared, we either accessed the results securely and directly from the assessment provider, or the school shared the data using the provided template via the NFER secure portal. Intervention schools were asked to provide scores for the entire year group, where at least one pupil was selected for tutoring. In the case of comparison schools, we requested test scores for the whole year group where this data was available. Where a school provided the raw score, we used the lookup tables supplied by the assessment provider to convert it to the standardised score.

In the case of both the baseline and the outcome (end-point) test, schools were only able to send us what they had. Not all schools tested (or tutored) in both subjects.

¹⁵ We gave schools the opportunity to tell us about other tests they were using, in order to see if there were any other standardised tests we could include in the analysis. However, no additional standardised tests were mentioned by participating schools.

¹⁶ <https://www.gov.uk/government/publications/2019-scaled-scores-at-key-stage-2>

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.¹⁷ 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 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 fifth¹⁸ of all pupil records were lost during the match and these pupils are therefore not included in the analysis.

Sample size

Sample size calculations: original assumptions

We used cluster randomised trial power calculations to provide an indication of the MDES for RQ1a¹⁹. We allowed for clustering of pupils within schools. The evaluation was not based on a randomised intervention, but instead relied on a quasi-experimental approach.

During the design phase, sample-size calculations were based on an average of 24 PP-eligible pupils²⁰ per primary school. Note that the intention to treat effect is more diluted, the lower the percentage of PP-eligible pupils per school who were selected for the intervention. Since our analysis focused on disadvantaged pupils, we do not produce separate estimates for all pupils.

We assumed an ICC of 0.15 and pre-post correlations of approximately 0.7. This is within the range of assumptions made in power calculations for other tutoring evaluations (e.g., the EEF evaluations of Affordable Maths, Catch-up Numeracy, Tutor Trust) and of secondary school intervention such as Embedding Formative Assessment.

Sample size calculations: numbers

Sample size calculations are shown in Table 8. The sample sizes we originally planned to achieve were based on pupils with: (1) baseline and (2) end-point assessment, and for whom we had (3) the required pupil data (for matching to tutoring attendance data). Based on the three data points, we originally intended to over-recruit schools by 20%, resulting in targets of:

- 478 primary schools:
 - 239 primary schools (120 intervention and 119 comparison) for English
 - 239 primary schools (120 intervention and 119 comparison) for maths

¹⁷ More information about the data collected, and the process of collecting it can be found in the IPE report, see section *Research methods* in the *IPE methods* section.

¹⁸ 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.

¹⁹ MDES are computed also for studies not based on a randomised intervention, such as QEDs.

²⁰ This assumption (average cluster size) was 19 PP-eligible pupils per school in the first version of the study plan as we assumed that there would be a relatively low or incomplete coverage of PP in the year groups participating either in terms of tutoring or testing. The assumption was revised to 24 PP-eligible pupils per primary school based on an extract of the TP pupil data in March 2021. This was subsequently revised again during the analysis stage, based on the actual data analysed, as explained at the end of this subsection.

The estimated MDES in the study plan was 0.11 for English and 0.115 for maths.²¹

The recruitment of the comparison sample was slower than required. As an initial step we reviewed the MDES calculations using the sample of schools recruited to the evaluation that shared pupil data by the end of August. The achieved sample sizes were the following at the end of August:

- 328 primary schools:
 - 174 primary schools (120 intervention and 54 comparison) for English
 - 154 primary schools (105 intervention and 49 comparison) for maths

The sample sizes used in the analysis stage (after the data had been cleaned and matched to the assessment data) are as follows:

- 291 primary schools²²:
 - 165 primary schools (117 intervention and 48 comparison) for English
 - 126 primary schools (84 intervention and 42 comparison) for maths

Upon checking the actual data collected and cleaned for analysis, the cluster size increased to 43 PP-eligible pupils per cluster in English and 42 in maths (from 24 in the study plan v.2 assumptions). The MDES for the achieved samples are 0.125 for English and 0.125 for maths. These are well within the effect sizes seen from small-scale tutoring trials. However, the calculations do not take into account the level of dilution that results from relatively low levels of PP-eligible pupils being selected for TP. If only a percentage 'X' of PP-eligible pupils actually receive the intervention, the MDES for those treated should be multiplied by 1/X.

²¹ In the first version of the study plan the MDES was 0.13 for both English and maths; this was revised based on updated information for the second version of the study plan, as described in this section.

²² Note Dec 2022: these figures have been updated to describe the number of schools analysed (i.e. schools with at least one pupil with non-missing baseline and end-point assessment data).

Statistical analysis

Propensity score matching

We used matching to control for school selection into TP by constructing a comparison group of non-TP schools that was similar in important, observable respects to the TP schools in the TP intervention sample. As explained in the *Recruitment* section, the challenges and additional burden faced by schools during the spring 2021 term made it difficult to recruit all required schools from the lists of matched ones and so we also recruited schools similar to TP schools in terms of motivation to participate in the NTP.

To create a sample of comparison schools, we applied the matching procedure laid out in the study plan and used the variables listed in Table 5 of the Study Plan, listed below for reference and in Appendix B.²³

1. Maths, Read and GPS KS2 in 2017/18²⁴
2. KS1 to KS2 value added attainment, at district level, in 2018/19.²⁵
3. Management = primary – Community, Academies, Foundation, Free Schools, Sponsored Academies, Voluntary schools.
4. School size, total number of pupils in 2018/19.
5. Ofsted, overall effectiveness, in 2018/19.
6. Urban/rural area.
7. Region (London, Government Office Region, and regional dummies).
8. Income Deprivation Affecting Children Index (IDACI) quintile, 2018/19 and interaction of IDACI tertiles with average attainment in the previous year.
9. Free School Meals (FSM) – percentage eligible in 2018/19.
10. English as an Additional Language (EAL) – percentage in 2018/19.
11. Special Educational Needs (SEN) – percentage in 2018/19.

In our case, the matching approach was complicated since it formed part of the recruitment strategy. It was carried out as a means of identifying the sample and therefore was prior to outcomes being known. The procedure, detailed in the study plan, identified several potential samples of comparison schools. Once the recruitment was completed, we did not re-match the samples, but instead checked the quality of the match using balance tables and the preliminary analysis outlined below.

To create the initial pool of potential matched comparison schools for recruitment we 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. The probability of treatment was estimated on the basis of the schools' characteristics listed above. Each treated unit was then matched without replacement with eight comparison units with the closest possible propensity score. We did not impose any caliper nor common support due to challenges in recruitment, which meant we needed to expand the pool of potential comparison schools.

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.

²³ The matching variables listed in Table 5 of the Study plan were slightly different from those listed in version 1 of the Study plan and are explained in page 46 of the final Study plan.

²⁴ 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. For Maths, Read and GPS KS2 the most recent year is 2017/18.

²⁵ In the study plan, we originally listed KS1 attainment. However, the variable was not available and we replaced it with KS1 to KS2 value added attainment.

Preliminary analysis: Weighting methods and placebo tests

This method used weighting after matching and recruitment, to control for school-level selection and bring the TP intervention and the comparison sample to balance. Weights are applied separately to the English and maths samples.

As outlined in the *Selection of the comparison group and identification assumptions* section, the aim of recruiting a sample of comparison schools that looked similar was abandoned in favour of simply recruiting a sufficient number of comparison schools, selected according to different criteria. Having a comparison group of schools that is closer to a convenience sample means that the school-level selection was fully reliant on subsequent re-weighting of the achieved sample. To address this, we constructed weights to bring the achieved comparison sample into line with the TP intervention sample in terms of observable characteristics. We constructed these weights using school-level variables to more tightly control for differences between the TP and comparison schools.²⁶

To limit sample loss in constructing the weights, we zero-imputed missing covariates and created missing dummies for those covariates.

Compared to the list in the study plan that identifies the list of variables used for matching at the initial stage, we used a slightly different list of variables to re-weight the samples to improve the quality of the weighting:

- School size, total number of pupils in previous year, in quartiles (planned as total number of pupils).
- Teacher-pupil ratio in 2018/19, in quartiles.
- Income Deprivation Affecting Children Index (IDACI) in levels, 2018/19 (planned in quintile and the interaction of IDACI tertiles with average attainment in the previous year).
- Lagged KS2 maths and English scores (planned as percentage of pupils who achieved expected standard in KS1 in 2018/19).
- Baseline standardised assessment instead of the percentage of pupils who achieved expected standard in KS1 in 2018/19 and KS1 to KS2 value added attainment, at district level, in 2018/19.

We applied two estimation approaches in pre-TP years as a form of placebo test and compared the approaches to determine how to proceed:

- a. inverse probability weighting, using school-level weights
- b. entropy balancing, using school-level weights.²⁷

These approaches are described in more detail below.

In the inverse probability weighting approach, the group of comparison units was used, but each observation was weighted. Inverse probability weights are based on propensity score, as shown below:

$$W_t = 1, W_c = \frac{PS}{1 - PS}$$

where W_t is the weight for a treated unit, W_c is the weight for a comparison unit, and PS is the propensity score.

Similarly, entropy balancing calculated weights. In this case, the weights are adjusted for known sample distributions, integrating covariate balance directly into the weights. Balance was introduced on the first (mean) moments of the covariate distributions only, due to the small sample size, and the procedure was set to iterate repeatedly until the variance of the weights could not be reduced further without undermining the balance constraints.

Entropy balance provides an alternative to inverse probability weighting. An advantage of entropy balancing over inverse probability weighting is that it leads to perfect balance; inverse probability weighting typically balances the data only approximately (unless the balancing problem is very simple). Perfect balance means that modelling the outcome (e.g.,

²⁶ Data used for the preliminary analysis are school-level data; we cannot therefore use pupil-level data to compute pupil-level weights.

²⁷ The preliminary analysis in the TP intervention sample data is run on aggregated school-level data. We do not have pupil-level data for KS2 in previous years, and thus weights can only be computed at the school level.

using regression adjustment) after the data have been balanced will lead to no refinements in the treatment effect estimate, implying that entropy balancing has the ‘doubly robust’ property (also see Zhao and Percival, 2017).

We simulated impacts on KS2 in pre-TP years using school-level variables to understand the sensitivity of the estimators and assess whether the approach was likely to be sufficiently sensitive (results reported in the *Preliminary Analysis* section). The chosen approach was the one that scores high on (a) not significant impact of TP in pre-intervention years, and (b) small size of the coefficient of interest, as this would indicate similar intervention and comparison samples before the intervention. The outcome of this analysis indicated entropy balancing was the approach that scored higher on (a) and (b).

Further preliminary analysis

To test whether propensity score matching controls for school selection, we initially conducted placebo tests for the TP intervention schools in the three years prior to the intervention using the data from the NFER Register of Schools.²⁸ The placebo test was conducted in March 2021, when we drew the lists of potential comparison schools. They showed similarity between intervention and matched comparison samples. (See the *School-level selection* section in the *Results*.)

However, because the achieved recruited sample of control schools was not only drawn from the list of matched schools, we performed another placebo test after completing the recruitment of the comparison schools and after we had finalised the sample used for the analysis. The placebo test was used to check the similarity of the recruited TP intervention and comparison schools (the comparison group including a combination of schools identified through matching, schools that completed an EoI or that signed a MoU) before the intervention. The check was conducted at the analysis stage in March 2022. The placebo testing was done for the following three preceding years: 2017/18/19,²⁹ using results for KS2 to demonstrate similarity of the achieved match in the years prior to the intervention.

We assessed the performance of the match by (a) comparing observed characteristics of TP schools and their recruited comparators, weighted and non-weighted (Tables 11a, 11b and 11c); (b) comparing baseline assessments of TP schools and their comparison schools (Tables 12a and 12b). This is a school-level analysis, but outcomes are considered for PP-eligible pupils.

Analyses: Overview

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 analysis (RQ1a), there was no single primary outcome identified by design.

The first RQ 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. Whilst 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’ amongst 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: weighting/regression and instrumental variable (IV) regression.

²⁸ The Register of Schools contains information about characteristics of schools from sources including [.gov/Get information About Schools](#).

²⁹ 2020 school-level data was not available at the time we shared the data with the NPD.

Outcome analysis RQ1

Weighting/Regression (RQ1a): 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 at least one year group in the school was selected to receive TP) and (ii) a categorical variable measuring the 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.

For the analysis, all TP sessions taken after the summer assessment date are not considered in the dosage analysis as they took place too late to influence the summer assessment. We controlled for the baseline measure of the outcome of interest. All school-level variables listed in Appendix B (variables used for entropy balance) were used as controls. Pupil-level controls included background variables, such as gender, year group, ethnicity, EAL, special educational needs and looked after children. They are also listed in Appendix B. Residuals were clustered at the school level to account for any common school-specific unobservable component. Regression was based on PP-eligible pupils in the TP intervention schools and their comparison schools, with and without the weights derived using entropy balance. The software used to run the model was Stata.

The coefficient on the TP indicator represents the estimated treatment effect, on an 'intention to treat' basis.

Each estimator has two outcomes of attainment, maths and English (reading). We originally planned to adjust for multiple testing using the Romano–Wolf (2005a;b) simulation approach, as implemented by the Stata program `rwolf` ado (Clarke, Romano and Wolf, 2020). However, the Romano–Wolf correction does not allow the inclusion of weights and does not support the multiple imputation (MI) of missing baseline assessments. In addition, the correction requires the same set of covariates and the same sample for both English and maths assessments (i.e., only schools that tested in both maths and English). In relation to the dosage analysis, the correction can only be applied within a treatment, and not across multiple treatments; hence it cannot be presented. These caveats lessen the value of this adjustment therefore we do not present the correction.

Instrumental variables (RQ1c): What is the impact of the intensity of TP (dosage) on the attainment of all PP-eligible pupils?

We used a second technique of IVs to provide estimates of TP that do not rely on the selection on observables assumption. 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 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 at the time of the summer 2021 assessment. 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 amongst 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 TP completion. In practice, we used the number of days passed between signing the MoU and the time of assessment (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 (compared to 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 assessment.

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, 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: the baseline 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 PP 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 22% of pupils in the maths sample of TP schools and for 38% of pupils in the English sample of TP schools.

Further analyses

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

Our approach to the above outcome analysis (RQ1a) provides an estimate of the impact on a subgroup of the eligible population, PP-eligible pupils, which does not coincide with the entire group of children who were selected to receive the intervention. RQ2 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, special educational needs, interaction with social service, prior attainment, English as first language 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.

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

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 pupils (rather than PP-eligible pupils or predicted TP pupils). Similar to RQ1a, we regressed the pupil-level outcome on two measures of TP – this time for all 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 RQ1a, 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 RQ1a estimates, as the TP impact was more diluted when considered across entire year groups with at least one pupil in receipt of tuition because the percentage of all pupils receiving TP in TP intervention schools would be lower than the percentage of PP-eligible pupils receiving TP. RQ3 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 effect of spillover (peer) effects that may have occurred. The purpose of this is to capture the overall impact of TP, though given that only a small proportion of each class was treated, it should be borne in mind that most pupils in TP schools were 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 controlled for the same school-level and pupil-level characteristics mentioned in RQ1a (PP-eligible pupil 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 including interaction terms between TP availability and the following variables:

1. School characteristics: Ofsted rating (high, outstanding and good vs low, inadequate and requires improvement); proportion of FSM (high vs low, defined on the basis of the median); type of school (academy/maintained); school size (by quartile).
2. Pupil characteristics: prior attainment; SEND vs not; Year group; English as an additional language, ethnicity and gender.³⁰
3. Other: geography (urban/rural; low/high IDACI).

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

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

A descriptive analysis (using the participation data collected from TPs) compared outcomes associated with different tutoring models and moderators among TP intervention schools. We did not propose any impact analysis within RQ6 since we could not observe the counterfactual treatment model among comparison schools. Instead, this element of the analysis summarised mean attainment among participating pupils in TP intervention schools according to the model of tutoring they experienced. Hence the coefficients may reflect how pupils were selected to receive different models 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 intervention schools and TP participants only to assess heterogeneity. In particular, we looked at the following variables at pupil level:

- The intervention: pupil-level and school-level dosage; 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 choose 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 (low/high buy-in schools); intensity of delivery (determined by sessions attended/number of weeks tutoring is spread over); completed versus scheduled sessions (determined as high if 80% of the sessions are completed, equivalent to 12 or more sessions out of 15).
- Tutors: Experience/qualifications; TP tutor training; shared characteristics with pupil/tutee (gender, ethnicity).
- Other: early/late delivery.

Missing data analysis

Missing baseline assessment data were imputed using multiple imputation (MI), a statistical technique which uses the distribution of observed data to estimate a set of plausible values for missing data. The missing values were replaced by the estimated plausible values by the estimation of multiple datasets. The variables used for MI of these baseline tests are the following: female, FSM ever in 6 years, ethnicity (White British, Asian, Black, unknown), EAL, SEN, looked after for 12 months, looked after since 31 March, looked after for 6 months; Ofsted ratings; school FSM percentage above median; urban, dummy for year groups; IDACI quintiles; school types; assessment providers. The results obtained from each dataset are combined using Rubin's rules to create a 'complete' dataset (Schafer, 1999).

³⁰ School attendance, listed in the study plan, could not be explored as the data was not available in the NPD in autumn 2021.

