Project Title: Overarching evaluation of Tuition Partners
Evaluation Protocol and Analysis Plan

Evaluator (institution): NFER, University of Westminster and Kantar
Principal investigator(s): Pippa Lord

PROJECT TITLE	Overarching evaluation of Tuition Partners
EVALUATOR (INSTITUTION)	NFER, University of Westminster and Kantar (referred to as the Evaluator)
PRINCIPAL INVESTIGATOR(S)	Pippa Lord
STUDY PLAN AUTHOR(S)	Pippa Lord, Ben Styles, Helen Poet, Veruska Oppedisano, Richard Dorsett, Ben Collins, Sarah Tang
STUDY DESIGN	Programme evaluation involving a quasi-experimental design (QED) and process evaluation
PUPIL AGE RANGE AND KEY STAGE	5-16 years (KS1, KS2, KS3 and KS4)
NUMBER OF SCHOOLS	Original targets for schools:191 primary English, 191 primary maths, 151 secondary English, 151 secondary maths (known as the evaluation sample). (In practice there may be some overlap between samples where a school is receiving tutoring in both subjects). These targets have been revised (June 2021) due to slow recruitment of control schools and grant an MDES of 0.19: 132 primary English, 121 primary maths, 105 secondary English and 104 secondary maths.
NUMBER OF PUPILS	Original targets for Pupil premium pupils (in the evaluation sample): • 3629 primary English, 3629 primary maths • 10570 secondary English, 10570 secondary maths Pupil premium pupils revised targets: • 3168 primary English; 2904 primary maths • 6720 secondary English; 6656 secondary maths
PRIMARY OUTCOME MEASURE AND SOURCE	Attainment in English and maths. Within primary schools this will be based on standardised assessments (Years 1-6). Within secondary schools this will be based on standardised assessments (Year 7-10) and on the teacher assessed grades that are replacing GCSEs in 2021 if available on NPD (Year 11). *[note that the analysis on the secondary school evaluation sample/Research Champion sample is no longer*

	possible due to achieved numbers – but the Year 11 population analysis will proceed subject to checks]
SECONDARY OUTCOME MEASURE AND SOURCE	Not applicable

Study Plan version history

VERSION	DATE	REASON FOR REVISION
2.0	23/11/2021	*Updated in response to the implications of the national lockdown involving partial school closures (4th Jan 21 – 5th March 21) on TP delivery and evaluation design. This includes the cancellation of the Year 6 SATs; the use of teacher assessed grades replacing GCSE assessments for Year 11s .* *The study plan has also be updated to reflect the achieved number of recruited schools to the impact evaluation.* *Changes since the first study plan are shown in blue font.*
1.0 [*original*]*	12/02/2021	*[leave blank for the original version]*

Table of contents

National Tutoring Programme: Tuition Partners – summary

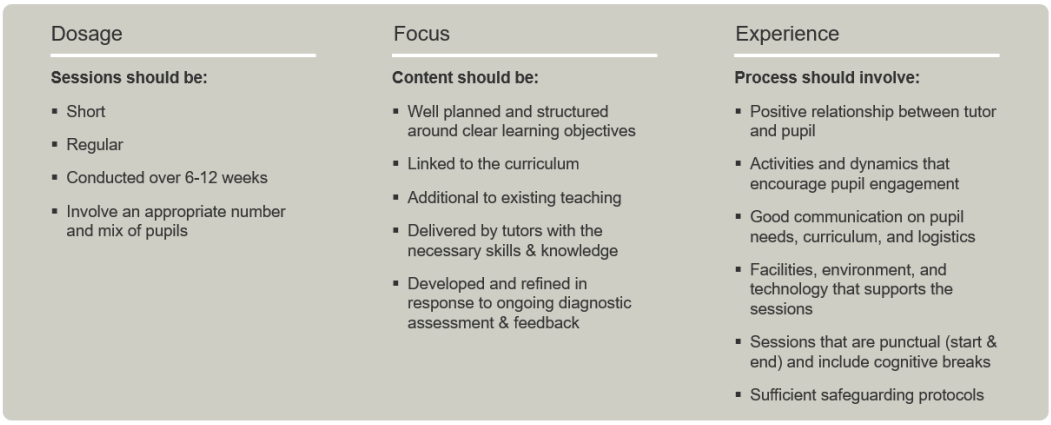
The National Tutoring Programme: Tuition Partners is designed to provide additional support to schools and teachers, to supplement classroom teaching, through subsidised, high quality tutoring for schools from an approved list of Tuition Partners.

The programme aims to support teachers and schools in providing a sustained response to the Covid-19 pandemic and to provider a longer-term contribution to closing the attainment gap. The programme is part of a wider UK-Government response to the pandemic.

The TIDieR framework below summarises the programme and intervention. Continue to page 7 for the full study plan content.

Table 1: TIDieR framework – programme summary

Aspect	Description
Programme	National Tutoring Programme: Tuition Partners (referred to in the evaluation and this study plan as the 'TP programme').
Why (rationale)	Research shows that pupils' learning has been affected by school closures due to Covid-19 (Coe *et al.,* 2020; Cullinane and Montacute, 2020). The programme is designed to provide additional support to schools to help disadvantaged pupils whose education has been most affected by school closures due to Covid-19. There is a large body of evidence that tutoring and small-group tuition is effective – particularly where it is targeted at pupils' specific needs; and that it can be particularly effective for disadvantaged pupils (Nickow, Oreopoulos and Quan, 2020; EEF, 2018a, 2018b; Torgerson *et al*., 2018; Dietrichson *et al*, 2017).
Who (recipients)	Disadvantaged pupils, including those eligible for Pupil Premium/Free School Meal funding. Participating schools are able to identify which of their pupils they feel will most benefit from additional tuition support. The programme aims to reach 235,000 pupils, across 6,000 state-maintained schools in England.
What (materials)	Tuition is provided to schools at a 75% subsidy. Tuition Partners receive 75% from the NTP TP programme funding. Tuition Partners receive guidance and support around: Best tutoring practice guidance for tutoring organisations: https://nationaltutoring.org.uk/resources/best-tutoring-practice-for-tutoring-organisations Best tutoring practice for schools: https://nationaltutoring.org.uk/resources/best-tutoring-practice-for-schools