Estimation of effect sizes

Estimates are presented as effect sizes, calculated using the Hedges' g formula. Formally, the effect sizes were 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 7: Timeline

Dates	Activity	Staff responsible / leading
Oct 2020	Project set up, logic model development, materials development, study plan development	Consortium
Early Nov 2020	TPs launch. TP evaluation guidance pack launch. TPs can start contacting schools.	NFER and EEF
Nov 2020–July 2021	Tutoring period (whole programme)	TPs
End Nov–Dec 2020	Study plan finalisation and publication	Consortium
Early Dec 2020	Submit NPD request	UoW
Early Dec 2020–end Jan 2021	Evaluation team contact TP schools to invite them to take part in the evaluation sample (impact evaluation)	NFER
By end Dec 2020	Schools conduct baseline assessments (prior to starting tuition)	NFER
Dec 2020	First population data uploads; compilation and checks	NFER
5th Jan–8th March 2021	National lockdown period – many pupils learning from home, schools only open to children of keyworkers and vulnerable children. TP provision predominantly online during this period.	
End March 2021	Cut-off date for evaluation sample recruitment	NFER
End March 2021	Second population data uploads; compilation and checks	NFER
March–April 2021	Draw comparison sample and placebo check	UoW
Mid-April 2021	Confirm schools to be contacted for the comparison group	NFER/UoW
Mid-April–May 2021	Recruit comparison schools	NFER
June/July 2021	Testing window for end-point standardised assessments	
July–Aug 2021	Summer term data collection from evaluation sample and Comparison schools	
End Aug 2021	Final population data uploads from TPs; compilation and checks. MI data collection from comparison schools.	NFER
Late July–mid-Aug 2021	Access assessment data from providers	NFER
Mid-Aug–Nov 2021	Data cleaning (matching MI/pupil data to standardised assessment data for impact evaluation sample). Send data to NPD to match in.	NFER
Nov–Dec 2021	NPD (unamended) data available and matched into dataset	NPD team/ UoW
Jan 2021–March 2022	Impact analysis	UoW
Feb–May 2022	Draft reporting	All
June–Sept 2022	Final reporting and revisions	All
October 2022	Publication	

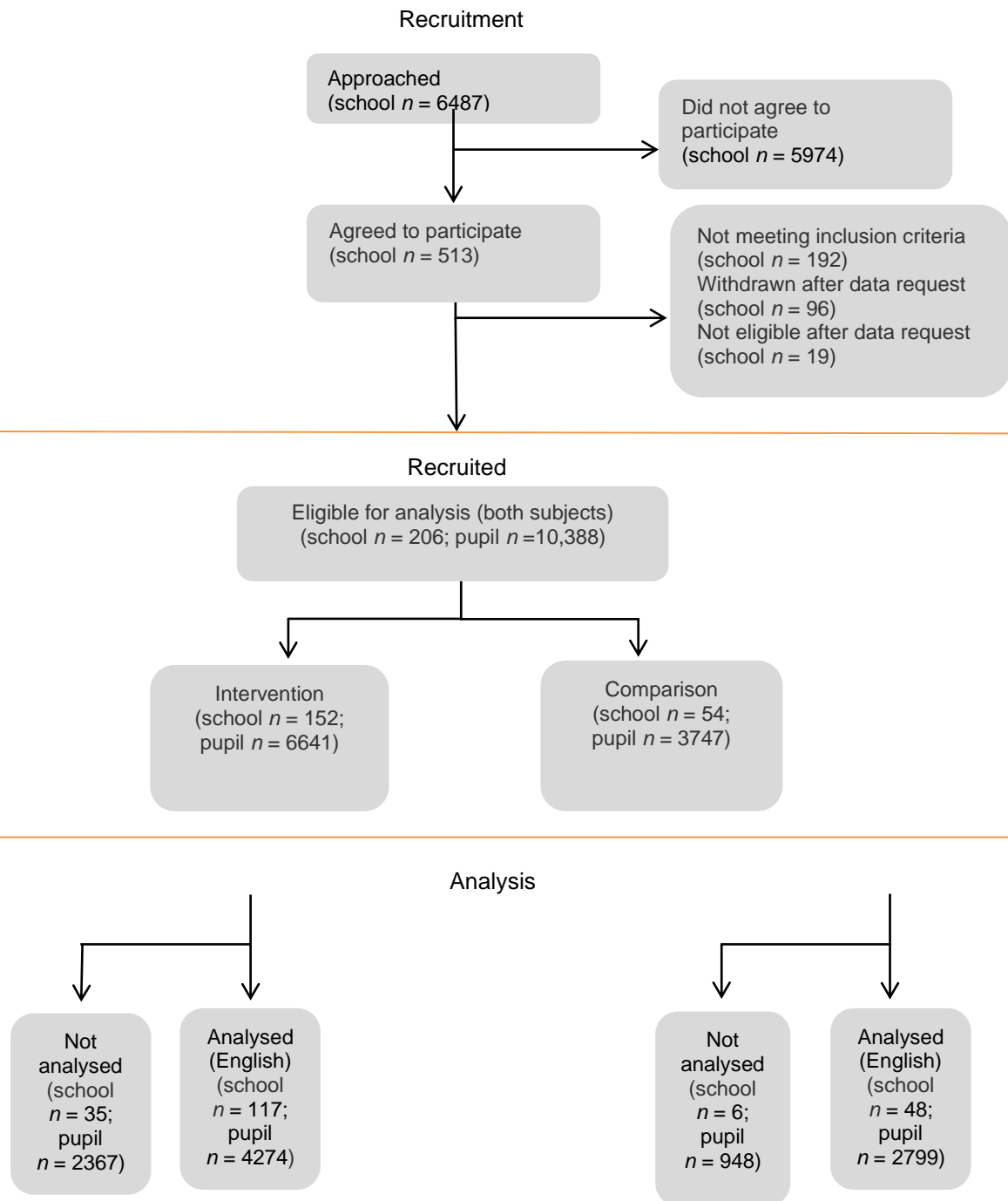
Impact evaluation results

Participant flow including losses and exclusions

Figures 1 and 2 provide details for the flow of participants through the study. Following the recruitment and the data collection, several schools were lost because of missing end-point testing. We did not require nor expect all schools to test on both subjects, but the loss associated with missing outcomes is substantial: 22% of schools in the TP intervention sample did not test in English and 45% did not test in maths. For the comparison sample, the sample loss is lower: 11% of schools did not test in English and 20% did not test in maths. The differential non-response across treatment groups points towards the likelihood of potentially unobserved differences between the two groups. However, as discussed later on, this may also be related to how data was requested from schools, particularly in the case of intervention schools, where the main focus was on the overlap of year groups testing and tutoring in the same subject.

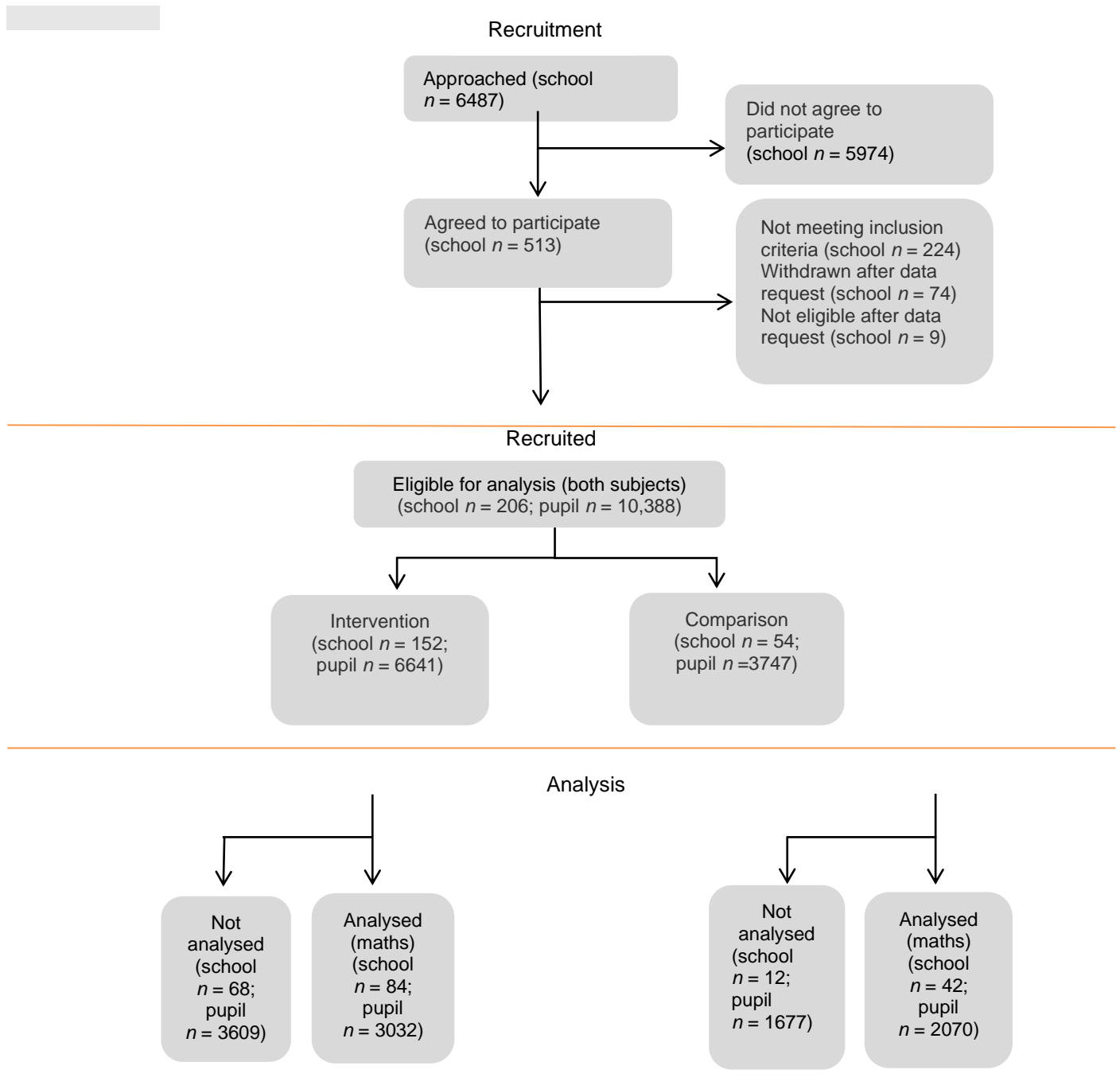
Schools and pupils with missing primary outcome measure were excluded from the analysis. Missing baseline scores were imputed using pupil-level and school-level information for the subset of pupils for whom we had outcome data, and as such are not reflected in the figures.

Figure 1: Participant flow diagram (2 arms) – English score for PP-eligible pupils³¹



³¹ The 'not analysed' figures for intervention and comparison groups are schools that were lost due to missing end-point assessments for the given subject. (Not all schools tested/tutored in both English and maths, and some schools' planned testing was disrupted due to Covid-related closures). Note: the number of analysed intervention schools for English in this participant flow was updated in December 2022 to reflect those that were actually analysed (i.e. schools with at least one pupil with non-missing baseline and end-point assessment data).

Figure 2: Participant flow diagram (2 arms) – maths score for PP-eligible pupils³²



³² The not-analysed figures for intervention and comparison groups are schools that were lost due to missing end-point assessments for the given subject. (Not all schools tested/tutored in both English and maths, and some schools' planned testing was disrupted due to Covid-related closures). The number of analysed schools represents schools with at least one pupil with non-missing baseline and end-point assessment data.

Table 8: Minimum detectable effect size at different stages, RQ1a, PP-eligible pupils

		Study plan		Analysis	
		English	Maths	English	Maths
MDES		0.11	0.115	0.125	0.125
Pre-test/post-test correlations	Level 1 and 2 (pupil and school)	0.70	0.70	0.60	0.63
Intracluster correlations (ICCs)	Level 1 (school)	0.15	0.15	0.15	0.13
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2-sided	2-sided	2-sided	2-sided
Average cluster size (PP-eligible pupils per school)		24	24	43	42
Number of schools	Intervention	120	105	117	84
	Comparison	54	49	48	42
	Total:	174	154	165	126
Number of pupils	Intervention	2880	2520	4274	3032
	Comparison	1296	1186	2799	2070
	Total:	4176	3706	7073	5102

Note: The table was updated in December 2022 to reflect the analysed school level sample sizes.

The difference between the assumed numbers of schools and children and the numbers analysed alters the power of the analysis. Table 8 presents this in the form of the minimum detectable effect size (MDES) for RQ1a. The MDES 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 in the study plan is 0.11 for English and 0.115 for maths.³³ This is set out in Table 8, along with the other assumptions used. Despite the sample loss due to missing outcomes, the recruitment target of intervention and comparison schools was almost met, except for intervention schools in maths, which was short of the target by 21 schools.

The top row of Table 8 presents the MDES for the analysis sample. In addition to reflecting the number of schools and pupils on which impacts are based, the observed ICC and correlation between regressors and the end-point assessment can now be included. The ICC is slightly lower than that assumed at the design stage (0.13 compared to 0.15) for maths. The pre-test/post-test pupil-level and school-level correlations are slightly lower (0.60 (English) and 0.63 (maths) compared to predicted 0.7).

³³ The MDES reported in the study plan are higher, as we erroneously computed them assuming 50% of pupil-level and school-level variance was accounted for, instead of 70%.

Together, these have the effect of slightly increasing the MDES to 0.125 for English and 0.125 for maths at the analysis stage.

Pupil-level attrition

Table 9 reports pupil-level attrition. Of the total sample, 36% of PP-eligible pupils in the TP intervention sample and 25% of PP-eligible pupils in the comparison sample did not test for English. 54% of PP-eligible pupils in the TP intervention sample and 45 % of PP-eligible pupils in the comparison sample did not test for maths. These pupil-level numbers mostly reflect schools that did not test in a specific subject. We did not expect all schools to test on both subjects. However, due to the way the attrition is calculated, schools that only tutored and tested in one subject are in the dataset as ‘missing’ for the other subject. Some pupils were lost because their assessment date was earlier than the first TP session.

When considering only schools with a testing regime in place in a specific subject, 24.4% of PP-eligible pupils did not test in maths among schools that submitted scores for maths and 20.4% in English among schools that submitted scores for English. The percentage of PP-eligible pupils that did not test in maths and English is higher for pupils in Years 4, 5 and 6 than for pupils in Years 1, 2 and 3. These statistics are reported in Tables 28 and 29 and discussed in the *Missing data* section. Pupil-level missingness in schools with non-missing outcomes is still substantial, because not all year groups in the same school tested in both subjects.

Even though the resulting pupil-level loss associated with missing outcomes or late TP delivery is substantial, the number of pupils analysed is higher than we anticipated in the study plan.

Table 9: Pupil-level attrition from the trial (RQ1a)

		Intervention	Comparison	Total
Number of pupils	Matched at analysis	6641	3747	10,388
	Analysed English	4274	2799	7073
	Analysed maths	3032	2070	5102
Pupil attrition (from matching to analysis)	Number English	2367	948	3315
	Percentage English	35.6%	25.3%	31.9%
	Number maths	3609	1677	5286
	Percentage maths	54.3%	44.8%	50.9%

Source: Primary school evaluation sample.

Preliminary analysis: Placebo tests for weighting methods

To check the similarity of TP schools and comparison schools, placebo testing was conducted on the sample of TP intervention and comparison schools on KS2 assessments for the academic years 2016/17 to 2018/19 and this is reported in Table 10. For each subject, we present regressions on KS2 scores, including the variables listed in the method section and in Appendix B. For each outcome and year, we present three specifications: one without weights, one with weights computed through inverse probability weighting and the third with weights computed with entropy balance.

Results indicate that, before the intervention, TP intervention schools did not have significantly different KS2 scores than matched comparison schools, except for the specification looking at maths scores in 2018/19, indicating that scores were significantly lower when estimated with the unweighted and with the inverse probability weighting specifications. Results using entropy balance are not significant. This provides support for the use of the comparison group as a means of estimating the counterfactual under the conditional independence assumption. In all the other cases, both weighted specifications present coefficients that are small and not significant.

For the main analysis, we proceeded with entropy balance as it performs better than inverse probability weighting with respect to the estimated placebo impacts for English (and performs no worse for maths). We proceeded with entropy balance for both samples because it also has the advantage, over inverse probability weighting, of no sample loss induced by perfect prediction. All regressions are estimated also without weights to check the sensitivity of the estimates to inclusion of weights.

Table 10: Placebo tests for the TP/comparison schools in years 2016/17 up to 2018/19

	KS2 RAW maths 2018/19		KS2 Ebal maths 2018/19		KS2 IPW maths 2018/19	
	Coef	S.E.	Coef	S.E.	Coef	S.E.
TP Intervention schools	-0.931*	(-2.27)	-0.508	(-1.41)	-0.974**	(-2.86)
Observations	121		121		115	
	KS2 RAW maths 2017/18		KS2 Ebal maths 2017/18		KS2 IPW maths 2017/18	
	Coef	S.E.	Coef	S.E.	Coef	S.E.
TP Intervention schools	0.585	(1.20)	0.470	(1.00)	0.402	(0.97)
Observations	118		118		110	
	KS2 RAW maths 2016/17		KS2 Ebal maths 2016/17		KS2 IPW maths 2016/17	
	Coef	S.E.	Coef	S.E.	Coef	S.E.
TP Intervention schools	0.217	(0.42)	0.127	(0.27)	0.262	(0.61)
Observations	117		117		113	
	KS2 RAW Eng 2018/19		KS2 Ebal Eng 2018/19		KS2 IPW Eng 2018/19	
	Coef	S.E.	Coef	S.E.	Coef	S.E.
TP Intervention schools	-0.197	(-0.47)	-0.247	(-0.57)	-0.274	(-0.67)
Observations	159		159		152	
	KS2 RAW Eng 2017/18		KS2 Ebal Eng 2017/18		KS2 IPW Eng 2017/18	
	Coef	S.E.	Coef	S.E.	Coef	S.E.
TP Intervention schools	-0.0184	(-0.05)	-0.0826	(-0.23)	0.0368	(0.11)
Observations	156		156		146	
	KS2 RAW Eng 2016/17		KS2 Ebal Eng 2016/17		KS2 IPW Eng 2016/17	
	Coef	S.E.	Coef	S.E.	Coef	S.E.
TP Intervention schools	-0.195	(-0.43)	-0.220	(-0.55)	-0.0177	(-0.05)
Observations	153		153		147	

Source: Primary school evaluation sample. Note: controls include school-level variables: percentage FSM, percentage EAL, percentage SEND, four quantiles of pupil-to-teacher ratio, four quantiles of pupil count, Ofsted ratings, rural, region, type of schools, lagged and missing KS2, KS1 to KS2 value added.

Pupil and school characteristics

School-level selection

Observable characteristics

Demographic data are presented in Tables 11a, 11b and 11c, with all figures rounded to three decimal places. Tables 11a and 11b present the comparison of observable characteristics between TP intervention and comparison schools, with and without weights, and with national averages. Table 11a refers to the maths sample and Table 11b to the English sample. We applied weighting, as described in the *Preliminary analysis* section.

Details of the school characteristics show small differences between TP intervention and comparison schools, for both maths and English (Tables 11a and 11b, respectively).

For maths (Table 11a), the only significant difference between TP intervention and comparison schools is pupil-to-teacher ratio in 2018, which is higher in TP intervention schools (21.47% vs 20.15% in the comparison sample and 20.6% in the national population). Significant differences disappear once we apply entropy balance weights. However, introducing weights results in a significantly higher percentage of schools from Yorkshire and the Humber in the comparison sample (39.8% versus 15.5% in TP intervention schools), and in the proportion of pupils in Year 2 (15.2% in comparison schools versus 13.2% in TP schools, and a significantly higher mean number of pupils in TP schools (mean of 285) than in comparison schools (mean of 242).

For English (Table 11b), a higher proportion of comparison schools were Voluntary Controlled schools (14.6%, compared to 6.0% and 7.9% in the TP intervention and nationwide respectively). The difference is no longer significant once we include entropy balance weights. Comparison schools have a significantly higher proportion of female pupils than TP intervention schools (51.1% versus 47.6% in the unweighted samples) and the difference remains significant once we apply weights. Comparison schools have a significantly higher proportion of pupils in Year 5 than TP intervention schools and the national average (22.1% versus 18.6% and 17.1% respectively). This is the result of more Year 5 data having been provided by comparison schools. Differences in gender and in Year 5 pupils remain after applying weights to the samples.

Table 11c presents the comparison of observable characteristics between the population of (all) primary TP schools, the national averages and the TP intervention samples (maths and English presented separately). The descriptives indicate that the TP primary school population (i.e., all of the primary schools that signed up to the programme) appears to be similar in terms of observable characteristics to primary schools nationally. Schools in the TP intervention samples look similar to both schools in the TP population and in the population of schools, except for some characteristics. For example, there are more schools in the TP intervention samples from urban areas than found in the TP population and nationally, more schools in the TP intervention groups from the East Midlands, the Yorkshire and the Humber and fewer schools from the North East, the South East and London than in the TP population and nationally. In terms of pupil composition, schools in the TP intervention samples have a slightly lower percentage of SEN pupils, a slightly higher IDACI score and a higher percentage of White British pupils than the TP population and nationally.

Table 11a1: Baseline characteristics of TP intervention schools, recruited comparison schools and national proportions, unweighted – maths sample

Unweighted specification							
Variable	National averages	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff: Comparison–TP intervention
Total pupil counts	290.2928	250.571	(150.647)	285.274	(146.089)	34.702	0.165
Pupil-to-teacher ratio 2018	20.602	20.154	(4.214)	21.472	(3.157)	1.318*	0.250
Ofsted 2018: Outstanding	0.154	0.071		0.107		0.036	0.088
Ofsted 2018: Good	0.709	0.690		0.655		-0.036	-0.053
Ofsted 2018: Inadequate	0.024	0.071		0.036		-0.036	-0.111
Ofsted 2018: Requires improvement	0.113	0.143		0.155		0.012	0.023
School types: Community schools	0.345	0.238		0.321		0.083	0.131
School types: Converter Academies	0.249	0.190		0.310		0.119	0.195
School types: Foundation schools	0.038	-		0.036		-0.012	-0.042
School types: Free schools	0.020	-		-		-0.036	-0.148
School types: Sponsored Academies	0.120	0.190		0.095		-0.095	-0.192
School types: Voluntary Aided Schools	0.150	0.143		0.155		0.012	0.023
School types: Voluntary Controlled Schools	0.079	0.119		0.060		-0.060	-0.147
Urban	0.784	0.762		0.810		0.048	0.081
Region: East Midlands	0.094	0.119		0.155		0.036	0.073
Region: East of England	0.126	0.071		0.131		0.060	0.139
Region: London	0.131	0.143		0.083		-0.060	-0.132
Region: North East	0.060	0.071		0.048		-0.024	-0.071
Region: North West	0.161	0.119		0.131		0.012	0.025
Region: South East	0.129	-		0.095		0.048	0.130
Region: South West	0.096	0.143		0.095		-0.048	-0.103
Region: West Midlands	0.102	0.071		0.095		0.024	0.060
Region: Yorkshire & the Humber	0.102	0.190		0.155		-0.036	-0.066
% EAL	0.129	0.161		0.127		-0.034	-0.111
% SEN	0.301	0.270		0.251		-0.019	-0.095
% FSM	0.263	0.310		0.300		-0.01	-0.039
% Female	0.480	0.510		0.483		-0.027	-0.205
Average IDACI scores	0.184	0.229	(0.107)	0.209	(0.094)	-0.021	-0.145
% White British	0.698	0.705		0.727		0.022	0.055
% Not White	0.255	0.280		0.253		-0.028	-0.070
% Year 1	0.144	0.120		0.130		0.010	0.087
% Year 2	0.155	0.131		0.132		0.001	0.010
% Year 3	0.153	0.177		0.176		-0.000	-0.003
% Year 4	0.167	0.175		0.173		-0.002	-0.016
% Year 5	0.171	0.211		0.192		-0.019	-0.130
% Year 6	0.175	0.187		0.187		-0.001	-0.005

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Source: Evaluation sample and population data from the NPD. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in maths

Table 11a2: Baseline characteristics of TP intervention schools, recruited comparison schools and national proportions, weighted – maths sample

Variable	Weighted						Diff in means: Comparison -TP intervention	Std diff: Comparison - TP intervention
	National averages	Means: Comparison n	St dev: Comparison n	Means: TP intervention n	St dev: TP intervention			
Total pupil counts	290.2928	242.216	(121.839)	285.274	(146.089)	43.058*	0.226	
Pupil-to-teacher ratio 2018	20.602	21.128	(2.590)	21.472	(3.157)	0.344	0.084	
Ofsted 2018: Outstanding	0.154	0.143		0.107		-0.036	-0.076	
Ofsted 2018: Good	0.709	0.655		0.655		-0.000	0.000	
Ofsted 2018: Inadequate	0.024	0.054		0.036		-0.018	-0.061	
Ofsted 2018: Requires improvement	0.113	0.136		0.155		0.018	0.038	
School types: Community schools	0.345	0.344		0.321		-0.022	-0.034	
School types: Converter Academies	0.249	0.282		0.310		0.028	0.043	
School types: Foundation schools	0.038	0.005		0.036		0.031	0.156	
School types: Free schools	0.020	0.041		0.012		-0.029	-0.127	
School types: Sponsored Academies	0.120	0.132		0.095		-0.037	-0.082	
School types: Voluntary Aided Schools	0.150	0.096		0.155		0.059	0.125	
School types: Voluntary Controlled Schools	0.079	0.089		0.060		-0.029	-0.078	
Urban	0.784	0.821		0.810		-0.012	-0.020	
Region: East Midlands	0.094	0.086		0.155		0.068	0.149	
Region: East of England	0.126	0.088		0.131		0.043	0.097	
Region: London	0.131	0.083		0.083		0.000	0.000	
Region: North East	0.060	0.042		0.048		0.005	0.020	
Region: North West	0.161	0.081		0.131		0.049	0.114	
Region: South East	0.129	0.055		0.095		0.040	0.107	
Region: South West	0.096	0.039		0.095		0.056	0.158	
Region: West Midlands	0.102	0.114		0.095		-0.019	-0.044	
Region: Yorkshire & the Humber	0.102	0.398		0.155		-0.243***	-0.395	
% EAL	0.129	0.105		0.127		0.021	0.083	
% SEN	0.301	0.256		0.251		-0.005	-0.030	
% FSM	0.263	0.256		0.251		-0.005	-0.095	
% Female	0.480	0.508		0.483		-0.025	-0.193	
Average IDACI scores	0.184	0.227	(0.101)	0.209	(0.094)	-0.019	-0.130	
% White British	0.698	0.766		0.727		-0.039	-0.108	
% Non white	0.255	0.223		0.253		0.030	0.085	
% Year 1	0.144	0.137		0.130		-0.007	-0.061	
% Year 2	0.155	0.154		0.132		-0.023*	-0.208	
% Year 3	0.153	0.172		0.176		0.004	0.052	
% Year 4	0.167	0.160		0.173		0.014	0.136	
% Year 5	0.171	0.192		0.192		-0.001	0.000	
% Year 6	0.175	0.184		0.187		0.003	0.033	
N	42	84		126				

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Evaluation sample and population data from the NPD. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in maths

Table 11b1: Baseline characteristics of TP intervention schools, recruited comparison schools and national proportions, unweighted – English sample

Unweighted specification							
Variable	National averages	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison – TP intervention	Std diff: Comparison – TP intervention
Total pupil counts	290.2928	252.625	(150.717)	284.547	(141.401)	31.922	0.154
Pupil-to-teacher ratio 2018	20.602	20.404	(4.079)	21.358	(3.221)	0.954	0.184
Ofsted 2018: Outstanding	0.154	0.104		0.085		-0.019	-0.045
Ofsted 2018: Good	0.709	0.688		0.650		-0.038	-0.057
Ofsted 2018: Inadequate	0.024	0.063		0.034		-0.028	-0.093
Ofsted 2018: Requires improvement	0.113	0.125		0.171		0.046	0.091
School types: Community schools	0.345	0.250		0.333		0.083	0.129
School types: Converter Academies	0.249	0.208		0.299		0.091	0.147
School types: Foundation schools	0.038	0.063		0.043		-0.020	-0.062
School types: Free schools	-	-		-		-0.025	-0.102
School types: Sponsored Academies	0.120	0.146		0.128		-0.018	-0.036
School types: Voluntary Aided Schools	0.150	0.125		0.111		-0.014	-0.030
School types: Voluntary Controlled Schools	0.079	0.146		0.060		-0.086*	-0.201
Urban	0.784	0.771		0.803		0.033	0.056
Region: East Midlands	0.094	0.125		0.128		0.003	0.007
Region: East of England	0.126	0.063		0.128		0.066	0.158
Region: London	0.131	0.125		0.094		-0.031	-0.070
Region: North East	0.060	0.063		0.043		-0.020	-0.062
Region: North West	0.161	0.104		0.145		0.041	0.088
Region: South East	0.129	0.063		0.094		0.032	0.083
Region: South West	0.096	0.167		0.103		-0.064	-0.132
Region: West Midlands	0.102	0.083		0.085		0.002	0.005
Region: Yorkshire & the Humber	0.102	0.188		0.171		-0.017	-0.030
% EAL	0.129	0.149		0.135		-0.014	-0.048
% SEN	0.301	0.259		0.253		-0.005	-0.029
% FSM		0.308		0.307		-0.001	-0.005
% Female	0.480	0.511		0.476		-0.035**	-0.255
Average IDACI scores	0.184	0.226	(0.108)	0.216	(0.099)	-0.010	-0.066
% White British	0.698	0.717		0.719		0.002	0.005
% Not White	0.255	0.270		0.262		-0.009	-0.022
% Year 1	0.144	0.122		0.135		0.013	0.115
% Year 2	0.155	0.124		0.139		0.015	0.142
% Year 3	0.153	0.179		0.167		-0.012	-0.142
% Year 4	0.167	0.168		0.181		0.014	0.130
% Year 5	0.171	0.221		0.186		-0.035**	-0.242
% Year 6	0.175	0.186		0.185		-0.001	-0.011

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Evaluation sample and population data from the NPD. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in English

Table 11b2: Baseline characteristics of TP intervention schools, recruited comparison schools and national proportions, weighted – English sample

Variable	Weighted specification					Diff in means: Comparison-TP intervention	Std diff: Comparison - TP intervention
	National averages	Means: Comparison	St dev: Comparison	Means: TP intervention	St dev: TP intervention		
Total pupil counts	290.2928	256.431	(146.890)	284.547	(141.401)	28.116	0.138
Pupil-to-teacher ratio 2018	20.602	21.101	(3.861)	21.358	(3.221)	0.256	0.051
Ofsted 2018: Outstanding	0.154	0.103		0.085		-0.017	-0.043
Ofsted 2018: Good	0.709	0.650		0.650		-0.000	0.000
Ofsted 2018: Inadequate	0.024	0.060		0.034		-0.026	-0.086
Ofsted 2018: Requires improvement	0.113	0.145		0.171		0.026	0.050
School types: Community schools	0.345	0.277		0.333		0.056	0.086
School types: Converter Academies	0.249	0.263		0.299		0.037	0.056
School types: Foundation schools	0.038	0.093		0.043		-0.051	-0.140
School types: Free schools	-	0.012		0.017		0.005	0.029
School types: Sponsored Academies	0.120	0.136		0.128		-0.008	-0.017
School types: Voluntary Aided Schools	0.150	0.093		0.111		0.018	0.042
School types: Voluntary Controlled Schools	0.079	0.083		0.060		-0.023	-0.063
Urban	0.784	0.796		0.803		0.007	0.012
Region: East Midlands	0.094	0.091		0.128		0.037	0.083
Region: East of England	0.126	0.081		0.128		0.047	0.108
Region: London	0.131	0.094		0.094		-0.000	0.000
Region: North East	0.060	0.025		0.043		0.018	0.070
Region: North West	0.161	0.090		0.145		0.055	0.120
Region: South East	0.129	0.055		0.094		0.039	0.105
Region: South West	0.096	0.189		0.103		-0.086	-0.172
Region: West Midlands	0.102	0.088		0.085		-0.003	-0.007
Region: Yorkshire & the Humber	0.102	0.243		0.171		-0.072	-0.125
% EAL	0.129	0.117		0.135		0.018	0.068
% SEN	0.301	0.240		0.253		0.013	0.082
% FSM	0.263	0.307		0.307		0.000	0.000
% Female	0.480	0.515		0.476		-0.039**	-0.278
Average IDACI scores	0.184	0.219	(0.100)	0.216	(0.099)	-0.003	-0.021
% White British	0.698	0.760		0.719		-0.041	-0.111
% Not white	0.255	0.224		0.262		0.038	0.104
% Year 1	0.144	0.132		0.135		0.003	0.027
% Year 2	0.155	0.123		0.139		0.016	0.160
% Year 3	0.153	0.178		0.167		-0.011	-0.137
% Year 4	0.167	0.168		0.181		0.013	0.137
% Year 5	0.171	0.218		0.186		-0.032**	-0.250
% Year 6	0.175	0.181		0.185		0.004	0.042
N		48		117		165	

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: Evaluation sample and population data from the NPD. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in English

Table 11c: Baseline characteristics of the population of TP primary schools, recruited TP intervention schools and national proportions, unweighted

Variable	National averages	Means: all TP primary schools	Means: TP intervention - Maths (unweighted)	Means: TP intervention - English (unweighted)
Total pupil counts	290.2928	273.678	285.274	284.547
Pupil-to-teacher ratio 2018	20.602	20.760	21.474	21.358
Ofsted 2018: Outstanding	0.154	0.145	0.107	0.085
Ofsted 2018: Good	0.709	0.657	0.655	0.650
Ofsted 2018: Inadequate	0.024	0.019	0.036	0.034
Ofsted 2018: Requires improvement	0.113	0.100	0.155	0.171
Ofsted 2018: Missing	0.031	0.080	0.048	0.060
School types: Community schools	0.345	0.334	0.321	0.333
School types: Converter Academies	0.249	0.231	0.310	0.299

School types: Foundation schools	0.038	0.034	0.036	0.043
School types: Free schools	0.020	0.016	-	-
School types: Sponsored Academies	0.120	0.106	0.095	0.128
School types: Voluntary Aided Schools	0.150	0.148	0.155	0.111
School types: Voluntary Controlled Schools	0.079	0.079	0.060	0.060
Urban	0.784	0.776	0.810	0.803
Region: East Midlands	0.094	0.093	0.155	0.128
Region: East of England	0.126	0.124	0.131	0.128
Region: London	0.131	0.120	0.083	0.094
Region: North East	0.060	0.050	0.048	0.043
Region: North West	0.161	0.159	0.131	0.145
Region: South East	0.129	0.129	0.095	0.094
Region: South West	0.096	0.087	0.095	0.103
Region: West Midlands	0.102	0.092	0.095	0.085
Region: Yorkshire & the Humber	0.102	0.095	0.155	0.171
% EAL	0.129	0.150	0.127	0.135
% SEN	0.301	0.269	0.251	0.253
% Female	0.480	0.489	0.483	0.476
Average IDACI scores	0.184	0.198	0.209	0.216
% White British	0.698	0.682	0.727	0.719
% Not white	0.255	0.284	0.253	0.262
% Year 1	0.144	0.130	0.130	0.135
% Year 2	0.155	0.140	0.132	0.139
% Year 3	0.153	0.162	0.176	0.167
% Year 4	0.167	0.178	0.173	0.181
% Year 5	0.171	0.182	0.192	0.186
% Year 6	0.175	0.184	0.187	0.185
N	17496	4211	84	117

Source: Primary school evaluation sample, population of TP primary schools and population data from the NPD. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Note: Counts lower than 3 had to be suppressed for statistical disclosure control (indicated as '-'). If there was only one count that had to be suppressed, we also suppressed the 2nd lowest count, so that the missing count could not be derived from the total. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in English.

Baseline assessment

As well as looking at differences in background characteristics, we also ran checks to compare the samples in terms of the standardised baseline assessment scores.

The results are shown in Table 12a for maths and Table 12b for English: TP intervention schools and comparison schools do not have statistically significantly different maths and English scores, at baseline. Histograms of these scores for all pupils and PP-eligible pupils can be found in Appendix C. Each graph shows the distribution of scores among all pupils and among PP-eligible pupils only, for the TP intervention and comparison groups.

Table 12a: Comparison of baseline assessments of TP intervention group and recruited comparison group, unweighted and weighted – maths sample

Unweighted						
Variable	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff: Comparison – TP intervention
Maths scores autumn assessment	93.911	(6.474)	94.98	(5.826)	1.069	0.123
Weighted specification						
Variable	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff: Comparison – TP intervention
Maths scores autumn assessment	94.795	(6.006)	94.98	(5.826)	0.185	0.022
Observations	42		84		126	

Source: Primary school evaluation sample. Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in maths.

Table 12b: Comparison of baseline assessments of TP intervention group and recruited comparison group, unweighted and weighted – English sample

Unweighted						
Variable	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff: Comparison – TP intervention
English scores autumn assessment	96.698	(5.716)	95.141	(5.520)	-1.557	-0.196
Weighted						
Variable	Means: Comparison	St dev: Comparison	Means: TP intervention	St dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff: Comparison – TP intervention
English scores autumn assessment	96.260	(5.057)	95.141	(5.520)	-1.020	-0.149
Observations	48		117		165	

Source: Primary school evaluation sample. Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in English.

School-level section: a summary

Overall, it appears that the TP intervention and comparison sample are similar among the majority of observable characteristics and the outcomes. We tested 35 characteristics and by chance we could have found that a small percentage of these tests were statistically significant. The majority of comparison schools were selected according to whether schools expressed an interest in the TP programme (either by signing an MoU or an EoI) suggesting that at least on the propensity to implement TP, comparison schools are motivated similarly to TP intervention schools. Together with the similarity of a wide range of observable characteristics we are confident in considering school-level selection to be accounted for by the recruitment strategy, the matching and weighting procedures.

Pupil-level selection: Observable characteristics

The programme was intended to reach disadvantaged pupils including those eligible for Pupil Premium funding, Free School Meals or those identified by schools as having an equivalent need for support. As noted in the participant

Selection section (see Methods), schools were able to decide which pupils would receive tutoring, and whilst 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 tutoring programme. Indeed, this had been the case in the online tutoring pilot in 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 around half of those that were selected for TP were PP-eligible pupils: the percentage of TP pupils who are eligible for PP is 52% in the maths sample and 46% in the English sample (Table 13a).

Table 13a: Number and percentage of PP-eligible pupils who are TP and TP who are eligible for PP in the TP intervention maths sample and English sample

Primary school sample (analytic sample for Maths), N: 126 schools						
	N TP pupils (of PP)	Total PP	% TP pupils (of PP)	N PP pupils (of TP)	Total TP	% PP pupils (of TP)
Total across year groups	937	5102	18.37%	937	1804	51.94%
Primary school sample (analytic sample for English), N: 165 schools						
Total across year groups	1449	7073	20.49%	1449	3145	46.07%

Source: Primary school evaluation sample. Note: The table was updated in December 2022 to reflect the analysed school level sample sizes.

For context, around a third of pupils were eligible for PP within the year groups in the TP schools in our sample (Table 13b): 33% of pupils in the maths sample; 32% of pupils in the English sample).³⁴ So, whilst 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 a fifth (20%) of PP-eligible pupils were selected for TP in the samples which means that pupil-level selection from schools is an issue (Table 13a). 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 amongst 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 in receipt of 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 conduct analysis based on predicted participation (RQ2), which would have provided an alternative approach to approximating the eligible group. However, as reported later on (under RQ2), this analysis could not proceed due to the poor predictive power of the model.