Aspect	Description
What (procedures)	The TP programme will enable participating schools to access high quality tutoring from an approved list of tutoring providers, who have passed a set of quality, safeguarding and evaluation standards specified by the Education Endowment Foundation (EEF) (eight criteria were used, and these are outlined in the section on the programme theory and design). There are five broad phases of the programme: design, develop, mobilise, deliver and legacy (see Figures 4 – 7). As part of the development phase (and ongoing throughout the year), TPs receive support from capacity building partners Nesta and Impetus. TPs are required to provide training for all their tutors. Tutoring is expected to be high quality to unlock the intended impacts on attainment. High Quality Tutoring is expected to entail: 
Who (provider)	The programme is being overseen by Programme Managers (EEF, Impetus and Nesta). The tutoring will be delivered by 33 Tuition Partners (TPs)[1], appointed by the NTP. Participating schools can access high quality tutoring from this approved list of Tuition Partners. Tuition Partners recruit and train tutors to deliver the tuition in schools. The tuition will be delivered by up to 20,000 tutors.

[1] 32 organisations were announced as Tuition Partners on 02/11/2020. The 33rd TP was announced in mid-November.

Aspect	Description
How (format)	A range of tutoring models will be funded, including those that are suitable for pupils with special educational needs and disabilities (SEND) and in alternative provision. Tuition is delivered either online or face-to-face. A majority of tuition will be delivered by a single tutor to three pupils (1:3 ratio), but a smaller amount is expected to be delivered on a 1:1 basis for pupils with SEND and other additional needs or on a 1:2 basis. Tuition is available in the following subjects: English/literacy, maths, science, humanities and modern foreign languages. Schools are entitled to one 15 hour block of tutoring in one subject per pupil. This evaluation is concerned with English and maths. It is expected that most pupils will receive tuition in English or maths, so the evaluation will assess outcomes in these subjects to maximise power, minimise testing burden and reduce the complexities of multiple testing outcomes.
Where (location)	State-maintained primary, secondary and special schools in England. Pupils in alternative provision can also access tutoring. Tuition is expected to be delivered in schools (before, during and after school), in additional to usual teaching. In certain circumstances, tuition can be delivered at home (for example, for pupils' who are shielding/medically vulnerable and are accessing their school learning from home). As a result of the national lockdown in January-March 2021, when schools only remained open to children of keyworkers and vulnerable pupils, the NTP made provision for online tutoring to be available for pupils at home during that time. This led to a greater proportion of online tutoring. The EEF assessed Tuition Partners' safeguarding arrangements for online at home tuition.
When and how much (dosage)	The TP programme takes place during the academic year 2020–2021: ▪ Initial programme activities to establish the implementation systems and engage participants will run from July 2020 onwards. ▪ Tutoring expected to commence from November 2020. ▪ Programme completion (year one) in July 2021. Schools can access 15 hours of tutoring per selected pupil during the 2020/21 academic year (min. 12 hours is considered as a completed block of tuition). Note – each pupil selected for tuition will be able to access one 15 hour block of tuition. Another impact of the lockdown in the spring 2021 term was that delivery shifted to later in the academic year. The NTP agreed that some tuition could continue into the summer holidays, but the expectation was that tutoring blocks are completed by the end of the summer term where possible, and if not, then tuition blocks should end by the end of the summer holidays (end of August 2021). Similarly, schools starting tuition from mid-May could opt to book 10 sessions per selected pupil, and for this to be considered a completed block.

Aspect	Description
Tailoring (adaptation)	Tuition Partners will be supported by 'capacity building partners' (Impetus and Nesta), to develop their organisation's capacity during the year of the programme. This may include developing their delivery/tuition models. The study will explore if and how TPs adapt their models during the year of the programme. The study will also explore how pupil learning outcomes vary by different types of tutoring, as well as by pupil, tutor and school characteristics.

About the National Tutoring Programme: Tuition Partners

This section of the study plan sets out the current context in schools, the government-funding response, and the rationale for the TP programme including evidence for one-to-one and small group tuition. It then outlines the range of stakeholders involved in the programme (providers and participants), a logic model depiction for the whole programme, and a theory of change focused on tutoring ('the intervention'). It highlights the key mechanisms and features of the programme and of tuition that might affect or moderate pupils' learning outcomes.

Context in schools: Covid-19 and the impact on pupils' learning

In response to the Covid-19 pandemic, the UK Government asked all schools in England to close in March 2020. Re-opening for some year groups was possible during June and July (where possible for Reception and Year 1, and for some Year 11s and Year 10s, as well as continued provision for children of keyworkers). All schools were asked to re-open from September 2020, and although schools now also have a duty to provide remote learning where needed, provision is still affected by the challenges of Covid-19 (for example, local closures, year groups or bubbles unable to be in school, teachers' teaching online at the same time as in class).

Research highlights that pupils are behind in their learning. In their review of the impact of school closures on attainment, the Education Endowment Foundation (EEF), reported projections that school closures will widen the attainment gap between disadvantaged children and their peers (with a median estimate of widening by 36%), likely reversing progress made to narrow the gap since 2011 (EEF, Coe *et al*., 2020). According to the report of a national survey (weighted sample) of school leaders and teachers published in September 2020, teachers estimated that their pupils were behind in their learning with the average reported estimate being three months behind (Sharp *et al*., 2020). This issue was more acutely reported in the most deprived schools, and over half of teachers estimated that the learning gap between disadvantaged pupils and their peers had widened. Reasons relate to schools' varied provision of remote learning – a particular challenge for the most deprived schools, and for pupils from disadvantaged backgrounds, particularly around access to IT, having suitable spaces to learn in, and access to other learning resources (Cullinane and Montacute, 2020; UCL, 2020; Sharp *et al*., 2020).

Evidence for one-to-one and small group tuition

In their review of the evidence on Covid-19 disruptions and the impact on attainment, the EEF suggested two key ways to support learning in these unprecedented times: i) to support effective remote learning to mitigate the extent to which the gap widens; and ii) sustained support to help disadvantaged pupils catch up. They particularly 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 tutoring and small-group tuition is effective – particularly where it is targeted at pupils' specific needs. The EEF toolkit pages on one-to-one tuition (EEF, 2018a) and on small group tuition (EEF, 2018b) show that both are effective interventions, and that training and support are important in the effectiveness of the tuition. Effect sizes vary across studies – with between three and six months additional progress being reported in studies of various one-to-one interventions; and in small group tuition the key finding across studies is that the smaller the group and the more aligned to pupils' needs, the more effective the intervention.

Meta-analyses have shown that tutoring programmes yield consistent and substantial positive impacts on learning outcomes: the EEF Teaching and Learning Toolkit meta-analysis

estimates the average effect size tutoring to be 0.3 SD for small group tuition and 0.37 SD for one-to-one tuition; Nickow, Oreopoulos and Quan (2020) found an overall pooled effect size estimate of 0.37 SD; Dietrichson *et al.* (2017) found a pooled effect size of 0.36 SD; and Ritter *et al.* (2009) found a pooled effect size of 0.30 SD. There is also evidence that tutoring can be particularly effective for disadvantaged pupils (Torgerson et al., 2018 and Dietrichson et al, 2017). However, it is expected that the effect sizes for pupils receiving TP may be smaller. Many of these studies will have been targeted at low-attainers and may be based on samples with a narrow range. The target group of TP (disadvantaged pupils) is expected to be more heterogeneous in terms of attainment.

Researchers also highlight that contributions to research on 'recovery' or 'catch up' should take into account lockdowns and absences throughout this academic year, patterns of recovery (Kuhfeld et al., 2020) and assumptions about different support strategies including online learning (Moss, 2020). Hence, any evaluation of tuition or tutoring should take into account wider context and 'moderators' that might affect the implementation of that tutoring. Moreover, research should also take into account that the counterfactual may be a very different 'business as usual' in the current climate – it is likely that pupils who do not receive TP will be provided with other forms of support by schools, and these may involve one-to-one or small group support (see funding response below).

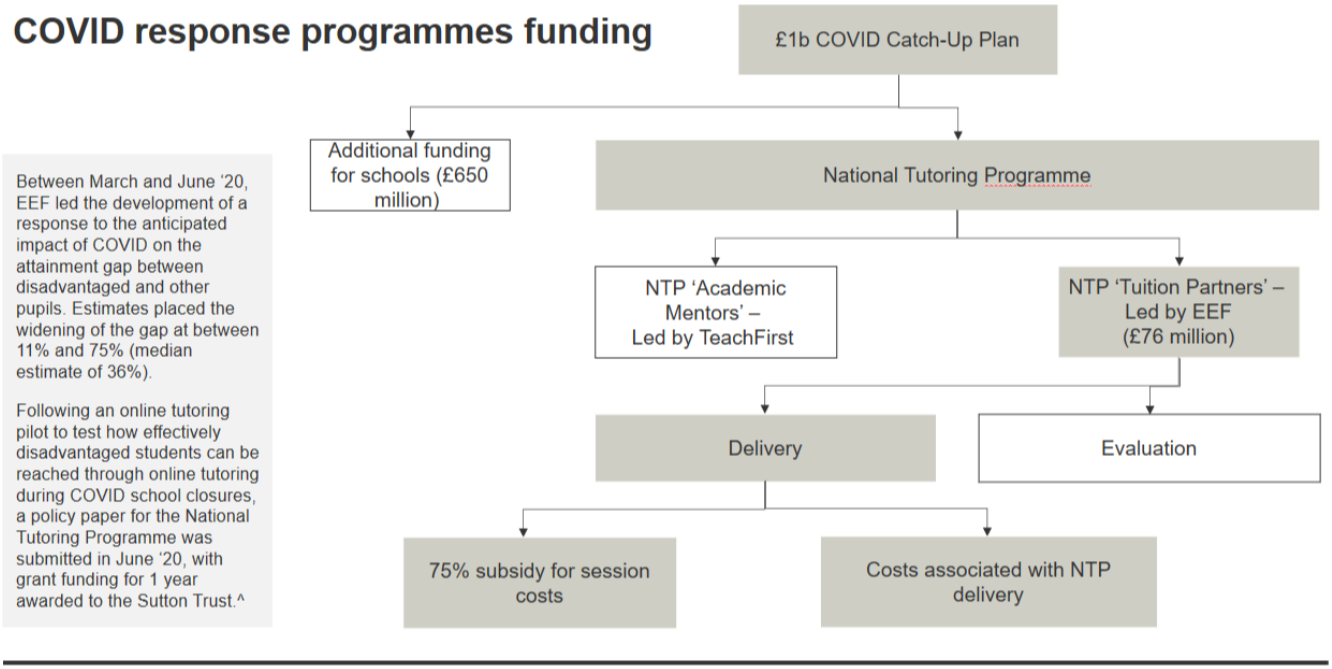
Current policy context: government funding to support catch up