In addition to the issue of dilution for the analysis on PP-eligible pupils, this is a consideration for the analysis on all pupils too. Table 13b shows that the proportion of all pupils in the intervention group (TP schools) who were selected for TP is 12% in the maths sample and 14% in the English sample. This suggests that dilution is also likely to be a problem in detecting significant impact of TP in RQ3 (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. Figures 3 and 4 (on dosage: see *Amount of tutoring received* section below) indicate that almost 10% of schools in the analytical sample did not tutor in one of the two subjects. We avoid trying to account for selection into subject by only selecting TP intervention schools irrespective of the subject they tutored in. However, it has to be acknowledged that this also contributes further to dilution.

³⁴ Note the percentage of pupils eligible for PP was lower in Years 1 and 2, and higher in Year 6.

Table 13b: Distribution of PP-eligible and TP pupils by year group in the TP intervention sample, the maths sample and English sample

Primary schoolsample (analytic sample for Maths), N: 126 schools							
Year	Non-PP	PP	Non-TP pupils	TP pupils	Total	PP %	TP pupils %
1	763	287	998	52	1,050	27	5
2	1,026	429	1,304	151	1,455	29	10
3	2,043	1,008	2,677	374	3,051	33	12
4	2,452	1,165	3,220	397	3,617	32	11
5	2,825	1,515	3,686	654	4,340	35	15
6	1,132	698	1,654	176	1,830	38	10
Total	10,241	5,102	13,589	1,804	15,393	33	12
Primary school sample (analytic sample for English), N: 165 schools							
Year	Non-PP	PP	Non-TP pupils	TP pupils	Total	PP %	TP pupils %
1	945	371	1,224	92	1,316	28	7
2	1,596	668	1,925	339	2,264	30	15
3	3,382	1,491	4,121	752	4,873	31	15
4	3,582	1,691	4,503	770	5,273	32	15
5	3,623	1,896	4,634	885	5,519	34	16
6	1,596	956	2,248	304	2,552	37	12
Total	14,724	7,073	18,734	3,145	21,879	32	14

Source: Primary school evaluation sample. We further explored the distribution of the baseline scores of pupils selected for TP versus pupils not selected for TP in TP schools to assess the ability composition, as measured by the baseline assessment, 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. Note: The table was updated in December 2022 to reflect the analysed school level sample sizes.

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 (i.e., that pupils with lower prior performance were selected to participate in TP by schools).

In more detail, the results shown in Table 14 present the regression on standardised summer assessments 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 baseline assessment (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 attainment quartile and TP status interaction coefficients. The size of the coefficients is associated with higher quartiles of baseline assessment (except for maths and PP-eligible pupils where the interaction with the highest quartile is not significant).

Table 14: Interaction of quartiles of baseline scores with pupils' TP status

	PP pupils						Non-PP 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 Maths quartiles:																			
Math baseline 2nd quart#TP	-2.635*	1.053	0.014				-2.741	1.412	0.055				-2.365*	1.040	0.025				
Math baseline 3rd quart#TP	-5.375**	1.507	0.001				-5.426**	1.580	0.001				-5.195***	1.459	0.001				
Math baseline 4th quart#TP	-2.552	2.624	0.333				-7.886*	2.466	0.002				-5.862*	2.270	0.011				
Receiving TP X English quartiles:																			
English baseline 2nd quart#TP				-2.321*	1.044	0.028				-2.680*	0.943	0.005					-2.524**	0.844	0.003
English baseline 3rd quart#TP				-4.663**	1.216	0.000				-2.635*	1.156	0.024					-3.593**	1.028	0.001
English baseline 4th quart#TP				-3.655*	1.393	0.010				-3.434*	1.293	0.009					-3.722**	1.090	0.001
Constant	60.939**	13.146	0.000	54.468**	10.475	0.000	88.239**	11.097	0.000	75.459**	7.711	0.000	74.251***	12.779	0.000	72.188**	10.596	0.000	
N	5102			7073			10,241			14,724			15393			21879			

Source: TP intervention sample. 1st quartile is the lowest performing. The table presents the coefficients of the interactions between quartiles of baseline ability and TP pupils in the sample of PP-eligible pupils, non-PP pupils and all pupils.

Amount of tutoring received

Table 15 presents the school-level average of blocks of tutoring pupils received by the time of the end-point assessment. 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, at pupil level, PP-eligible pupils received on average 73% of a block of 12 hours in English and 74% of a block in maths (i.e., 8.8 hours of English tutoring and 8.9 hours of maths by the time of the assessment). The numbers slightly change to 8.9 for English and 8.6 for maths when all pupils are included in the sample. 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 11th June 2021 (note: 41% of booking rows in the full dataset across all year groups did not provide detailed dates per session).

Table 15: Pupil-level average of blocks of tutoring received

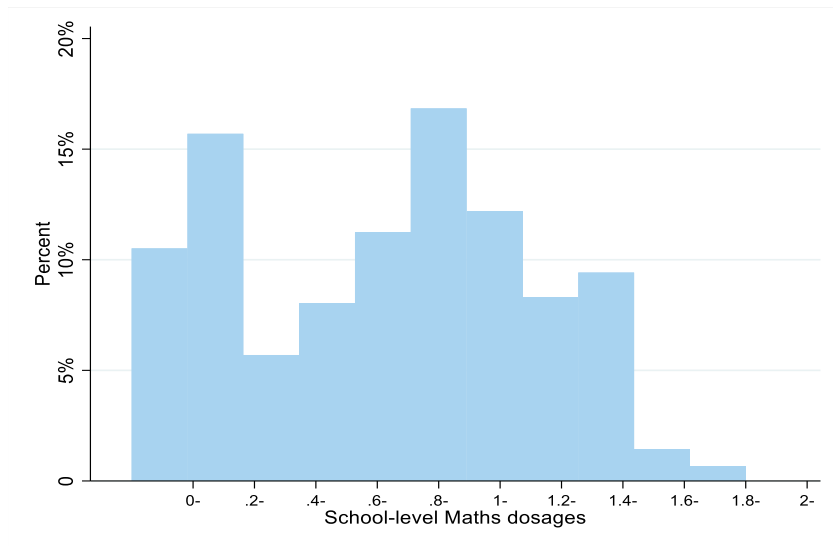
		Obs	Avg blocks	Std dev
PP-eligible pupils	Subject maths	703	0.74	0.58
	Subject English	1110	0.73	0.57
All pupils	Subject maths	1579	0.72	0.57
	Subject English	2330	0.74	0.60

Source: TP intervention sample.

Figures 3 and 4 show the distribution of school-level dosage in the two samples, based on pupils selected for TP. There are some schools with zero dosage in one of the two subjects, which is explained by the fact that not all pupils in a school had TP for both subjects and by the fact that pupils with missing or blank dosage are recorded as zero.³⁵

³⁵ For the maths sample, 10 schools (10.05% of pupils) have zero school-level dosages. TP pupils in these 10 schools have online dosage '0' and face-to-face dosage blank. For the English sample, 12 schools (9.57% of pupils) have zero school-level dosages.

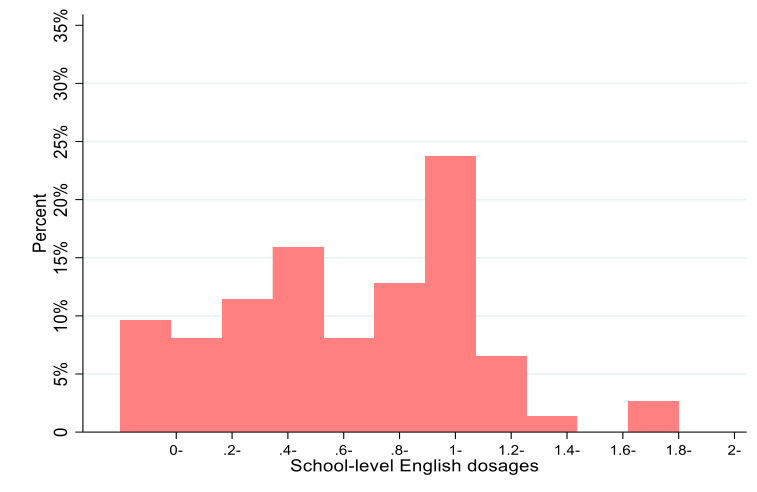
Figure 3: Distribution of school-level dosage on all TP pupils in terms of blocks of TP for maths (where one block corresponds to 12 hours of tutoring), continuous variable



Source: TP intervention sample.

Note: The x-axis shows the percent of pupils selected for TP. The first bar (shown to the left of zero) represents all schools with zero dosage, the second bar represents schools with dosage >0 and up to 0.2 and so on.

Figure 4: Distribution of school-level dosage on all TP pupils in terms of blocks of TP for English (where one block corresponds to 12 hours of tutoring), continuous variable



Source: TP intervention sample.

Note: The x-axis shows the percent of pupils selected for TP. The first bar (shown to the left of zero) represents all schools with zero dosage, the second bar represents schools with dosage >0 and up to 0.2 and so on.

Outcomes and analysis

Outcome analysis

Outcomes

We ran checks to compare the samples in terms of the standardised assessment scores. The results are shown in Table 16a for maths and Table 16b for English: TP intervention schools and comparison schools do not have statistically significantly different maths and English end-point assessments. Sets of histograms provide a graphical inspection of the distributions of standardised assessments in the analysed TP and comparison schools. Histograms of these scores for all pupils and PP-eligible pupils can be found in Figure 5 (maths) and Figure 6 (English). Each graph shows the distribution of scores among all pupils and among PP-eligible pupils only, for the TP intervention and comparison groups. They do not indicate the presence of different patterns of testing at end-point between TP intervention schools and comparison schools in both English and maths, for all pupils and PP-eligible pupils. The distribution of scores for English is similar between TP intervention and comparison schools for all pupils. For PP-eligible pupils, it appears that the distribution of scores is more centred around the mean for PP-eligible pupils in TP intervention schools.

Table 16a: Comparison of outcomes of TP intervention group and recruited comparison group, unweighted and weighted – maths sample

		Unweighted				
Variable	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff
Maths scores summer assessment	96.942	(6.639)	97.902	(7.537)	0.960	0.096
		Weighted specification				
Variable	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff
Maths scores summer assessment	98.216	(6.095)	97.902	(7.537)	-0.313	-0.032
Observations	42		84		126	

Source: Primary school evaluation sample. Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in maths.

Table 16b: Comparison of outcomes of TP intervention group and recruited comparison group, unweighted and weighted – English sample

		Unweighted				
Variable	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff
English scores summer assessment	99.401	(6.550)	98.462	(7.067)	-0.940	-0.098
		Weighted				
Variable	Means: Comparison	Std dev: Comparison	Means: TP intervention	Std dev: TP intervention	Diff in means: Comparison–TP intervention	Std diff
English scores summer assessment	99.271	(6.067)	98.462	(7.067)	-0.809	-0.087
Observations	48		117		165	

Source: Primary school evaluation sample. Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Note (Dec 2022): The numbers in this table have been updated to reflect the analysed sample in English.

Figure 5: Histograms of the distributions of summer maths assessment, all pupils and PP-eligible pupils, unweighted

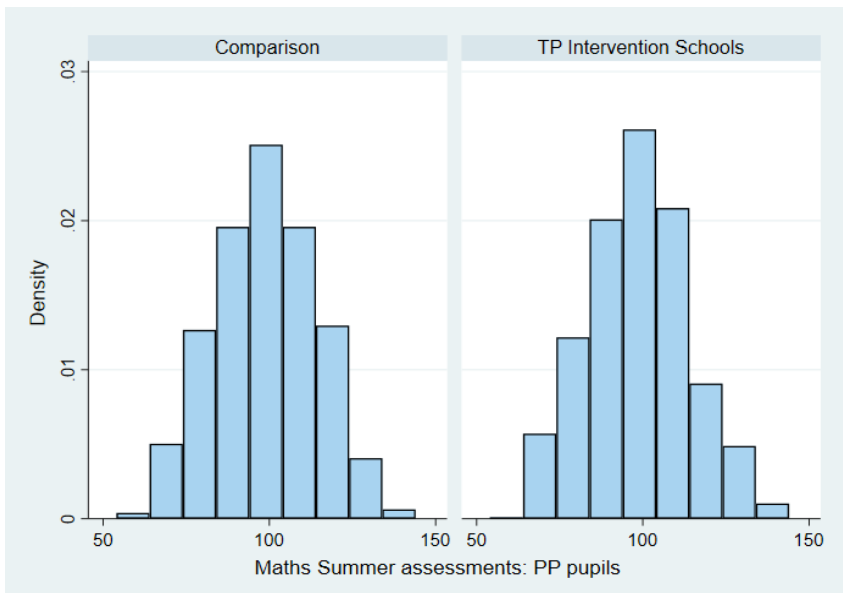
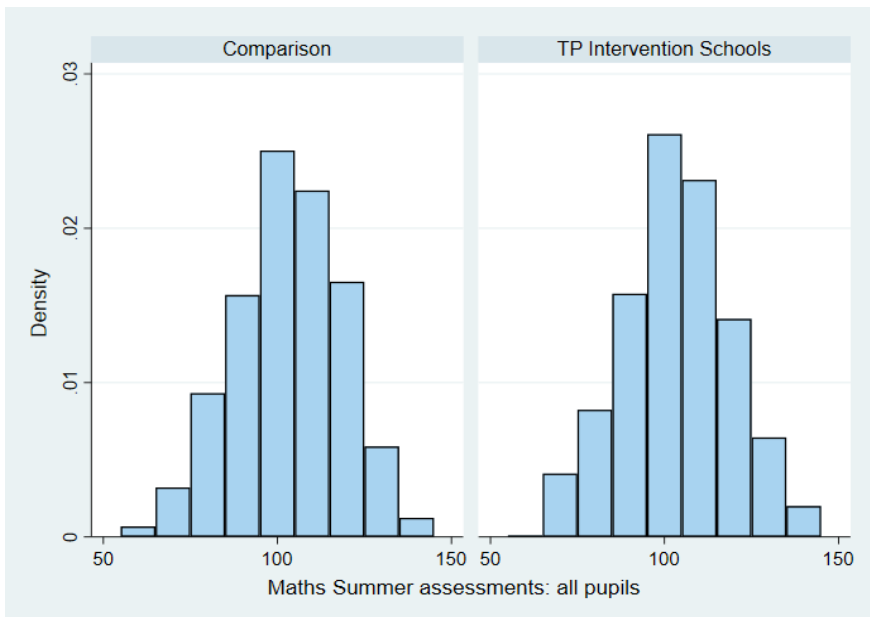
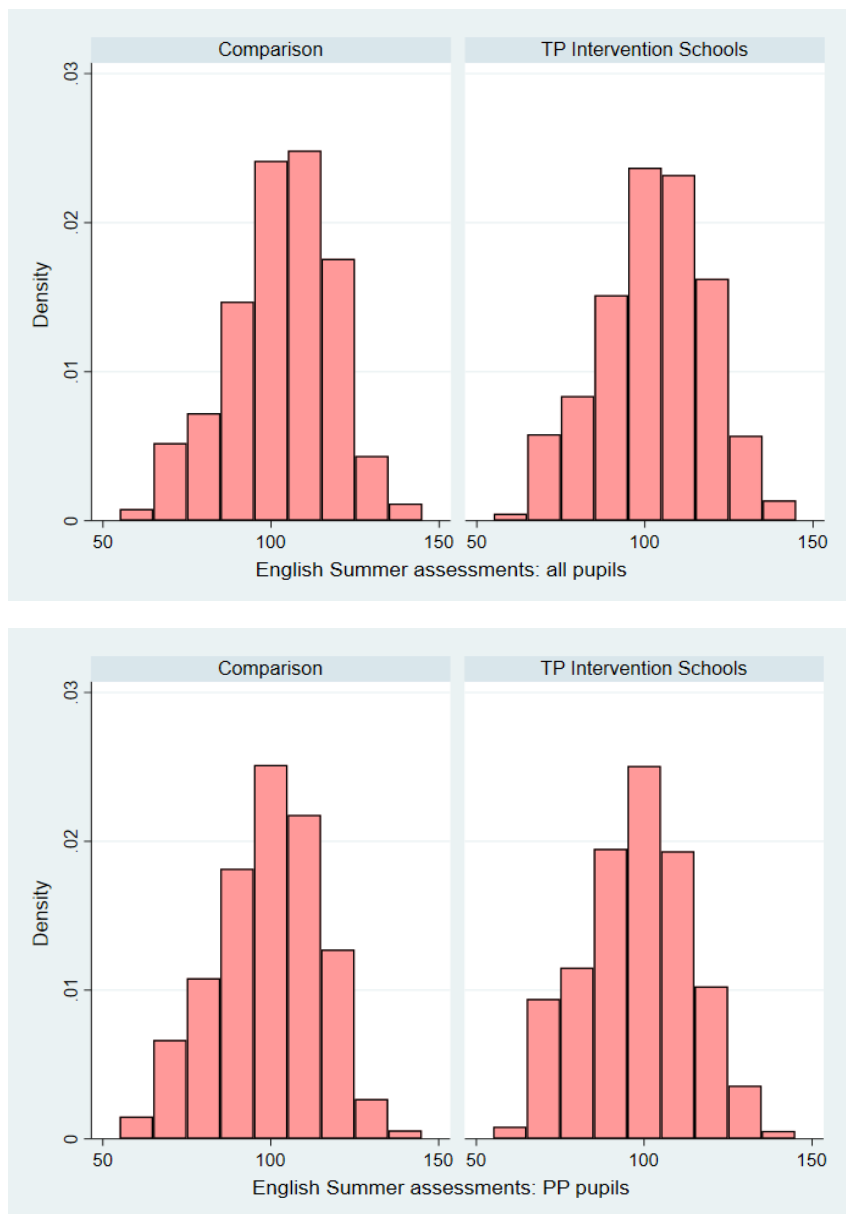


Figure 6: Histograms of the distributions of summer English assessment, all pupils and PP-eligible pupils, unweighted



Weighting/Regression (RQ1a): 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 (Table 17) and (ii) a categorical variable measuring the number of blocks of hours completed on average at the school level by the time of the assessment (dosage) (Table 18).

Results in Table 17 show the impact of TP availability in the school in maths and in English on PP-eligible pupils. TP is measured with a dummy equal to one if TP is available at school level. The coefficient of TP is not statistically significant and different from zero in the preferred weighted specification, nor in the unweighted specification for both maths and English. The size of the coefficient is less than one in all specifications. This converts to an effect size of -0.026 (CI: -0.030 to 0.079) for maths and -0.024 (CI: -0.100 to 0.052) for English (see Table 32). The effect sizes are the equivalent of no additional months' progress compared to PP-eligible pupils in comparison schools. These results must be considered in the context of dilution, which is that not all of those being analysed here (PP-eligible pupils) were selected for TP. The data for the primary school TP intervention sample indicated that in this group on average only 20% of PP-eligible pupils participated in TP and about 50% of TP pupils are eligible for PP: 'dilution' (see Table 13a). Therefore, this analysis does not properly represent the population of pupils who benefitted from the intervention. Selection into TP subject also further contributes to dilution (see below).