The UK government has launched a one-off universal £650 million catch-up premium for the 2020 to 2021 academic year, to help schools to provide catch up activities to help pupils make up for lost teaching time. (This is part of a £1 billion Covid-19 catch up package to support schools.) The government also launched a £350 million National Tutoring Programme to provide additional, targeted support for those children and young people who need the most help (for example, the disadvantaged and vulnerable groups that will have been affected most). The National Tutoring Programme (NTP) aims to support schools in providing a sustained response to the coronavirus pandemic and to provide a longer-term contribution to closing the attainment gap. It will do this by increasing the availability of high quality tuition available to disadvantaged schools and students. There are two components of the NTP:

1. Academic Mentors (AM)[2] which places trained graduates in schools in the most disadvantaged areas to provide intensive support to their pupils.
2. Tuition Partners (TP) programme which offers schools access to subsidised, high quality tuition from an approved list of providers.

[2] Note, members of the evaluation team involved in evaluating Tuition Partners, also have a strand of work to evaluate AM. A separate study plan will be published about that.

Figure 1: Covid response funding streams



Who is involved in the TP programme?: providers and participants

The following five stakeholder-levels are involved in the TP programme:

- Programme Managers at EEF, Impetus and Nesta are leading the design and development of the TP programme in the academic year 2020–2021[3]. Programme Managers at EEF are responsible for selecting Tuition Partners based on the quality of, and evidence behind, their model, coordinating safeguarding[4] and due diligence checks, managing the relationship and grant agreement with each Tuition Partner, and collecting data aggregate monitoring data from Tuition Partners in order to report to the DfE on reach. Impetus and Nesta are supporting Tuition Partners to build their capacity for delivery and impact, through workshops, 1:1 support and peer learning. The focus of support is on best practice for delivering impactful tuition at scale, and troubleshooting growth.
- Tuition Partners are organisations that have passed a set of standards to deliver tutoring for the TP programme. There is an approved list of 33 Tuition Partners, selected based on the quality of their model and value for money, evidence, and potential to scale to support large numbers of schools. Tuition Partners might be existing tutoring providers that have experience of working with schools, or other organisations, such as charities, local authorities or universities who are able to design a new programme to meet the NTP standards. Tuition Partners are responsible for recruiting and training tutors for the TP programme, managing the tutors, communication with schools, scheduling tutoring sessions, and for providing

[3] Note, the design and delivery of the whole of the National Tutoring Programme in its first year will be led by a collaboration of five charities - the Education Endowment Foundation, Sutton Trust, Impetus, Nesta and Teach First - supported by the KPMG Foundation.
[4] EEF developed the safeguarding criteria and oversaw the process but commissioned external experts to conduct the necessary checks.

monitoring data to Programme Managers (and for the evaluation – see later in the study plan).

- Tutors are responsible for delivering tuition to participating pupils in schools. They will deliver up to 15 hours tuition per pupil. Around 20,000 are being recruited to the TP programme over the course of the year. Tutors can work with one or more Tuition Partner, and are expected to undertake any TP training provided.
- Around 6,000 state-maintained primary, secondary and special schools can access tuition through the TP programme. Schools can choose which Tuition Partner(s) they wish to work with, and are responsible for identifying pupils for tuition – in which year groups and which subjects.
- The programme is targeted at disadvantaged pupils, including those eligible for Pupil Premium. Participating schools are able to identify which of their pupils they feel will most benefit from additional tuition support, as outlined in the guidance to TPs[5]: 'The focus of the NTP is on supporting disadvantaged pupils aged 5-16. Schools should therefore be asked to focus on disadvantaged pupils, including pupils eligible for Pupil Premium funding, Free School Meals or those identified by schools as having an equivalent need for support. Participating schools will be able to decide which of their pupils will most benefit from additional support.' Pupils selected for tuition will take part in up to 15 hours tuition in one subject through the TP programme. Pupils can be in Year 1 – Year 11. In initial targets, the programme was expected to reach around 235,000 pupils in the academic year 2020-21. (Note that as the programme progressed, targets were revised (upwards) – such revisions will be reported in the implementation section of the evaluation report.)

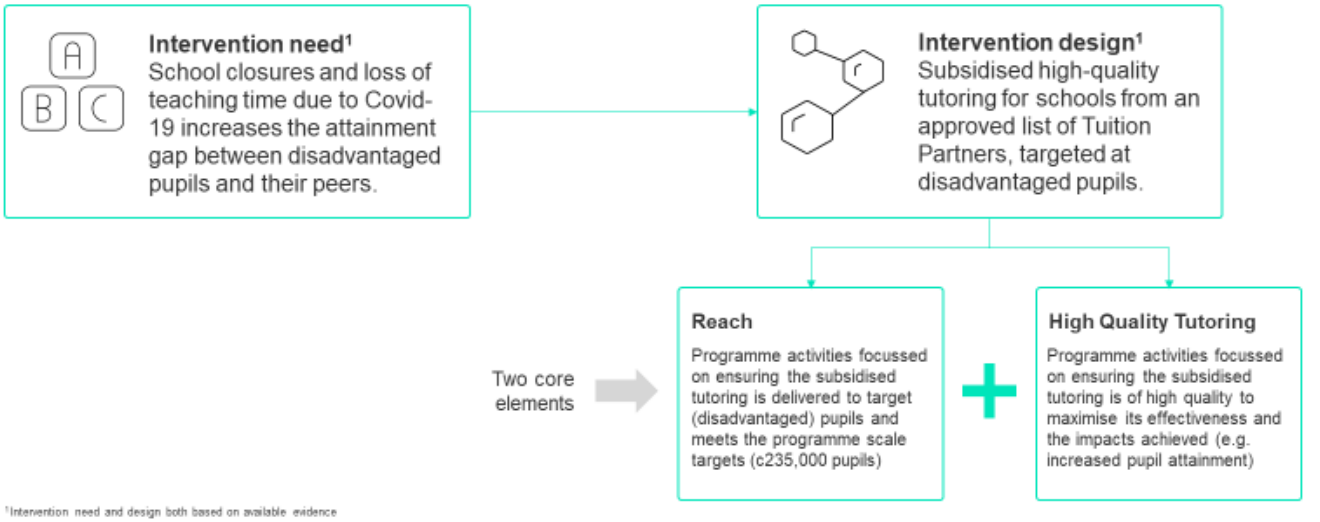
What does the programme entail?: programme theory and design[6]

Tuition Partners is designed to address the effects of the loss of teaching time from school closures due to COVID-19. Through subsidised high quality tutoring, disadvantaged pupils will receive additional teaching to help raise their attainment and reduce the gap to their peers.

There are two core elements of the underpinning programme theory: Reach and High Quality Tutoring. The programme has been designed to ensure the subsidised tutoring reaches the desired type and number of pupils and that the teaching they receive is of high quality to maximise the impact on the attainment gap to their peers.

[5] NTP Guidance on signing up schools to Tuition Partners [accessed 25/06/2021]
[6] An accompanying Logic Model document provides a detailed depiction of the programme design and intentions, the delivery activities, and phase by phase logic models. Key elements have been extracted here.

Figure 2: Programme theory – intervention need and design



High quality tutoring consists of a number of components, across three conceptual areas: dosage, focus, experience – and these informed the implementation and process evaluation.

- **Dosage** – sessions should be:
    - Short
    - Regular
    - Conducted over 6-12 weeks
    - Involve an appropriate number and mix of pupils
- **Focus** – content should be:
    - Well planned and structured around clear learning objectives
    - Linked to the curriculum
    - Additional to existing teaching
    - Delivered by tutors with the necessary skills & knowledge[7]
    - Developed and refined in response to ongoing diagnostic assessment & feedback[8]
- **Experience** – process should involve:
    - Positive relationship between tutor and pupil
    - Activities and dynamics that encourage pupil engagement
    - Good communication on pupil needs, curriculum, and logistics
    - Facilities, environment and technology that supports the sessions

[7] Includes: (i) theoretical knowledge, (ii) applied knowledge (including relevance from beyond teaching/tutoring), and (iii) pedagogical knowledge
[8] Including reflection on previous sessions and in response to external monitoring

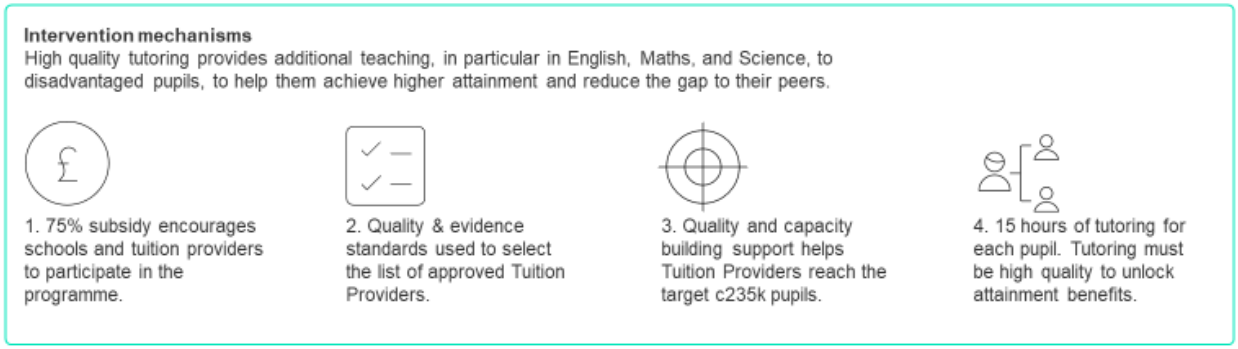
- Sessions that are punctual (start & end) and include cognitive breaks
- Sufficient safeguarding protocols

Ultimately, the TP programme seeks to reduce the attainment gap between disadvantaged pupils and their peers. But the programme is also designed to shape the tuition market and deliver evidence of what works. The specific intended programme impacts are:

- Improved attainment for disadvantaged pupils
- Increased amount of and better targeted support for disadvantaged pupils
- Increased capacity and quality in the tutoring sector
- TPs retain partnerships with schools and tutors in 2021/22 and beyond
- Greater teaching capacity in schools – tutors are retained and some train as teachers
- Generate an evidence base of the effectiveness of tutoring

The programme involves four broad mechanisms to achieve these intended impacts, as outlined in Figure 3 below.