Table 17: Impact of TP measured with a 0/1 dummy indicating the availability of TP or not on the population of PP-eligible pupils

	Weighted						Unweighted					
	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 intervention	-0.711	0.828	0.392	-0.314	0.580	0.589	-0.410	0.854	0.632	-0.395	0.634	0.534
Constant	24.015	13.287	0.073	17.260	10.451	0.101	14.685	12.239	0.233	11.868	10.345	0.253
N	5102			7073			5102			7073		

Source: Primary school evaluation sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEN, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

The dosage analysis in Table 18 presents the school-level impact of TP on PP-eligible pupils when measured with dosage, computed at school level, and divided in categories indicating 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.³⁶ As noted later on in the report, tutoring is a pupil-level intervention and therefore the pupil-level analyses on dosage are likely to be more useful to schools and tutoring providers, particularly given the level of dilution. It has to be noted that we have not controlled for selection into dosage category, we cannot control for that level of selection and therefore we cannot view the dosage relationship as causal.

For English, levels of dosage are not correlated with attainment. For maths, in the weighted specification, levels of dosage less than one block are significantly positively correlated with attainment, and being a comparison school is also positively correlated with attainment. The pattern is similar in the unweighted specification, except for comparison schools not being significantly correlated with maths assessment. The result associated with comparison schools in the maths sample is not consistent with what we might expect, but as dosage is aggregated at school level, the measure cannot properly pick up the relationship between pupil-level dosage and attainment. Results should also be interpreted with caution given the aggregate nature of the dosage variable.

Overall, these results indicate that we have not been able to detect any significant impact of the availability of TP on PP-eligible pupils' attainment. Estimated coefficients are small and noisy (large standard errors), as dilution makes it hard to pick up the impact of the intervention on the sample of PP-eligible pupils.

Table 18: Association between TP measured with dosage and attainment on the population of PP-eligible pupils

	Weighted						Unweighted						
	Maths			English			Maths			English			
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	
Comparison schools	2.126*	1.001	0.036				1.726	0.936	0.068				
≤1 block of dosage	2.424*	1.175	0.041				2.039*	0.985	0.041				
>1 block of dosage	1.583	1.258	0.211				1.688	1.094	0.125				
Comparison schools				-0.388	0.757	0.609					-0.187	0.820	0.820
≤1 block of dosage				-1.751	0.957	0.069					-1.722	0.935	0.067
>1 block of dosage				0.138	1.694	0.935					0.548	1.486	0.713
Constant	24.842	13.315	0.065	16.834	3	0.111	15.293	11.931	0.202	10.364	9.787	0.291	
N	5102			7073			5102			7073			

Source: Primary school evaluation sample. Note: Dosage measured as less than one block of 12 hours and more than one block of 12 hours. Baseline for dosage is the sample of TP schools with zero dosage. Controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEN, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

³⁶ The calculation of dosage includes pupils with any TP hours, not just at least one hour. It also includes true zeros (non-blank).

Instrumental variables (RQ1c): What is the impact of the intensity of TP (dosage) on all PP-eligible pupils' attainment?

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 end-point 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 PP status. 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 percentage τ of a 'worst-case' benchmark (BM).

The results of the weak instrument test on the analysis on PP-eligible pupils and on TP pupils are reported in Table 19. For maths, the F statistic is less than the worst-case BM (37.4 with $\tau = 5\%$), indicating that the instrument is weak and that early date of signing up was not a good instrument in relation to dosage for maths. For English, the F statistic is higher than the worst-case BM (15.062 with $\tau = 20\%$), indicating that the instrument is not weak for the analysis on PP-eligible pupils and that early date of signing up was a good instrument in relation to dosage for English.

The different results for the two subjects indicate that the schools in the intervention group that provided high dosage by the time of the end-point assessment in English, are those that signed up earlier for the programme. However, in maths the pattern was more varied and some of the intervention schools provided high dosage in maths even if they signed up for the programme close to the end-point assessment date. Moreover, the different effect between the maths and English samples may be due to the fact that because of missing data in sessions dates, the maths sample used for this analysis includes only 22% of TP schools, while the English sample includes 38% of TP schools.

We proceeded with this part of the analysis for English (PP-eligible pupils and TP pupils), but it was not possible to do so for maths.

Table 19: Weak instrument test: PP-eligible pupils and TP pupils only

Weak IV tests: Maths – PP			Weak IV tests: English – PP		
Montiel–Pflueger robust	Weak instrument test		Montiel–Pflueger robust	Weak instrument test	
Effective F statistic:	0.098		Effective F statistic:	15.8	
Confidence-level alpha:	5%		Confidence-level alpha:	5%	
Critical values	TOLS	LIML	Critical values	TOLS	LIML
% of worst case bias			% of worst case bias		
$\tau = 5\%$	37.418	37.418	$\tau = 5\%$	37.418	37.418
$\tau = 10\%$	23.109	23.109	$\tau = 10\%$	23.109	23.109
$\tau = 20\%$	15.062	15.062	$\tau = 20\%$	15.062	15.062
$\tau = 30\%$	12.039	12.039	$\tau = 30\%$	12.039	12.039
Weak IV tests: maths – TP			Weak IV tests: English – TP		
Montiel–Pflueger robust	Weak instrument test		Montiel–Pflueger robust	Weak instrument test	
Effective F statistic:	0.028		Effective F statistic:	19.003	
Confidence-level alpha:	5%		Confidence-level alpha:	5%	
Critical values	TOLS	LIML	Critical values	TOLS	LIML
% of worst case bias			% of worst case bias		
$\tau = 5\%$	37.418	37.418	$\tau = 5\%$	37.418	37.418
$\tau = 10\%$	23.109	23.109	$\tau = 10\%$	23.109	23.109
$\tau = 20\%$	15.062	15.062	$\tau = 20\%$	15.062	15.062
$\tau = 30\%$	12.039	12.039	$\tau = 30\%$	12.039	12.039

Source: TP intervention sample.

Dosage is measured in blocks of sessions of 12 hours. PP-eligible pupils in TP schools are included in this analysis, with pupil-level dosage of TP participation (those that did not receive TP have dosage of zero) and a second analysis

using just TP pupils. Focusing on the sample of TP pupils in TP schools allows us to more closely identify individuals who benefit from the intervention and the association with the programme.

Table 20a shows the first and second stages of the IV regression of dosage of tutoring on English assessment for PP-eligible pupils and for all TP pupils. The first stage indicates that the instrument is significant at 1% level in explaining variation in dosage. The second stage suggests that higher dosage of tutoring is positively correlated with higher English assessment scores after tutoring. The effect is slightly larger when we restrict the analysis to the sample of all TP pupils. Amongst PP-eligible pupils, the impact of receiving one 12-hour block of tutoring is an eight-point increase in the end-point English assessment. Amongst pupils that participated in TP, this effect is slightly larger – a ten-point increase in the end-point English assessment. The effect is slightly more than 1 standard deviation increase (Table 16b indicates that the standard deviation of the end-point assessment is 7.5), with confidence intervals small enough to make the effect significant at 1% levels.

As discussed in the *Statistical analysis* section, we cannot fully test the hypothesis that the instrument is uncorrelated with the outcome of interest, the assessment. As a check, we estimated the 2SLS with a placebo outcome of the baseline assessment for PP-eligible pupils. Results, reported in Table 20b, indicate that dosage is not significantly correlated with baseline assessment (and the point estimate is much smaller than when using the final assessment), providing some reassurance that the hypothesis may hold and that higher dosage of tutoring (more hours) is positively correlated with a higher score in the English assessment after tutoring.

This analysis indicates that there is a positive and strong effect of early sign-up to the TP programme on number of hours of tutoring received and a positive effect of TP on English (reading) scores for those that received more hours of tutoring due to the early sign-up of the school to the programme.

Table 20a: IV regression of dosage of tutoring on English end-point assessment

	PP-eligible pupils						TP pupils					
	Coef	IV reg		Coef	First stage		Coef	IV reg		Coef	First stage	
		S.E.	p-value		S.E.	p-value		S.E.	p-value		S.E.	p-value
Pupil-level English dosage	8.406*	3.291	0.011			0.000	10.613***	2.486	0.000			
Days from MoU date to English assessment date				0.004 ***	0.001	0.005				0.005***	0.001	0.000
Constant	37.394*	15.735	0.017	-2.445**	0.870		67.113***	15.803	0.000	-2.039*	0.947	0.031
F-stat				27.2						19		
Observations	768			768			1625			1625		

Source: TP intervention sample.

Table 20b: IV regression of dosage of tutoring on baseline English assessment (PP-eligible pupils)

	IV reg			First stage		
	Coef	S.E.	p-value	Coef	S.E.	p-value
Pupil-level English dosage	-1.328	3.825	0.729			
Days from MoU date to English assessment date				0.004***	0.001	0.000
Constant	-0.523	14.797	0.972	-2.449**	0.873	0.005
Observations	768			768		

Source: TP intervention sample.

Further analyses

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

Before proceeding with RQ2, we estimated the participation equation on all pupils in TP schools to assess its predictive power. This involved modelling the probability of pupil participation in TP schools, using various markers of disadvantage recorded in the NPD (socio-economic status measured by FSM, SEND). Results in Table 21 show the pupil-level participation equations for TP maths and TP English. Lower baseline assessment scores, being eligible for FSM and having SEN are positively and significantly correlated with predicted participation. The pseudo *R*-squared is not provided, as the model is estimated with imputed data.

Table 22 shows the predictive power of the model, by tabulating pupils predicted to participate versus pupils who actually participated: 41% (maths) and 26% (English) of pupils predicted to participate did not actually participate in the intervention, while 41% (maths) and 64% (English) of those predicted not to participate were selected to receive the intervention. These numbers 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.

Table 21: Estimation of TP participation using Logit

	Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value
Math scores autumn assessment	-0.038***	0.004	0.000			
English scores autumn assessment				-0.038***	0.004	0.000
Missing autumn maths score	0.040	0.480	0.934			
Missing autumn English score				0.292	0.395	0.460
Not matched to Census 2021	0.530	0.419	0.205	0.559	0.377	0.139
Female: Yes	-0.041	0.044	0.348	0.056	0.038	0.141
Missing female in Census	-0.306	0.964	0.751	-0.320	0.792	0.686
FSM: Yes	0.673***	0.118	0.000	0.557***	0.091	0.000
Ethnicity (unknown as base):						
White British	0.337	0.335	0.314	0.228	0.177	0.197
Asian	0.211	0.397	0.595	0.051	0.239	0.832
Black	0.382	0.375	0.309	0.469*	0.235	0.046
Other ethnicities	0.272	0.339	0.422	0.106	0.173	0.540
EAL (No as base):						
EAL: Yes	-0.133	0.083	0.108	-0.194**	0.073	0.008
EAL: Missing	0.514	0.471	0.274	0.062	0.329	0.850
Year group (Y1 as base):						
Y2	0.335	0.273	0.220	0.256	0.213	0.229
Y3	0.103	0.278	0.713	0.210	0.272	0.440
Y4	-0.122	0.260	0.640	0.099	0.260	0.705
Y5	0.306	0.288	0.287	0.433	0.263	0.100
Y6	-0.043	0.305	0.888	0.156	0.290	0.592
SEN (No as base):						
SEN: Yes	-0.463***	0.101	0.000	-0.315***	0.094	0.001
Looked after for 12 months	0.869	0.869	0.317	0.420	0.664	0.527
Looked after since 31 March	0.545	0.681	0.424	0.021	0.513	0.967
Looked after for 6 months	-0.923	0.964	0.338	0.422	0.704	0.549
Ofsted 2018 (outstanding as base):						
Good	-0.435	0.353	0.218	-0.261	0.374	0.484
Inadequate	0.010	0.531	0.985	-0.292	0.578	0.613
Requires improvement	-0.009	0.512	0.986	0.252	0.459	0.583
Ofsted missing	0.363	0.434	0.403	0.005	0.480	0.991
School %FSM high vs low	0.088	0.368	0.810	0.314	0.336	0.350
Urban vs rural:						
Urban	-0.533	0.308	0.083	0.044	0.297	0.884
Pupil-level IDACI quintiles (Q1 as base):						
Q2	-0.102	0.123	0.405	-0.205	0.105	0.051
Q3	-0.132	0.134	0.327	-0.098	0.109	0.369
Q4	0.097	0.147	0.507	0.140	0.106	0.186
Q5	0.230	0.153	0.131	0.155	0.121	0.201
IDACI Missing	-0.081	0.771	0.916	-0.123	0.564	0.827
Region (East midlands as base):						
East of England	1.172**	0.372	0.002	0.831*	0.340	0.015
London	1.758***	0.495	0.000	1.644***	0.349	0.000
North East	1.171**	0.428	0.006	1.089*	0.445	0.014
North West	0.931*	0.431	0.031	0.556	0.381	0.144
South East	0.077	0.483	0.873	0.231	0.389	0.553
South West	2.044***	0.427	0.000	0.852*	0.353	0.016
West Midlands	1.578*	0.647	0.015	0.344	0.341	0.314
Yorkshire and the Humber	1.413***	0.349	0.000	1.134***	0.299	0.000
Missing	1.520	1.009	0.132	-0.814	0.519	0.117
School types (Community schools as base):						
Converter Academies	-0.443	0.236	0.060	-0.148	0.233	0.527
Foundation School	-0.550	0.549	0.317	-0.500	0.278	0.072
Free Schools	0.538	0.911	0.555	0.429	0.521	0.410
Sponsored Academies	-0.073	0.558	0.895	-0.367	0.332	0.269
Voluntary Aided School	-0.157	0.431	0.716	-0.112	0.469	0.812
Voluntary Controlled School	-0.655	0.337	0.052	-0.045	0.249	0.855
School-level average:						
Maths autumn assessment	0.007	0.021	0.748			
School %FSM	1.795	1.213	0.139	0.741	1.218	0.543
School %EAL	-1.403*	0.550	0.011	-1.299**	0.488	0.008

	Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value
School %SEN	-0.593	1.065	0.578	-0.176	1.278	0.891
School-level IDACI tertile (T1 as base)						
T2	-0.327	0.330	0.322	-0.270	0.337	0.424
T3	-1.119**	0.433	0.010	-0.645	0.436	0.139
Pupil-to-teacher ratio 2018 tertiles (T1 as base):						
T2	-0.502	0.279	0.071	-0.179	0.242	0.458
T3	-0.200	0.252	0.427	0.017	0.261	0.949
Tertile missing	-2.207*	0.967	0.022	-0.288	0.468	0.538
School-level average KS2 Maths 2018 tertiles (T1 as base):						
T2	0.293	0.256	0.253			
T3	-0.144	0.361	0.690			
Tertile missing	-1.067	0.718	0.138			
Provider types: Maths autumn assessments (GL Assessment as base):						
NFER	-0.217	0.779	0.781			
Old SATs	-0.207	0.830	0.803			
Other	0.000	.	.			
Renaissance Learning	-0.318	0.833	0.702			
Rising Stars	0.225	0.739	0.761			
Provider missing	-0.534	0.944	0.572			
Provider types: Maths summer assessments (GL Assessment as base):						
NFER	-0.061	0.714	0.932			
Old SATs	-0.153	0.733	0.835			
Renaissance Learning	0.796	0.777	0.305			
Rising Stars	-0.112	0.688	0.871			
Provider missing	-0.004	0.644	0.994			
School-level average English autumn assessment				0.032	0.019	0.089
School-level average KS2 Reading 2018 tertiles (T1 as base):						
T2				0.013	0.241	0.958
T3				-0.411	0.353	0.244
Tertile missing				-0.225	0.471	0.633
Provider types: English autumn assessments (GL Assessment as base):						
NFER				0.056	0.568	0.922
Old SATs				-0.021	0.614	0.973
Other				0.423	0.847	0.618
Renaissance Learning				-0.586	0.613	0.339
Rising Stars				-0.157	0.538	0.771
Provider missing				-0.724	0.692	0.296
Provider types: English summer assessments (GL Assessment as base):						
NFER				0.078	0.512	0.879
Old SATs				-0.299	0.633	0.637
Other				0.000	.	.
Renaissance Learning				0.999	0.589	0.090
Rising Stars				0.422	0.529	0.424
Provider missing				0.045	0.474	0.925
Constant	-0.236	2.703	0.930	-3.426	2.659	0.198
N	11875			16813		

Source: TP intervention sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, Looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEN, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

Table 22: Predictive power of participation model: percentages of predicted TP participations by actual TP participations; adjusted by multiple imputation estimation³⁷

Maths:	Percentage	Standard error
Predicted non-TP & actual non-TP	59	0.003
Predicted non-TP & actual TP	41	0.008
Predicted TP & actual non-TP	41	0.003
Predicted TP & actual TP	59	0.008
English:	Percentage	Standard error
Predicted non-TP & actual non-TP	74	0.003
Predicted non-TP & actual TP	64	0.008
Predicted TP & actual non-TP	26	0.003
Predicted TP & actual TP	36	0.008

Source: TP intervention sample.

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

Similar to the approach for RQ1a (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 (Table 23) and (ii) a categorical variable measuring the percentage of hours completed by the time of the assessment (dosage) again at the school level (Table 24).

Results in Table 23 show that the impact of TP on maths and English on the population of all pupils is not statistically significant, in the preferred weighted specification. Results are the same in the unweighted specification. Estimated coefficients are less than one and have large standard errors. The effect sizes are 0.050 (–0.03 to 0.133) for maths and 0.001 (–0.063 to 0.065) for English (see Table 32), which according to the EEF guidance represents one month's additional progress in maths, but no additional month's progress in English. However, there is uncertainty around these estimates, with the positive maths result being consistent with a null (0 months) or slightly larger positive effect (2 months) and the English result being consistent with small positive (1 month) or small negative effect (–1 months). Although the month's progress conversion suggests a positive effect in maths, we note that these results are not statistically significant. We also highlight the caveats around this analysis, in particular the high level of dilution in the all-pupils analysis, with only 12% pupils selected for maths TP and 14% selected for English. In this analysis, therefore, most of the pupils analysed were not selected for, or receiving, TP. This finding is discussed further in the *Interpretation* section of the *Conclusion*.

Table 23: Impact of TP measured with a 0/1 dummy indicating the availability of TP or not on all pupils

	Weighted						Unweighted					
	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 intervention	0.769	0.688	0.266	0.204	0.496	0.682	0.839	0.707	0.238	0.012	0.543	0.982
Constant	38.405**	13.200	0.004	37.037***	9.700	0.000	28.293*	12.637	0.027	30.880**	9.864	0.002
N	15,393			21,879			15,393			21,879		

Source: Primary school evaluation sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, Looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEN, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

The school-level dosage analysis in Table 24 presents the impact of TP on all pupils when measured with dosage, computed at school level. Dosage was divided into categories indicating 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 level of dosage (see Figures 3 and 4 in RQ1a for histograms of the distribution of dosage).

For maths and English, in the weighted specification, levels of dosage at school level are not significantly correlated with outcomes when looking at the impact on all pupils. Only less than one block of dosage is negatively correlated with English assessment scores with respect to no dosage. Similar to the PP-eligible pupil analysis (RQ1a), the results are not consistent with what we might expect, and this may be due to the fact that dosage is aggregated at the school level.

³⁷ Table 22 shows the distribution of pupils in each category and not the numbers, as the data is imputed and counts would be misleading.

The results on the full sample of the population consider all pupils, which includes all TP pupils whether they are eligible for PP or not. The issue of dilution becomes even more problematic in the analysis on all pupils as only a small percentage of pupils among all pupils were selected to receive TP (see Table 13a), as TP pupils are a subset of each year group. It is therefore harder to detect impacts of TP when looking at all pupils.

We point towards the pupil-level dosage analysis reported earlier (RQ1c: What is the impact of the intensity (dosage) of TP on the attainment of PP-eligible pupils?) and later (RQ6: How do outcomes vary among TP pupils, by model of tutoring?), which are likely to be more useful for schools and tuition providers.

Table 24: Association between TP measured with dosage and attainment on all pupils

	Weighted						Unweighted					
	Maths			English			Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
Comparison schools	-0.338	0.887	0.704				-0.614	0.861	0.477			
≤1 block of dosage	0.945	1.104	0.394				0.614	0.974	0.530			
>1 block of dosage	0.374	1.210	0.758				-0.039	1.047	0.970			
Comparison schools				-0.867	0.636	0.175				-0.562	0.669	0.403
≤1				-1.476*	0.737	0.047				-1.458*	0.701	0.039
>1				-0.336	1.166	0.774				0.094	1.093	0.932
Constant	39.783**	13.517	0.004	38.358***	9.664	0.000	29.714*	12.884	0.023	31.226**	9.616	0.001
N	15,393			21,879			15393			21,879		

Source: Primary school evaluation sample. Note: Dosage measured as less than 1 block of 12 hours and more than a block of 12 hours. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, Looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEN, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 maths and Read 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

Moderator analysis:

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

Tables 25a (maths) and 25b (English) present the results of the interaction between TP schools and a set of school-level, pupil-level and geographic characteristics, for PP-eligible pupils.

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

We attempted to estimate margins but as the data are multiple imputed there is an identifiability issue (empty cells in one of the strates of the combination of the covariates). With multiple imputed data, the empty cell cannot be reweighted as it would be in a dataset without multiple imputed data. The solution would imply to present each interaction as a separate model, which is a departure from the study plan; hence we present the estimates of the interaction analysis and not the marginal effects.

The findings indicate that:

- None of the regressions with the interactions of TP and school-level characteristics are significant for maths. For English, high Ofsted rating is associated with higher English scores in TP intervention schools compared to comparison schools.
- Being female is associated with lower maths scores in TP schools compared to comparison schools. There is no evidence in the IPE to suggest that tutoring had particular challenges for girls.³⁸
- Having missed the baseline assessment is associated with lower maths scores in TP schools compared to comparison schools.
- Being Asian and Being Black are associated with lower English scores in TP schools compared to comparison schools. There is no evidence in the IPE to suggest that tutoring had particular challenges for ethnic minorities.
- TP schools in more disadvantaged areas (those with an IDACI score above the median) are associated with better maths and English assessments than comparison schools in the same type of area.
- TP schools in rural areas perform better in maths than comparison schools in rural areas.

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 around 20% of PP-eligible pupils participated in TP (Table 13a)).

³⁸ We also ran the model separately for males and females, and TP is significant and negatively correlated with maths in the sample of females only, and not significant in the sample of males.

Table 25a: How does the association of TP availability with attainment vary among PP-eligible pupils, by school and pupil characteristics in the maths sample?

	Interact w/ school level			Interact w/ pupil level			Interact w/ geography		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
TP intervention	0.347	2.546	0.892	-2.157	6.119	0.725	-6.355**	2.243	0.005
TP intervention # Ofsted rating high vs low	-0.701	1.986	0.725						
TP intervention # Ofsted rating missing	-2.492	3.498	0.478						
TP intervention # School %FSM high vs low	-2.759	2.027	0.176						
TP intervention # Maintained schools	0.274	2.003	0.891						
TP intervention # School type missing	0.000	.	.						
TP intervention # Q2 size school	4.375	2.535	0.087						
TP intervention # Q3 size school	2.039	2.697	0.451						
TP intervention # Q4 size school	0.845	2.535	0.739						
TP school X pupil level:									
TP intervention # Math scores autumn assessment				-0.017	0.058	0.768			
TP intervention # missing Math scores autumn assessment				-7.370*	3.190	0.023			
TP intervention # Female				-1.246*	0.541	0.023			
TP intervention # SEN ever				-0.457	1.170	0.697			
TP intervention # EAL				1.289	0.959	0.181			
TP intervention # EAL missing				-0.695	3.296	0.834			
TP intervention # White British				2.446	2.896	0.401			
TP intervention # Asian				3.855	3.401	0.260			
TP intervention # Black				-0.829	3.263	0.800			
TP intervention # Other ethnicities				2.214	3.069	0.473			
TP intervention # Year group = 2				2.648	3.754	0.482			
TP intervention # Year group = 3				0.551	3.089	0.859			
TP intervention # Year group = 4				2.506	3.311	0.451			
TP intervention # Year group = 5				2.105	3.125	0.502			
TP intervention # Year group = 6				0.051	3.597	0.989			
TP school X Geography									
TP intervention # Rural							5.512*	2.429	0.025
TP intervention # IDACI rank high vs low							1.916*	0.867	0.029
Constant	25.532*	11.531	0.029	20.675	10.923	0.061	25.203*	11.319	0.028
Observations	5102			5102			5102		

Source: Primary school evaluation sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, Looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEN, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

Note that we cannot estimate margins in the primary school analysis because of multiple imputed data. Therefore, the variables in this table are to be interpreted with respect to the omitted categories. The omitted categories (males, low Ofsted rating, low FSM, smallest school size, no SEN, no EAL and so on) are all in the constant.

Table 25b: How does the association of TP availability with attainment vary among PP-eligible pupils, by school and pupil characteristics in the English sample?

	Interact w/ school level			Interact w/ pupil level			Interact w/ geography		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
TP intervention	-3.920	2.317	0.093	3.004	5.887	0.611	0.406	2.335	0.862
TP intervention # Ofsted rating high vs low	4.482*	1.840	0.016						
TP intervention # Ofsted rating missing	5.166	2.926	0.079						
TP intervention # School %FSM high vs low	0.582	1.607	0.718						
TP intervention # Maintained schools	2.040	1.551	0.190						
TP intervention # School type missing	0.000	.	.						
TP intervention # Q2 size school	3.634	2.370	0.127						
TP intervention # Q3 size school	-4.383	2.240	0.052						
TP intervention # Q4 size school	-0.651	2.123	0.760						
TP school X pupil level:									
TP intervention # English scores autumn assessment				-0.001	0.036	0.969			
TP intervention # missing English scores autumn assessment				0.701	2.997	0.816			
TP intervention # Female				-0.620	0.545	0.257			
TP intervention # SEN ever				-1.069	0.870	0.221			
TP intervention # EAL				0.348	0.852	0.683			
TP intervention # EAL missing				-7.118	3.617	0.056			
TP intervention # White British				-5.960	3.222	0.066			
TP intervention # Asian				-6.881*	3.438	0.047			
TP intervention # Black				-7.432*	3.082	0.017			
TP intervention # Other ethnicities				-6.117	3.434	0.077			
TP intervention # Year group = 2				3.395	3.281	0.302			
TP intervention # Year group = 3				4.001	2.945	0.176			
TP intervention # Year group = 4				4.729	2.795	0.093			
TP intervention # Year group = 5				2.799	2.960	0.346			
TP intervention # Year group = 6				1.535	2.621	0.559			
TP school X Geography									
TP intervention # Rural							-1.517	2.370	0.523
TP intervention # IDACI rank high vs low							1.702*	0.750	0.025
Constant	27.356**	8.429	0.001	14.184	10.829	0.192	16.051	9.414	0.090
Observations	7073			7073			7073		

Source: Primary school evaluation sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEN, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

Note that we cannot estimate margins in the primary school analysis, because of multiple imputed data. Therefore, the variables in the table are to be interpreted with respect of the omitted categories. The omitted categories (males, low Ofsted rating, low FSM, smallest school size, no SEN, no EAL and so on) are all in the constant.

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 comparison schools. Consequently, the results reported here should be considered as descriptive findings that may be indicative (but not causal). The benefit of this analysis over RQ5 is that this is on the pupils that were selected for the intervention (i.e., TP pupils).

In this analysis, each observation is identified as a 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 Table 26a (looking at pupil-level TP characteristics on maths and English scores); and in Table 26b at the school level (school-level TP characteristics); and at the tutor level (tutor-level characteristics). We control for the same set of controls used in previous specifications and listed in Appendix B. For each subject analysis, the sample includes only pupils who received tutoring in that specific subject.

The key findings are summarised below.

- Higher pupil-level dosage is associated with higher English scores. Pupil-level dosage is measured as a continuous variable to avoid capturing non-linearities associated with the definition of dosage in blocks. School-level dosage higher than a block is correlated with higher maths scores (Table 26b).
- Group size is positively associated with scores for English: attending face-to-face sessions with two other pupils is associated with higher English scores than attending a session alone. For online sessions: attending the session with another one or two pupils is associated with better scores than a pupil-to-teacher ratio of 1:1 for English.
- Group size is not associated with a difference in maths scores in either online or face-to-face sessions.
- Scheduling face-to-face TP sessions to all take place within school hours is associated with better English scores than sessions delivered in a combination of both during and outside schooling hours. The timing of delivery made no difference for maths tuition.
- Online sessions are associated with better English scores than face-to-face sessions in English. There is no association between mode of tutoring and maths scores.
- Delivery of tutoring sessions concentrated over a short timeframe (intensity)³⁹ is positively correlated with maths scores. There is no difference for English.
- The number of sessions schools buy for pupils (these bought sessions may or may not be completed), high if above the median, is positively correlated with higher English scores.
- When looking at buy-in at school level (Table 26b), in schools that buy a higher number of sessions, with high equal to one if above the median, pupils have lower maths scores. It appears from the earlier checks for RQ1c (dosage IV) that the relationship between signing up and completing sessions is not straightforward for maths and therefore this result is not one of particular concern at this stage given the TP delivery continued for several months after the assessments.

³⁹ 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 missingness.

When looking at tutor characteristics and the tutor's highest qualification (we selected postgraduate degree as base as the highest possible qualification):

- tutors with an undergraduate degree are associated with higher performance in maths than tutors with a postgraduate degree. Tutors with a highest qualification of qualified teacher status and a PGCE are associated with higher performance in English compared with tutors with a postgraduate degree. It appears that having specialised postgraduate qualifications (rather than undergraduate or QTS) may not deliver better tutoring to primary school children.
- tutors who had received on-going tutor training are associated with higher scores in maths.
- the comparison between tutor and pupil characteristics shows that sharing the same ethnicity is associated with lower English scores.

Table 26a: How do outcomes vary among TP pupils, by model of tutoring, pupil-level TP characteristics?

	Maths, pupil level clustered residuals			English, pupil level clustered residuals		
	Coef	S.E.	P-value	Coef	S.E.	P-value
Math scores autumn assessment	0.637***	0.038	0.000			
English scores autumn assessment				0.498***	0.026	0.000
Pupil-level dosage:						
TP Maths dosages	1.744	1.593	0.274			
TP English dosages				3.414***	0.875	0.000
Delivery mode (face-to-face as base):						
Online delivery	4.486	5.683	0.430	8.405***	2.548	0.001
F2F timing (all during lessons as base):						
mixed				-17.836***	2.639	0.000
Online timing (all during lessons as base):						
all outside of lessons	0.894	2.024	0.659	-3.787	2.723	0.165
mixed	-0.316	0.984	0.748	-3.275	1.820	0.072
F2F tutor-pupil ratio (1:1 as base):						
1:2	-0.055	5.061	0.991	-2.425	8.857	0.784
1:3	7.767	6.682	0.246	11.881*	5.835	0.042
Online tutor-pupil ratio (1:1 as base):						
1:2	5.574	3.116	0.074	6.103**	2.228	0.006
1:3 and below	1.967	3.140	0.531	5.398**	1.929	0.005
Pupil-level bought hours (low as base):						
High buy-in	-2.009	1.373	0.144	6.272***	1.377	0.000
Maths Completed vs scheduled: high =1	-1.242	1.172	0.290			
English Completed vs scheduled: high =1				0.813	0.972	0.403
Intensity	216.183**	77.789	0.006	-42.620	29.973	0.155
Early delivery	1.218	0.972	0.211	-0.546	0.710	0.442
Constant	12.054	9.338	0.198	30.544***	7.576	0.000
N	739			1622		

Source: TP intervention sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEND, Looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEND, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. We control for missing delivery information but not report the coefficients. Pupil-level clustered residuals.

Table 26b: How do outcomes vary among TP pupils, by model of tutoring, school-level and tutor-level TP characteristics?

	Maths, school-level clustered residuals			English school-level clustered residuals			Maths, tutor-level clustered residuals			English tutor-level clustered residuals		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
Math scores autumn assessment	0.639***	0.058	0.000				0.625***	0.054	0.000			
English scores autumn assessment				0.535***	0.044	0.000				0.496***	0.067	0.000
School-level maths dosages (zero as base):												
≤1	2.408	1.483	0.110									
>1	3.280*	1.563	0.040									
School-level English dosages (zero as base):												
≤1				-3.624	2.186	0.102						
>1				-0.771	2.352	0.744						
School-level bought hours (low as base):												
High buy-in	-3.146*	1.241	0.014	-1.217	1.615	0.453						
Tutor qualifications (Postgrad as base):												
Undergraduate							2.435*	1.082	0.025	2.945	1.666	0.079
QTS							-1.558	1.472	0.291	5.866*	2.348	0.014
A levels							2.099	1.345	0.120	-1.384	3.073	0.653
Missing/Other							3.156	1.756	0.073	1.567	3.678	0.671
PGCE										5.490*	2.326	0.020
On-going							11.396***	3.367	0.001	3.132	2.213	0.159
Tutor: same ethnicity (no as base)												
Same							1.476	1.228	0.230	-3.454*	1.603	0.033
Tutor: same gender (no as base)												
Same							-0.224	0.737	0.761	-0.258	1.351	0.849
Constant	28.870***	6.552	0.000	48.979***	5.093	0.000	28.846***	6.804	0.000	49.200***	6.832	0.000
N	750			1869			680			906		

Source: TP intervention sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEN, Looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEND, pupil-to-teacher ratio 2018 tertiles, School-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. We control for missing delivery information but do not report the coefficients. School-level and tutor-level clustered residuals.

Missing data analysis

The extent of missingness in the end-point assessment is indicated in the flow-diagrams shown in Figures 1 and 2 and it is quite large. As noted in the *Participant Selection* section, we did not require nor expect schools to test in both subjects to be eligible for the evaluation.⁴⁰ However, due to the way the analysis was constructed, schools that provided test data for only one subject are indicated as having missing data for the other subjects here. In this section, we analyse the school-level and pupil-level pattern of missingness to explore that the pattern is not different between TP intervention and comparison schools. In Table 27 we present the results of the regression on a dummy equal to 1 if the school did not provide test scores in maths (columns 1 and 2) or English (columns 3 and 4) and zero otherwise, on a set of school-level and pupil-level characteristics.

The regressions indicate that school-level characteristics are not in general significant in explaining the pattern of missingness across schools testing. The exception is 'TP intervention', which suggests that the odds of having missing maths scores decreased by a factor of 0.35 for maths and by a factor of 0.25 for English with respect to the sample of comparison schools. Voluntary Aided Schools are significantly more likely not to have provided test data in maths. Inadequate Ofsted rating and foundation schools perfectly predict missing English outcomes and their coefficients are not displayed.

The exploration of pupil-level characteristics indicates that schools with a higher percentage of SEN pupils are significantly more likely to have provided test results in English. For both subjects, a higher percentage of pupils of Black ethnicity is associated with a higher probability of supplying test results. For English, also a higher percentage of pupils of Asian ethnicity is associated with a higher probability of supplying test results. A higher percentage of pupils in Year 3 on roll is associated with a higher probability of not supplying test results, and for English, in particular, also Years 1, 4 and 5.

In general, school-level observable characteristics can explain between 17.1% (for maths) and 23.8% (for English) of the school-level variability in missing testing and only a few variables are significant in explaining the pattern.

⁴⁰ It is possible that some schools only opted to send us test data for the subject in which they were tutoring. Although we opened the options for schools to supply data for more year groups, initial requests to schools were for year group/subject combinations participating in TP. We do not have information about whether the schools did or did not test in the other subject.

Table 27: School-level drop-out model (explore the extent of missingness by counting the observations for which the assessment variables are missing, and evidence of the potential mechanism (cross-tabulation and 'drop-out' model)

	Schools missing in maths			School missing in maths (with controls)			Schools missing in English			School missing in English (with controls)		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
TP intervention schools	-1.046**	0.394	0.008	-1.121*	0.468	0.017	-0.952	0.489	0.052	-1.389**	0.527	0.008
School-level average KS2 Maths 2018 tertiles (T1 as base):												
T2	-0.118	0.385	0.760	-0.097	0.406	0.812						
T3	0.248	0.499	0.620	0.056	0.577	0.922						
Tertile missing	-0.433	0.920	0.638	-0.614	1.193	0.607						
School-level average KS2 Read 2018 tertiles (T1 as base):												
T2							-0.707	0.510	0.165	-0.659	0.636	0.301
T3							0.012	0.602	0.984	-0.418	0.694	0.547
Tertile missing							-0.941	1.163	0.419	-0.373	1.157	0.747
Total pupil counts	0.001	0.001	0.403	0.001	0.002	0.461	0.001	0.001	0.673	0.000	0.002	0.813
Pupil-to-teacher ratio	-0.025	0.048	0.610	-0.060	0.054	0.266	-0.026	0.059	0.655	-0.066	0.074	0.373
Ofsted 2018 (outstanding as base):												
Good	0.781	0.568	0.169	0.841	0.615	0.172	0.876	0.593	0.139	0.853	0.655	0.193
Inadequate	2.160	1.185	0.068	2.357*	1.193	0.048	0.000	.	.	0.000	.	.
Requires improvement	0.709	0.717	0.322	0.733	0.772	0.342	0.852	0.786	0.278	0.707	0.867	0.415
Missing	0.532	1.002	0.595	0.685	1.123	0.542	2.379	1.676	0.156	2.666	2.282	0.243
School types (Community schools as base):												
Converter Academies	0.105	0.428	0.807	0.165	0.468	0.725	-0.387	0.531	0.466	-0.469	0.594	0.430
Foundation School	0.200	0.863	0.817	0.378	1.116	0.735	0.000	.	.	0.000	.	.
Free Schools	0.230	1.045	0.825	-0.312	1.052	0.767	0.241	1.019	0.813	0.379	1.387	0.784
Sponsored Academies	-0.083	0.550	0.880	-0.094	0.645	0.885	-0.264	0.718	0.713	-0.594	0.992	0.549
Voluntary Aided School	1.427*	0.610	0.019	2.312**	0.724	0.001	0.195	0.634	0.758	0.540	0.763	0.480
Voluntary Controlled School	0.143	0.741	0.847	0.135	0.752	0.857	0.777	1.164	0.504	1.064	1.353	0.432
Missing	-0.243	1.958	0.901	-1.493	2.027	0.461	-2.747	2.017	0.173	-6.307*	2.668	0.018
Urban vs Rural												
Urban	-0.001	0.489	0.999	0.139	0.673	0.836	-0.207	0.602	0.731	-0.444	0.749	0.553
Region (East Midlands as base):												
East of England	-0.975	0.709	0.169	-1.465	0.788	0.063	-1.236	0.940	0.188	-2.320*	1.085	0.032
London	-1.527	0.798	0.056	-0.774	1.063	0.467	-1.505	1.005	0.134	-1.919	1.219	0.115
North East	-0.723	0.950	0.446	-1.281	0.992	0.196	-1.026	1.209	0.396	-2.304	1.399	0.100
North West	-0.950	0.742	0.200	-1.131	0.916	0.217	-0.308	0.972	0.751	-0.479	1.041	0.646
South East	-1.270	0.786	0.106	-1.763	0.930	0.058	-0.797	1.076	0.459	-1.790	1.204	0.137
South West	-1.379	0.735	0.061	-1.823	0.933	0.051	-0.893	0.939	0.342	-1.763	1.059	0.096
West Midlands	-1.203	0.782	0.124	-1.232	0.887	0.165	-1.314	1.000	0.189	-2.288*	1.091	0.036
Yorkshire and the Humber	-0.709	0.697	0.309	-0.860	0.817	0.293	0.373	1.036	0.719	-0.058	1.051	0.956
Missing region	0.000	.	.	0.000	.	.	0.000	.	.	0.000	.	.
Pupil-to-teacher ratio missing	0.000	.	.	0.000	.	.	0.000	.	.	0.000	.	.
Total pupil count missing	0.000	.	.	0.000	.	.	0.000	.	.	0.000	.	.