Figure 3: Programme theory – intervention mechanisms

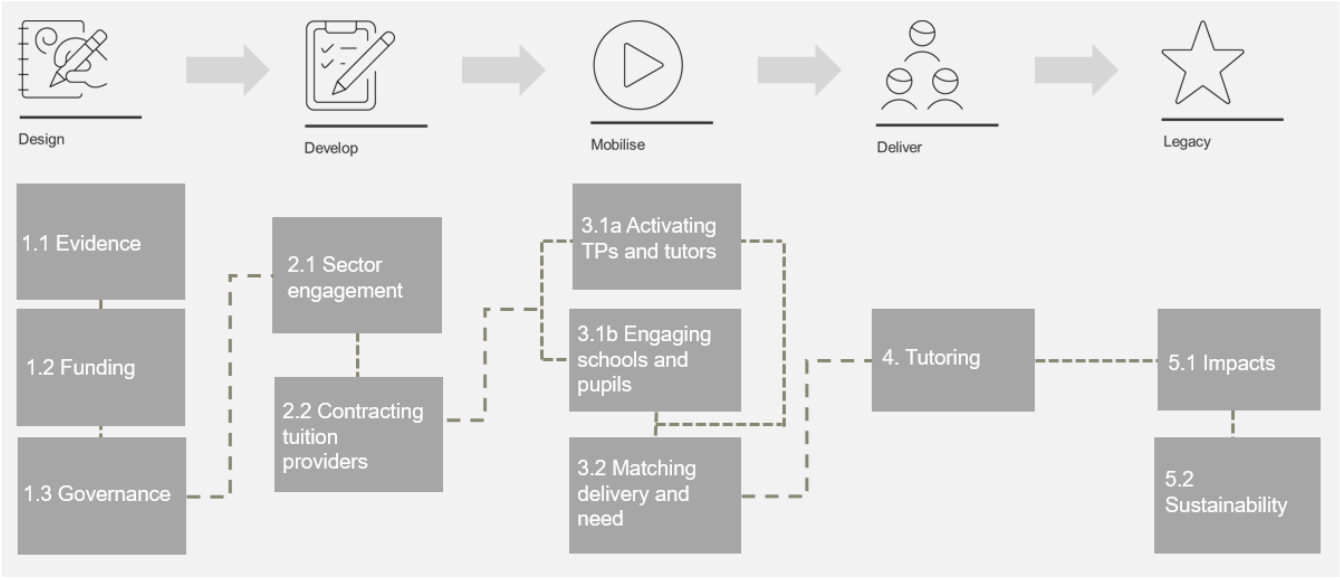


There are five broad phases of programme implementation:

- **Design:** Identify the need for an intervention to address the impact of COVID-19 on attainment amongst disadvantaged pupils; and based on current evidence, design a programme to obtain funding and participation from relevant organisations and individuals.
- **Develop:** Establish the necessary infrastructure for programme delivery, including sector engagement to map tutoring supply and school demand, invite interest from schools and applications from tutoring providers, including a three-part assessment of the suitability, quality and capacity of providers.
- **Mobilise:** Further activities to put in place the necessary resources, processes, guidance, training, standards, and reporting protocols, to ensure sufficient delivery capability, quality and scale amongst tutoring providers, participation of schools, and suitable matching of provision and supply.
- **Deliver:** Delivery of tutoring for each selected disadvantaged pupil (either in person or online, and in one to one or small group settings), facilitated by actions across tutoring providers, schools and teachers, and programme managers, plus pupils and parents.
- **Legacy:** Programme impacts on pupil attainment and non-cognitive effects; positive effects on the scale and quality of the tutoring sector; establishing connections between tutoring providers and state schools; and contributing to the evidence base on the effectiveness of tutoring. Long term sustainability of systems and effects.

Within these phases, there are a series of sub-phases and underlying activities that build to achieve the intended impacts, as shown in Figure 4. Figures 4a – c provide further detail of each phase – and can be found in Appendix C.

Figure 4: programme process diagram



Note: These are five discrete phases but due to the rolling nature of the programme and implementation, some phases are concurrent. For example, delivery may begin in some schools whilst mobilisation activities are continuing for some tuition partners.

As set out in the Implementation and Process Evaluation section, the study will have a focus on exploring quality – both at a system level, and at the level of the tutoring itself (as set out above in terms of dosage, focus and experience). At the system level, it is worth noting that the design and development phases of the programme involved establishing quality criteria, informed by the conceptual areas of high quality tutoring as well as other aspects of quality which programme managers used to assess organisations' provision, in order to select Tuition Partners for the programme. The eight criteria used to assess organisations covered the following areas:

- experience of working with schools
- tutor recruitment and qualifications
- tutor training
- systems and processes for school and pupil communication
- experience of working with disadvantaged pupils
- monitoring delivery
- quality assurance and tracking progress
- evidence of impact.

Programme and school context: spring and summer 2021

This new section of the study plan briefly documents some programmatic changes made in response to the continuing Covid-19 pandemic in 2021. Further details will be included in the evaluation report.

In early January 2021 (04/01/21), part-way through the tutoring programme, another national lockdown was announced, and schools were told to close to pupils other than those whose parent(s) were keyworkers, or who were identified as vulnerable pupils. All other pupils

returned to home-schooling/remote learning. This had a significant effect on the delivery of tutoring, which had been planned to expand in earnest in January 2021. A number of key changes were made to delivery from this point by the NTP in order to respond to the changing circumstances. These included:

- delivering online tutoring at home in specific circumstances during the school closures (early January 2021); this included some TPs who previously only offered face-to-face tutoring introducing an online offer – this change in provision had to be approved by EEF for each TP including providing appropriate safeguarding arrangements
- allowing greater flexibility in online/face-to-face delivery targets (mid-January 2021)
- expanding online at home into weekend provision (end of January 2021)
- extending the NTP TP into the summer holidays (early March 2021) (noting that most schools chose to delay delivery until they re-opened rather than move to online tutoring)
- shorter blocks of 10 hours of tuition accepted in specific circumstances (mid May 2021) (noting that schools that had not yet started tuition could therefore book blocks to fit in sessions before the end of term).

Reach targets were also amended as the programme progressed – both upwards and allowing greater flexibilities for TPs in key stage and regional targets – and these will be reported in the evaluation report. These numbers in themselves did not affect the evaluation design. In addition, it is worth noting that disruption in schools continued throughout the year as a result of the pandemic. Pupil absences due to self-isolation affected individuals as well as whole classes/year groups missing school. This has an impact on programme delivery. The disruption had an impact on evaluation implementation particularly in terms of schools' using and providing assessment data, and their capacity to be able to engage in an evaluation.

Policy context: spring 2021

In addition to announcing that schools would close to many pupils, the government also announced that the summer exams – including GCSEs and Year 6 statutory assessments - would be cancelled. On 25th February 2021, it was confirmed that GCSEs would be awarded based on teacher assessed grades.