	Schools missing in maths			School missing in maths (with controls)			Schools missing in English			School missing in English (with controls)		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
% EAL				3.109	2.480	0.210				4.254	3.053	0.163
% SEN				-0.703	1.524	0.644				-6.210**	1.957	0.002
% Female				3.423	2.659	0.198				0.175	2.604	0.946
Average IDACI scores				-1.177	2.255	0.602				-0.825	3.214	0.797
% White British				-4.424	3.980	0.266				-6.904	3.979	0.083
% Asian				-7.785	4.678	0.096				-11.243*	4.733	0.018
% Black				-12.285*	5.329	0.021				-10.999*	4.912	0.025
% Other ethnic				-3.183	4.162	0.444				-7.380	4.187	0.078
% Unknown ethnic				-3.866	7.075	0.585				-0.356	8.525	0.967
% Year 1				4.412	4.079	0.279				17.709***	5.322	0.001
% Year 2				5.063	3.861	0.190				7.074	4.845	0.144
% Year 3				14.413**	5.415	0.008				20.451***	5.964	0.001
% Year 4				1.620	5.017	0.747				10.257*	5.124	0.045
% Year 5				4.751	2.956	0.108				8.741*	4.108	0.033
% Year 6				0.000	.	.				0.000	.	.
% Not White				0.000	.	.				0.000	.	.
Constant	1.571	1.357	0.247	0.730	3.295	0.825	2.933*	1.471	0.046	3.853	3.452	0.264
Observations	206			206			191			191		
Pseudo R-squared	0.090			0.171			0.115			0.238		

Source: Primary school evaluation sample. Note: controls include pupil-level variables: baseline scores, gender, FSM, dummy for ethnic group (White, Black, Asian, other), EAL, SEND, looked after, year group and pupil-level IDACI quintile. School-level variables: region, Ofsted rating, urban, school with high FSM (1 if above median), school type, school-level IDACI tertile, %FSM, %EAL, %SEND, pupil-to-teacher ratio 2018 tertiles, school-level average KS2 Maths and Reading 2018 tertiles, assessment provider for baseline and end-point assessment. School-level clustered residuals.

We also explored the pattern of missingness among pupils in schools with non-missing assessment. Table 28 shows the percentage of pupils with missing tests in schools that submitted test results for the subject. This table indicates that 24.4% of pupils did not have test data for maths in schools that submitted maths test data and 20.4% did not have test data for English in schools that submitted English test data.

Table 29 presents the distribution of pupils with missing assessments in schools with non-missing assessment by year group. The percentage of pupils that did not test in maths and English is higher for pupils in Years 4, 5 and 6 than for pupils in Years 1, 2 and 3. This shows that schools did not provide test results for pupils across all year groups.

Table 28: PP-eligible pupil-level drop-out in summer assessment in non-missing schools

	Maths		English	
	<i>N</i>	%	<i>N</i>	%
Not missing	5242	75.599	7076	79.622
Missing summer scores	1692	24.401	1811	20.378
Total	6934	100.000	8887	100.000
Observations	6934		8887	

Source: Primary school evaluation sample.

Table 29: PP-eligible Pupil-level drop-out in summer assessment in non-missing schools, by year group

English non-missing schools				Maths non-missing schools			
Year group	Comparison	TP intervention	Total	Year group	Comparison	TP intervention	Total
Y1, <i>N</i>	91	93	184	Y1, <i>N</i>	120	19	139
Y1, %	17.01	7.71	10.57	Y1, %	13.94	2.31	8.26
Y2, <i>N</i>	79	115	194	Y2, <i>N</i>	142	89	231
Y2, %	14.77	9.54	11.14	Y2, %	16.49	10.83	13.73
Y3, <i>N</i>	76	188	264	Y3, <i>N</i>	131	102	233
Y3, %	14.21	15.59	15.16	Y3, %	15.21	12.41	13.84
Y4, <i>N</i>	67	306	373	Y4, <i>N</i>	130	195	325
Y4, %	12.52	25.37	21.42	Y4, %	15.1	23.72	19.31
Y5, <i>N</i>	64	243	307	Y5, <i>N</i>	148	183	331
Y5, %	11.96	20.15	17.63	Y5, %	17.19	22.26	19.67
Y6, <i>N</i>	158	261	419	Y6, <i>N</i>	190	234	424
Y6, %	29.53	21.64	24.07	Y6, %	22.07	28.47	25.19
Total	535	1206	1741	Total	861	822	1683
	100	100	100		100	100	100

Source: Primary school evaluation sample.

When exploring the impact of observable characteristics on the probability of a pupil missing the test (Table 30), missing baseline assessment is dropped as perfectly collinear with missing outcomes (i.e., pupils who did not have tests for a subject at baseline did not have the test for the same subject at end-point). As not all schools tested in both subjects, the pattern points towards the fact that some classes or year groups within a school did not test, or supply the test results to the evaluation, on both subjects. This is also confirmed by the fact that some year group dummies are significant in explaining missingness in the outcome. The pattern of pupil-level missingness is not significantly correlated with TP status. As for school-level missingness, pupils without end-point assessment were removed from the analysis, and for pupils with baseline assessment we imputed missing values with multiple imputation using all pupil-level characteristics to impute missing baseline values.

Table 30: Pupil-level drop-out in non-missing schools

	Missing in maths summer			Missing in English summer		
	Coef	S.E.	p-value	Coef	S.E.	p-value
TP schools	-0.916	0.508	0.071	-0.367	0.354	0.301
Autumn_maths_score_standardised	-0.005	0.004	0.242			
Autumn_English_score_standardised				-0.013**	0.005	0.006
Female	-0.057	0.067	0.393	-0.144	0.086	0.094
FSM ever in 6 years	0.000	.	.	0.000	.	.
English as Additional Language	0.229	0.198	0.247	-0.093	0.199	0.639
Year group (Y1 as base):						
Year group = 2	0.426	0.389	0.274	0.393	0.424	0.354
Year group = 3	-0.695*	0.286	0.015	-0.882**	0.310	0.004
Year group = 4	-0.625	0.327	0.055	-0.596*	0.296	0.044
Year group = 5	-0.810**	0.301	0.007	-0.685	0.392	0.080
Year group = 6	0.058	0.419	0.890	0.461	0.447	0.302
Ethnicity (unknown as base):						
White British	-0.158	0.378	0.677	-0.437	0.345	0.205
Asian	0.036	0.436	0.934	-0.034	0.330	0.917
Black	-0.586	0.443	0.185	-0.741*	0.351	0.035
Other ethnicities	-0.339	0.390	0.385	-0.470	0.324	0.147
SEN ever	0.121	0.107	0.258	0.188	0.121	0.120
Looked after since 31 March	0.236	0.885	0.790	0.583	0.917	0.525
Looked after for 6 months	-1.203	1.422	0.397	0.405	1.142	0.723
Looked after for 12 months	1.456	1.057	0.168	-0.765	0.761	0.315
%FSM (high vs low)	-0.762	0.585	0.192	-1.328*	0.557	0.017
Pupil counts 2021	0.002	0.002	0.160	0.001	0.001	0.602
Pupil-to-teacher ratio 2018	0.179*	0.072	0.013	0.029	0.049	0.558
KS1 to KS2 valued added	-0.274	0.707	0.698	-0.533	0.510	0.296
Ofsted 2018 (Outstanding as base):						
Good	-0.476	0.921	0.605	0.547	0.494	0.269
Inadequate	-2.814*	1.340	0.036	0.968	0.690	0.161
Requires improvement	0.372	0.972	0.702	1.919**	0.691	0.006
Missing	-3.638**	1.227	0.003	-0.560	0.707	0.429
IDACI (Q1 as base):						
Q2	0.478	0.316	0.131	0.940**	0.336	0.005
Q3	-0.117	0.324	0.719	0.560	0.316	0.077
Q4	-0.140	0.308	0.650	0.152	0.340	0.655
Q5	-0.275	0.387	0.478	0.017	0.327	0.959
Missing	-1.093	0.897	0.223	1.179	1.244	0.343
Region (East Midlands as base):						
East of England	-1.279	1.159	0.270	1.280*	0.617	0.038
London	0.232	1.214	0.849	2.389**	0.926	0.010
North East	3.661**	1.199	0.002	1.808	0.946	0.056
North West	1.440	1.035	0.164	1.328	0.935	0.156
South East	1.615	1.156	0.162	2.313***	0.690	0.001
South West	-0.513	0.736	0.486	1.963**	0.723	0.007
West Midlands	1.253	1.080	0.246	1.885**	0.727	0.010
Yorkshire and the Humber	1.373	0.873	0.116	1.191*	0.592	0.044
Missing	6.289*	2.759	0.023	1.625	1.072	0.129
School types (community as base):						
Converter Academies	0.814	0.689	0.238	-0.631	0.384	0.101
Foundation School	1.755	0.989	0.076	-0.332	0.776	0.669
Free Schools	-0.970	0.962	0.313	-1.851*	0.784	0.018
Sponsored Academies	1.600*	0.783	0.041	-0.007	0.559	0.990
Voluntary Aided School	-0.108	0.907	0.905	-0.094	0.905	0.918
Voluntary Controlled School	1.101	1.051	0.295	-2.863**	1.014	0.005
Missing in pupil-to-teacher ratio	-2.181	2.154	0.311	0.656	0.562	0.243
Constant	22.434	70.505	0.750	50.740	51.004	0.320
Observations	5929			7389		
Pseudo R-squared	0.242			0.159		

Source: Primary school evaluation sample. School-level clustered residuals.

Extra analysis on the sample: Difference-in-difference analysis in TP schools

There are some limitations of the evidence presented in the results section. The proportion of PP-eligible pupils doing TP is very low, hence the PP-eligible analysis cannot pin down the impact of participation in the intervention. Moreover, the fact that we could not predict TP uptake through observable characteristics means we could not investigate the effect on pupils predicted to be in receipt of TP. Given this considerable dilution on the school-level analyses, here we propose an extra analysis, not specified in the study plan, that exploits the fact that we have repeated measures of assessments (baseline and end-point) as an alternative way to explore the effect of the programme at the pupil level. The analysis looks at the differences in TP pupils vs non-TP pupils at end-point relative to TP pupils vs non-TP pupils at baseline, using the sample of TP schools only. TP treatment is assigned at pupil level: the interaction between TP pupils in TP schools and time, equal to one for the summer assessment (end-point) and zero for the autumn assessment (baseline) should identify any change associated with TP in summer assessments.

If pupils progress as expected, then their standardised score remains the same from the baseline assessment to the end-point assessment. If they improve relative to their peers, their score increases (and vice versa). The baseline before tutoring measures a similar curriculum domain to that measured in the end-point summer assessment. Thus, the two tests can be regarded as repeated measures.

The main assumption behind this strategy is that, in the absence of TP, scores would be expected to follow a similar trend across both TP and non-TP pupils. However, in this case we do not have any other pre-treatment data additional to the baseline, so we are not able to test if the parallel trends assumption holds and there may be characteristics of pupils in receipt of TP that mean they progress at a different rate from their peers.

The results of this analysis are reported in Table 31. The coefficient of interest is the interaction between TP pupils and the summer end-point time dummy. In the sample with all pupils and with PP-eligible pupils only, for both maths and English, the interaction is not significant (shown as: Time # TP pupils). Although the interaction is not significant we note that this analysis shows that all pupils on average did better in the summer than in the autumn.

Table 31: Difference-in-difference impact of pupil-level TP in TP schools

	All pupils						PP-eligible pupils					
	Maths			English			Maths			English		
	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value	Coef	S.E.	p-value
Time # TP pupils	0.878	0.883	0.323	0.909	0.708	0.202	1.239	1.121	0.273	0.731	0.84	0.388
Time	3.071***	0.531	0.000	2.587***	0.358	0.000	2.330***	0.531	0.000	2.292***	0.42	0.000
TP pupils	-6.789**	0.655	0.000	-8.186***	0.703	0.000	-5.244***	0.886	0.000	-5.585***	0.83	0.000
Constant	24.982	12.92	0.057	11.892	12.52	0.344	18.781	10.68	0.083	15.886*	7.90	0.047
Observations	19,412	4		27,925	6		6426	6		8837	5	

Source: TP intervention sample. Time is one for summer assessment, zero for autumn assessment. School-level clustered residuals.

Estimation of effect sizes

The estimates of RQa1 (impact of availability of TP on PP-eligible pupils) were an effect size of -0.026 for maths (95% CI: -0.030 to 0.079) and -0.024 for English (95% CI: -0.100 to 0.052) as shown in Table 32. This is the equivalent of no additional months' progress compared to the comparison group. The estimates of RQ3 (impact of availability of TP on all pupils) were an effect size of 0.05 for maths (95% CI: -0.030 to 0.133) and 0.001 for English (95% CI: -0.063 to 0.065). The all-pupils estimates convert to one month's additional progress in maths, and zero additional months' progress in English. However, these effect sizes did not reach conventional levels of statistical significance and are subject to high dilution, so these estimates should be interpreted with caution. Limitations are discussed in detail in the *Conclusions* section.

Table 32: Effect size of research questions (RQs) on PP-eligible pupils (RQ1a) and all pupils (RQ3)

Outcome	Intervention group	Comparison group	Effect size		
	<i>n</i> (missing)	<i>n</i> (missing)	Total <i>n</i> (intervention: comparison)	Hedges <i>g</i> (95% CI)	<i>p</i> -value
Maths assessment (PP-eligible pupils: RQ1a)	3032 (3609)	2070 (1677)	5102 (5286)	-0.026 (-0.030 to 0.079)	0.631
English assessment (PP-eligible pupils: RQ1a)	4274 (2367)	2799 (948)	7073 (3315)	-0.024 (-0.100 to 0.052)	0.533
Maths assessment (all pupils: RQ3)	8899 (9100)	6494 (3994)	15,393 (13,094)	0.050 (-0.030 to 0.133)	0.236
English assessment (all pupils: RQ3)	13,509 (4490)	8370 (2118)	21,879 (6608)	0.001 (-0.063 to 0.065)	0.982

Conclusion

Table 33: Summary of findings

Finding
<p>On average, 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 (no evidence of an effect in English or in maths). This result has a low security rating. A particular challenge is that, on average, only approximately 20% of pupils eligible for Pupil Premium were selected for tutoring, meaning a large proportion of pupils eligible for Pupil Premium were included in the analysis who 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.</p>
<p>Similar analysis on all pupils found that pupils in schools that received TP made, on average, similar progress in English compared to all pupils in comparison schools (no evidence of an effect), and an additional one month's progress in maths compared to pupils in comparison schools. However, there is uncertainty around these estimates, with the positive maths result being consistent with a null (0 months) or slightly larger positive effect (2 months) and the English result being consistent with small positive (1 month) or small negative effect (-1 months). Furthermore, this analysis was subject to even further dilution: on average, only 12% (for maths) and 14% (for English) of pupils in the analysed schools were selected for tutoring. Given this context, it is unlikely that any of these differences were due to TP.</p>
<p>In the sample of TP schools, completing a 12-hour block of tutoring (compared to zero hours) was related to higher English scores amongst pupils eligible for Pupil Premium that received more tutoring due to the early sign-up of the school. An equivalent analysis for maths was not able to proceed.</p>
<p>A different analysis within TP schools showed that pupils who received more hours of tutoring were associated with higher English scores on average than pupils who received fewer hours of tutoring. However, this was not the case for maths, where receiving more hours of tutoring was not associated with higher maths scores. These results are associations and are not necessarily causal estimates of impact; there may be other explanations for the results.</p>

Interpretation

This study aimed to evaluate the impact of the TP programme on pupil attainment, and was designed to do so using a QED design involving several estimators of impact.

The analysis on PP-eligible pupils did not find evidence of an effect; that is, PP-eligible pupils in TP schools made similar progress to those in comparison schools, though it needs to be borne in mind that only a small proportion of pupils in the intervention schools actually received the programme. The effect sizes for English and maths are close to zero (equivalent to no additional months' progress). The analyses were subject to high dilution (i.e., they were registering the outcomes of a large proportion of pupils who were not selected for TP).

A small positive – but non-significant – effect was detected for maths in the analysis on the availability of TP to all pupils. However, with such high dilution, the effect would need to be strong amongst the small proportion of primary school pupils who were selected for tutoring (and there is no indication that this was the case elsewhere in the analysis), and/or there would need to be strong spillover effects amongst the rest of the primary school pupils (but the evaluators are not aware of a mechanism for this in TP). Another possibility is that the primary school pupils not selected for TP may have received some other form of maths support that was having an effect. However, this was not a finding in the TP IPE report, and this would need to be more so, indeed differentially so, than any such support being provided in the comparison group (and there is no reason to suspect this).

As this is a pupil-level intervention, the pupil-level dosage analyses may be more useful to schools and TPs than the school-level analyses above. It is encouraging to see that the pupil-level dosage estimates were positively correlated with English attainment in the sample of TP schools. The dosage IV analysis in RQ1c indicates that larger dosage is positively correlated with English attainment. This is a local effect estimate; that is, for those pupils that received more hours of tutoring due to the early sign-up of their school to the programme. The effect is only found in the English sample, as for maths the equivalent analysis could not proceed because earlier sign-up to the programme was not correlated with higher dosage in the maths sample. In maths, the pattern was more varied and some of the intervention schools provided high dosage in maths even if they signed up for the programme close to the end-point assessment date. Moreover, only 22% of the maths sample had data to inform the instrument, while for English 38% pupils had data available to inform this analysis. Consistent with the result for English, the moderation analysis on tutoring model (RQ6) indicates that higher pupil-level dosage is correlated with better English scores, but not maths (note that this analysis is not necessarily causal). These positive associations between the amount of tutoring received and English attainment

scores are in line with the evidence on tutoring in the EEF Toolkit, which notes more evidence is available on English (reading) tutoring at primary stage. The IV analysis in English (RQ1c), controlling for unobservable differences, provides a robust check against the obvious counterargument that these associations are just observing that more able pupils tend to attend more sessions.