This has implications for the evaluation design. Indeed, at the time of writing, the reliability and validity of teacher-assessed grades remains unknown. More details can be found below in the impact evaluation section.

It also had implications for programme delivery in terms of how schools selected pupils for tutoring; for example, fewer Year 6 and Year 11 pupils were identified for tutoring. Programme reach will be explored in the evaluation report. These numbers in themselves did not affect the evaluation design.

About the evaluation

Evaluation rationale and aims

Given that national policy is aiming to support pupils' learning recovery in these unprecedented times, an evaluation of that support is important not only in terms of whether it is effective in supporting pupils' learning this year, but also in terms of how the programme as a whole is working, and any improvements needed to inform future tuition programme(s).

The evaluation aims to quantify the overall impact of the TP programme on pupil attainment/learning outcomes, and will look at how this varies by different types of tutoring, pupil, and school characteristics. The evaluation will also look into the experiences of schools, tutors and pupils in order to improve the delivery of similar programmes in the future.

The evidence generated through this evaluation will be used to help Tuition Partners and tutors design and deliver future high quality tuition that makes a difference to pupils' attainment. It will also help schools in further academic years to better target and engage the pupils who would benefit from tutoring. 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 will need to interpreted in light of context. The evidence will be specific to implementation during 2020/21 and when interpreting the results we will be mindful that they may not be fully generalisable to future years of the programme or to tutoring more widely.

Evaluation design overview

The evaluation is designed around three workstreams:

1. Data management – including collation and analysis of monitoring data including number of schools and pupils taking part in tutoring, characteristics of those schools and pupils (such as Pupil Premium), and attendance at tutoring sessions.

2. Implementation and process evaluation (IPE) – including both programme-level evaluation and evaluation at the level of the tutoring itself; using a logic model as the framework for design and analysis; incorporating views via surveys and interviews from the range of stakeholders involved (NTP TP Programme Managers, Tuition Partners, school leads, classroom teachers, tutors and pupils); collecting information on costs; and taking a formative approach so that learning is fed back into the programme during the course of the year.

3. Impact evaluation – assessing the impact of tutoring on pupils' maths and literacy attainment outcomes, using a quasi-experimental design (QED) involving a comparison group.

The IPE and impact analysis will also include an exploration (descriptive analysis) of the tutoring models and moderators, and whether these are associated with outcomes (such as mode of delivery, tutors' experience/qualifications).

When designing the evaluation, a number of issues were considered including: defining the research questions that could be answered; the appropriateness of a QED and how best to operationalise a comparison group design; scale and scope (note, the evaluation involves both

population data for reporting on the monitoring data provided by TPs about all participants[9], and evaluation sample data for analysing and reporting on impact[10]); burden on schools and incentives (see Appendix A); use of NPD data and other assessment data; and the timescales of the programme and how best to provide formative feedback throughout the evaluation. These issues are discussed further where relevant in the study plan.

(Note, although our three workstreams deliberately put the IPE evaluation to the foreground, acknowledging that this a whole programme evaluation, this study plan leads on the impact design in order to follow the conceptual outline of the study plan template and ensure due focus and pre-specification of the QED study design.)

Evaluation research questions – overview

The overall research questions for the evaluation are as follows (note further sub-research questions are detailed in the Impact section and Implementation and Process Evaluation section of this study plan):

Impact RQ: What is the impact of TP on learning outcomes for pupils? (this will be investigated through a number of estimators of impact, in both English and maths, in both primary and secondary schools).

Moderator RQs: Does the impact vary by school and pupil characteristics, and by different models of tutoring (e.g. face-to-face vs online; 1-1 vs small group)? (these will be investigated through comparison models where characteristics can be measured in both intervention and comparison groups; and in exploratory descriptive analysis where characteristics are observable in the intervention group).

Implementation RQ: How (well) was TP delivered and what are the implications for the programme theory, design, and effects? (this will be investigated through a number of qualitative and quantitative research activities with programme participants and beneficiaries across the five phases of the programme).

[9] I.e. approx. 6,000 schools and 230,000 pupils.

[10] I.e. from a sample of schools that provide assessment data for the evaluation, involving 191 primary schools in English, 191 primary schools in maths, 151 secondary schools in English, 151 secondary schools in maths; and 3629 primary pupils in English, 3629 primary pupils in maths, 10570 secondary pupils in English, and 10570 secondary pupils in maths (these numbers include TP and comparison group schools and pupils).

Impact evaluation

Overview

The primary objective of the impact evaluation is to determine what difference, if any, is made by TP to attainment outcomes (maths and literacy). The impact evaluation uses observational methods (supplemented by other approaches where possible) rather than a randomised controlled trial (RCT), due to the need to maximise reach to as many schools and pupils as possible, as soon as possible[11].

We will include analysis of attainment outcomes controlling for i) pupil characteristics, including gender, ethnicity, English as an additional language (EAL), year group, prior attainment, pupil premium (eligible vs. not), SEND vs. not, school attendance; ii) school characteristics such as education stage (primary vs. secondary), Ofsted rating (using the four-point scale), proportion of FSM (high vs. low), and iii) geography (London, GRO and regions).

We will examine whether estimated impacts vary according to school characteristics (primary vs. secondary; high vs. low Ofsted rating; proportion of FSM; pupil characteristics (prior attainment; whether SEND; key stage or age; attendance; English as additional language; ethnicity; gender) and geography.

We will also run analyses to compare outcomes associated with different tutoring models among TP schools in the evaluation sample. The differences considered will include mode of delivery (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 (low/high buy-in schools); intensity of delivery (determined by sessions attended/number of weeks tutoring is spread over). We will also examine variation in outcomes by tutor characteristics (experience/qualifications; TP tutor training; shared characteristics with pupil/tutee (gender, ethnicity)).