The issue of dilution is important to understand here (also see the discussion on limitations, below). It has not been possible to precisely identify the counterfactual at a pupil level; that is, it was not possible, using the data available, to accurately select a group of pupils who would have participated in 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 analysis on PP-eligible pupils or the analysis on all pupils) did not in fact participate in TP, the dilution of any effect means the analyses were underpowered. In addition to the pupil-level dilution, the shift in delivery to later in the academic year was one of the reasons that not all of the pupils that received TP had completed a full block of tutoring at the time of the assessment, which is a further challenge to the analysis. On average at pupil-level, PP-eligible pupils had received only 73% of a block of 12 hours in English (approx. 8.8 hours) and 74% of a block in maths (approx. 8.9 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 the *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 during school time; disruption where whole-class bubbles had to isolate; and where tutors failed to establish a good rapport with pupils.

Initially the evaluators had some concerns about the composition of the comparison group because, in order to recruit enough schools, there needed to be a more flexible approach. This meant that during recruitment schools that could be considered similar in terms of observed school characteristics, or schools that were similar in terms of their motivation to join the TP programme (schools that signed up or expressed interest in TP but did not go on to deliver any tuition) were accepted to participate. The matching and weighting approach used indicated good balance of observable characteristics across the groups, plus some of the unobservable motivation or interest in tutoring may be accounted for in part of the comparison group. Weidmann and Miratrix's (2020) paper on selection bias further supports the idea that school-level selection bias is not likely to be a large problem in this study. The concern is centred on pupil selection.

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 pandemic-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 primary schools 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 selected to receive tutoring, and because schools selected pupils for tutoring based on characteristics that were unobservable in the available data. 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 assessment scores in English in the primary school sample.

Limitations and lessons learned

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 these strategies was very successful due to the way pupils were selected for TP. Firstly, only one fifth of PP-eligible pupils in the TP intervention sample were identified to take part in the programme. It should be noted that this low proportion is 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. Secondly, 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 and all-pupils analyses would be able to detect an effect, despite having a suitable MDES.

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 analysis of primary school data, the evaluation originally planned to select comparison schools from lists of schools matched to TP intervention schools. However, due to the challenges associated with the recruitment of all required schools from the lists of matched comparison schools, schools that signed up an MoU or an EoI but did not start the delivery of the TP intervention were approached to participate in addition to the matched comparison sample. These schools are similar to TP schools in terms of motivation to participate in the NTP. Balance checks indicate that these schools are also similar to TP intervention schools in terms of observable characteristics. 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 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 address pupil-level selection bias severely limited the conclusions that can be drawn.

To counterbalance this limitation, the evaluators proposed an additional difference-in-difference analysis on TP schools only, that looked at TP at pupil level. While the analysis has the limitations that it cannot control for pupil-level selection in TP and it cannot provide evidence in favour of the parallel trend hypothesis between TP and non-TP pupils before the intervention, it exploits the fact that the intervention is at pupil level. However, the evaluation did not detect any significant impact of the intervention using this strategy. While this additional analysis did not find any indication of impact of TP, it did at least provide some reassurance that all pupils in TP schools (whether they participated in TP or not) on average performed better in the summer tests than they did at baseline.

The analysis reported here is based on the participation and monitoring data supplied by schools and TPs. The assessment data supplied for this analysis was supplied by schools, often with the support of Assessment Providers. We, the evaluators, were able to work closely with schools to obtain this data; however, schools were not always able to supply everything (e.g., in some instances there were limitations on what test data was available). The participation data which was also used in this analysis came from the dataset for the 'population' of participating schools. This dataset was collated by TPs and held information about the tutoring itself (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 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 spring 2021 lockdown period 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. As noted above, the data available on dosage shows that, on average at school-level, pupils had completed just over two-thirds of a 12-hour block by the time of the assessment. Incomplete data (41% of rows had missing session dates) means that it is not possible to tell how accurate this figure is across the whole sample.

Measurement attrition for end-point maths and English attainment was relatively high, but it does not seem to depend on school-level characteristics, except for “TP schools” which are associated with lower missingness than non-TP schools. Similarly, pupil-level missingness in schools that submitted test results is uncorrelated with TP status, suggesting attrition should not be problematic.

Future research and publications

Given the little robust evidence for the impact of the intervention on maths and English attainment, which is limited to IV analysis on TP dosage on English attainment in the sample of TP schools, 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 TPs. 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 first year of delivery, which is evaluated here. In the second year, 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](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].
- 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 environment*. 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>.
- Zhao, Q. and Percival, D., 2017. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1). <https://doi.org/10.1515/jci-2016-0010>.

Appendix A: Security classification of trial findings

OUTCOME: *English*

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

Threats to validity	Risk rating	Comments
Threat 1: Confounding	High	Schools are well-balanced on observable characteristics, but pupil selection cannot be fully controlled for. The analysis focusing on pupils eligible for Pupil Premium is heavily diluted because only approximately 20% of these pupils were selected for tutoring, meaning that a large proportion of pupils included in the analysis did not receive tutoring. This makes it hard to detect any effect that may (or may not) be present. This is acknowledged by the evaluation team who performed and clearly reported a comprehensive analysis using all available data.
Threat 2: Concurrent Interventions	Moderate	It is unknown what other interventions were implemented in comparison schools but given the context of learning recovery during the Covid-19 pandemic, comparison schools are likely to have taken up initiatives and practices to support their pupils. The evaluation has not been able to account for these (except for participation in the Academic Mentoring programme). If comparison schools took up other support during the TP delivery period, this would make any impact of TP harder to detect.

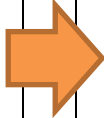

Threat 3: Experimental effects	No information available	
Threat 4: Implementation fidelity	Moderate	<p>Pupils received on average fewer hours of tutoring by the time of the end-point assessment than had been anticipated (at a pupil-level average for PP-eligible pupils, 8.8 hours in English compared to a minimum of 12 hours expected). This was in part due to delivery shifting to later in the academic year because of school closures to most pupils. The accompanying implementation and process evaluation report indicates that other aspects of the programme were implemented with moderate to high level of fidelity.</p> <p>Receiving less of the intervention than intended would likely mean that any measurable effect is smaller.</p>
Threat 5: Missing Data	Low/Moderate	<p>High attrition has already resulted in the loss of a padlock.</p> <p>Not all schools were expected to test on both subjects, but the loss associated with missing outcomes is substantial: 22% of schools in the TP intervention sample did not test in English. For the comparison sample, the sample loss is lower: 11% of schools did not test in English. The differential non-response across treatment groups points towards the likelihood of potentially unobserved differences between the two groups. However, this may also be related to how data was requested from schools, particularly in the case of intervention schools, where the main focus was on the overlap of year groups testing and tutoring in the same subject.</p> <p>Attrition does not otherwise depend on school- or pupil-level characteristics. After weighting, TP and comparison schools were well balanced on observable characteristics.</p>
Threat 6: Measurement of Outcomes	Low	<p>Several standardised assessments across multiple year groups have been amalgamated, which may increase measurement error and muddy the domain of learning that is being measured. However, the assessments used were all well-validated, standardised measurements, that are aligned to the national curriculum taught in English schools. Test type was included in the models, and distributions were normal.</p> <p>Schools administered the assessments and were not blind to treatment condition, that is, schools were aware whether pupils were receiving TP or not. However, as these are routine assessments that the school was already planning to implement, there is no reason to believe that their administration would have been biased.</p>
Threat 7: Selective reporting	Low	The analysis followed the published study plan. Where minor deviations were made, these were clearly reported.

- **Initial padlock score:** 2 padlocks – Propensity score matching and re-weighting design, low MDES, but high attrition.
- **Reason for adjustment for threats to validity:** minus 1 padlock – There is a high risk of confounding, moderate risk of concurrent interventions and moderate risk of implementation fidelity. All threats are likely to have the same direction of bias – the issue of dilution, the possibility of comparison schools taking up similar interventions in the evaluation period, and pupils receiving less tutoring than the optimal dosage all mean that it is more difficult for the evaluation to detect any impact (either a positive or a negative one). The level of attrition has already dropped the padlock rating to 2 ahead of considering threats to internal validity, however, it is important to note that schools are balanced on observable characteristics after attrition is taken into account. To reflect an overall assessment of the security of the findings,

only one further padlock is dropped here. The threats to internal validity are, however, essential to keep in mind for the interpretation of the findings.

- **Final padlock score:** initial score adjusted for threats to validity = 1 padlock

OUTCOME: *Maths*

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

Threats to validity	Risk rating	Comments
Threat 1: Confounding	High	Schools are well-balanced on observable characteristics, but pupil selection cannot be fully controlled for. The analysis focusing on pupils eligible for Pupil Premium is heavily diluted because only approximately 20% of these pupils were selected for tutoring, meaning that a large proportion of pupils included in the analysis did not receive tutoring. This makes it hard to detect any effect that may (or may not) be present. This is acknowledged by the evaluation team who performed and clearly reported a comprehensive analysis using all available data.
Threat 2: Concurrent Interventions	Moderate	It is unknown what other interventions were implemented in comparison schools but given the context of learning recovery during the Covid-19 pandemic, comparison schools are likely to have taken up initiatives and practices to support their pupils. The evaluation has not been able to account for these (except for participation in the Academic Mentoring programme). If comparison schools took up other support during the TP delivery period, this would make any impact of TP harder to detect.
Threat 3: Experimental effects	No information available	
Threat 4: Implementation fidelity	Moderate	Pupils received on average fewer hours of tutoring by the time of the end-point assessment than had been anticipated (at a pupil-level average for PP-eligible pupils, 8.9 hours in

		<p>maths compared to a minimum of 12 hours expected). This was in part due to delivery shifting to later in the academic year because of school closures to most pupils. The accompanying implementation and process evaluation report indicates that other aspects of the programme were implemented with moderate to high level of fidelity.</p> <p>Receiving less of the intervention than intended would likely mean that any measurable effect is smaller.</p>
Threat 5: Missing Data	Low/Moderate	<p>High attrition has already resulted in the loss of padlocks.</p> <p>Not all schools were expected to test on both subjects, but the loss associated with missing outcomes is substantial: 45% of schools in the TP intervention sample did not test in maths. For the comparison sample, the sample loss is lower: 20% of schools did not test in maths. The differential non-response across treatment groups points towards the likelihood of potentially unobserved differences between the two groups. However, this may also be related to how data was requested from schools, particularly in the case of intervention schools, where the main focus was on the overlap of year groups testing and tutoring in the same subject.</p> <p>Attrition does not otherwise depend on school- or pupil-level characteristics. After weighting, TP and comparison schools were well balanced on observable characteristics.</p>
Threat 6: Measurement of Outcomes	Low	<p>Several standardised assessments across multiple year groups have been amalgamated, which may increase measurement error and muddy the domain of learning that is being measured. However, the assessments used were all well-validated, standardised measurements, that are aligned to the national curriculum taught in English schools. Test type was included in the models, and distributions were normal.</p> <p>Schools administered the assessments and were not blind to treatment condition, that is, schools were aware whether pupils were receiving TP or not. However, as these are routine assessments that the school was already planning to implement, there is no reason to believe that their administration would have been biased.</p>
Threat 7: Selective reporting	Low	<p>The analysis followed the published study plan. Where minor deviations were made, these were clearly reported.</p>

- **Initial padlock score:** 1 padlock – Propensity score matching and re-weighting design, low MDES, but high attrition.
- **Reason for adjustment for threats to validity:** minus 0 padlocks – There is a high risk of confounding, moderate risk of concurrent interventions and moderate risk of implementation fidelity. All threats are likely to have the same direction of bias – the issue of dilution, the possibility of comparison schools taking up similar interventions in the evaluation period, and pupils receiving less tutoring than the optimal dosage all mean that it is more difficult for the evaluation to detect any impact (either a positive or a negative one). The level of attrition has already dropped the padlock rating to 1 ahead of considering threats to internal validity, however, it is important to note that schools are balanced on observable characteristics after attrition is taken into account. To reflect an overall assessment of the security of the findings, only one further padlock is dropped here. The threats to internal validity are, however, essential to keep in mind for the interpretation of the findings.
- **Final padlock score:** initial score adjusted for threats to validity = 1 padlock

Appendix B

Variables listed in Table 5 of the Study Plan, used for matching:

1. KS1 to KS2 value added attainment, at district level, in 2018/19.
2. Management/school type secondary – Community, Academies, Foundation, Free schools, Sponsored Academies, Voluntary school, Studio schools, University Technical college.
3. School size, total number of pupils in previous year.
4. Ofsted, overall effectiveness, in 2018/19.
5. Urban/rural area.
6. Region (London, Government Office Region, and regional dummies).
7. IDACI quintile, 2018/19 and interaction of IDACI tertiles with average attainment in the previous year.
8. Free School Meals (FSM) – percentage eligible in 2018/19.
9. English as an Additional Language (EAL) – percentage in 2018/19.
10. Special Educational Needs (SEN) – percentage in 2018/19.
11. KS2 Maths average 2017/18, KS2 GPS average 2017/18, KS2 Read average 2017/18

Variables used to compute entropy balance weights and included as controls in all regressions:

Maths sample:

- School average maths autumn scores.
- % of missing maths autumn score.
- FSM – percentage eligible in current year in 2020/21.
- English as an Additional Language (EAL) – percentage in in current year in 2020/21.
- Special Educational Needs (SEN) – percentage in in current year in 2020/21.
- IDACI, average at school level.
- Pupil-to-teacher ratio in tertiles, in 2018/19.
- Ofsted, overall effectiveness, in 2018/19.
- London dummy.
- Urban/rural area.
- Management/school type primary
- KS2 maths average in 2018/19, tertiles.
- Missing KS2 maths average in 2018/19.

English sample:

- School average English autumn scores.
- Missing English autumn score.
- FSM – percentage eligible in current year in 2020/21.
- English as an Additional Language (EAL) – percentage in in current year in 2020/21.
- Special Educational Needs (SEN) – percentage in in current year in 2020/21.
- IDACI, average at school level
- Pupil-to-teacher ratio in tertiles, in 2018/19

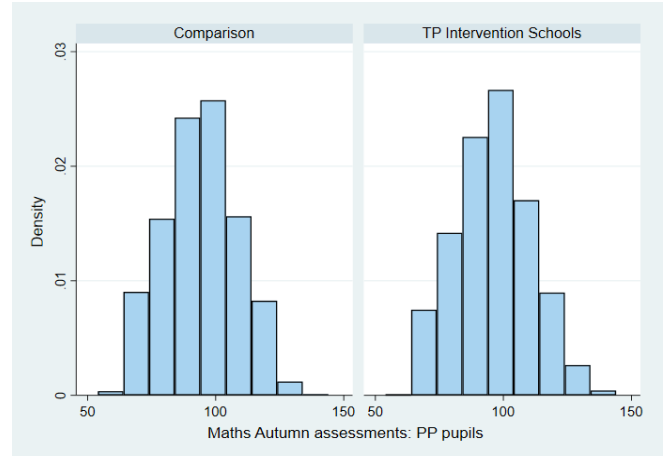
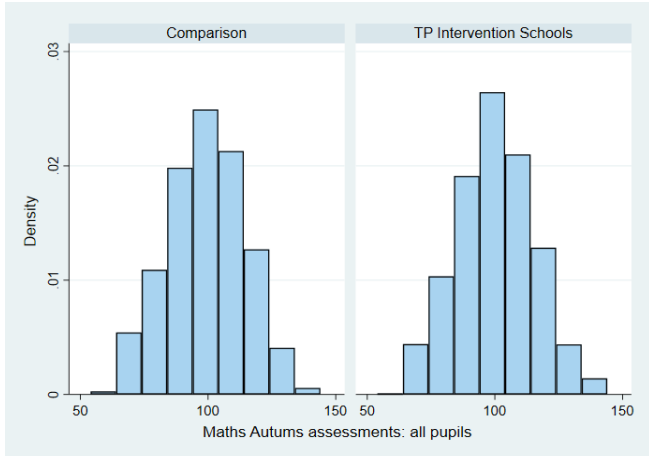
- Ofsted, overall effectiveness, in 2018/19.
- Derived regions (1. East Midlands + West Midlands 2. East of England + South East + South West 3. North East + North West + Yorkshire)
- Management/school type primary
- KS2 read average in 2018/19, tertiles
- Missing KS2 read average in 2018/19.

Pupil-level controls used in regressions:

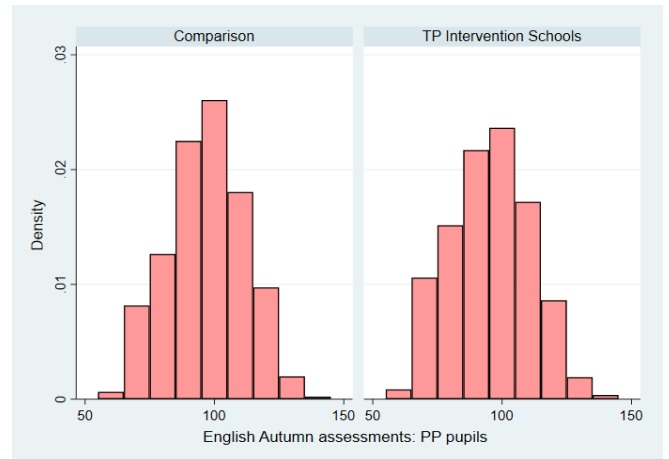
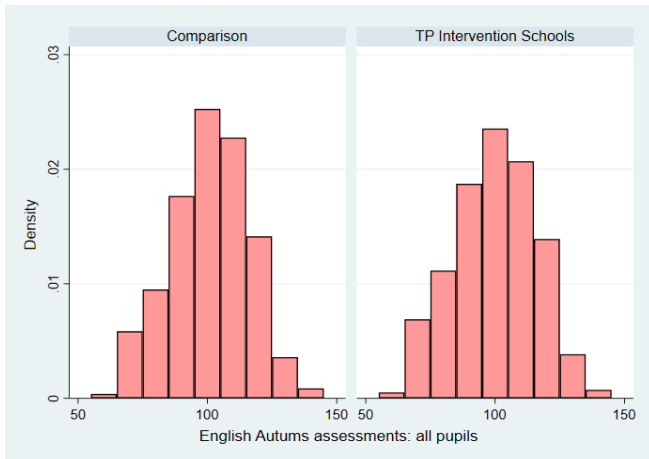
- Gender, and gender missing
- Ethnicity (White British, Asian, Black, Other and unknown)
- EAL and EAL missing
- Year group dummies and year group missing
- SEN and SEN missing
- Looked after for 12 months, since 31st March and for 6 months.
- IDACI quintiles and missing IDACI
- Assessment providers

Appendix C

Distribution of baseline assessments for maths, all pupils and PP-eligible pupils



Distribution of baseline assessments for English, all pupils and PP-eligible pupils



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

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

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



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

<https://educationendowmentfoundation.org.uk>

 @EducEndowFoundn

 Facebook.com/EducEndowFoundn