Summary of changes to the research questions and method (October 2021)

As outlined in the outcomes section below, originally we intended to conduct population analysis at both primary and secondary. Population analysis refers to the population of [Year 11] pupils participating in TP in schools in England compared with a comparison group selected from those not participating. However with the **cancellation of the Key Stage 2 tests in summer 2021, we are no longer able to conduct population analysis on the primary schools**. We still intend to conduct the population analysis on the secondary schools using the teacher assessed GCSE grades awarded in 2021 as we understand that this will be available on the National Pupil Database (NPD)[12]. We are concerned that the process for grade determination may mean that it is difficult to detect any potential impact of the TP

[11] There was an opportunity to include a small RCT within the evaluation, if a TP or TPs did not have capacity to satisfy demand in the spring (2021) term and planned two consecutive waves of (compressed) delivery. In this scenario the TP could have incorporated pupil-randomised waitlist designs. The outcome measure would have been attainment, using standardised tests as outlined in the QED section above. In contrast to the overarching evaluation, which will look at impact on attainment across the programme, the RCT would have investigated the impact of single TPs on attainment. TPs were made aware of the possibility of randomisation during the November webinar about the evaluation. The changes to delivery due to the January lockdown meant this option was no longer feasible.

12 We understand that while the GCSE grades will be available on NPD, it will not be possible to link the grades to school identifiers (e.g. name, post code, unique reference number (URN), DfE number), which will limit the amount of analysis we are able to conduct in relation to school characteristics. We have listed our intended analysis by school characteristics in this study plan, but as we learn more about what is possible with regards to analysing the 2021 GCSE results for this project, this may need to be updated.

programme for the reasons outlined in the analysis section[13]. We outline a number of checks in the analysis section that will be carried out on the Year 11 grades prior to proceeding with the analysis, as well as caveats to consider when reporting the results, and have agreed this exploratory approach with the EEF.

We also succeeded in recruiting an 'evaluation sample' for both primary and secondary phases of education. However, due to a number of challenges, in particular the lockdown in spring 2021, we were not able to recruit enough **comparison secondary schools** so **certain research questions will remain unanswered for this sample** (see 'Sample size calculations: revised calculations').

Original projections of delivery patterns indicated that we would be able to look at the impact of completing whole blocks of tuition. Once the tuition delivery was shifted to later in the academic year due to the lockdown in spring 2021, including later blocks with 10 sessions rather than 15, it became apparent that we would need to look at the **influence of dosage** more closely. As this change to the design was introduced part-way through the year, we are reliant on the data recorded by the TPs. A short consultation was conducted with TPs in April 2021 which showed that almost all TPs said they would be able to supply the date of each tutoring session *or* the total amount of tuition by a given date at a pupil level. We allowed TPs to decide which piece of information (session dates or total amount by 11th June[14]) to supply, with a request that where possible they supplied actual session dates. At the time of updating the study plan the data from TPs is still being processed but it appears that only half of TPs have been able to provide dosage data in the way they had said. We will review the quality and completeness of the data prior to running the analysis and we may need to run the analysis on a sub-set.

Research questions

The research questions are summarised in the following table (Table 2). There are nine research questions, two of which are sub-divided.

In six cases, there are four outcomes: maths, primary; English, primary; maths, secondary; English, secondary.

Research questions 1b, 4a and 4b involve analysis on the population of TP pupils, and as explained above, this is only possible at secondary level now using Year 11 results, due to changes in the summer assessments (i.e. cancellation of Year 6 national assessments).

For each research question, the outcome/phase combinations are identified by a suffix: mp, ep, ms, es, respectively. Hence RQ1a.mp estimates the impact of TP availability on primary school PP pupils' maths attainment. For each specific research question, Table 2 gives the outcome, the phase, the type of pupil for which impacts are estimated, the type of school used in the analysis and whether this uses the evaluation sample or the population of schools. The final column provides a brief description.

Rather than a single primary outcome, each estimator has two or four outcomes (attainment by phase [where applicable] and subject). We will adjust for multiple testing using the Romano-Wolf (2005a,b) simulation approach, as implemented by the Stata program rwolf ado (Clarke,

[13] The EEF is not currently using teacher-assessed grades in other evaluations.
[14] The 11th June was chosen for two reasons: based on the revised delivery projections available at the time; and it was one week before the deadline for schools to submit the Year 11 teacher assessed grades (assuming the teacher assessed grades would need to undergo some internal quality checks and time for submission).

Romano and Wolf, 2020). Impact estimates will be presented with their 95% confidence intervals and Romano-Wolf p-values.

The table retains the analysis originally planned for the secondary school Research Champion sample for completeness. These analyses (see shaded rows, marked: sample*) will no longer proceed due to the low numbers recruited to the secondary school comparison sample (see 'Sample size calculations: revised calculations').

Table 2: Impact research questions						
Impact estimates						
RQ1a: What is the impact of TP availability on PP pupils' attainment?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ1a.mp	maths	primary	PP	TP/non-TP	sample	We focus on PP pupils since they are expected to form a large proportion of the eligible group and can be identified within TP and comparison schools. Using PP pupils avoids the complication of pupil selection as a result of school decision and pupil choice. The drawback is that the resulting estimate relates to PP pupils only rather than to participants as a whole. This analysis will be on PP pupils in year groups where at least one pupil is in receipt of TP (and equivalent year groups in comparison schools). This analysis will use standardised assessments administered at baseline as covariates for primary schools and standardised assessments for the outcome measure at endpoint.
RQ1a.ep	English	primary	PP	TP/non-TP	sample	
RQ1a.ms	*maths*	*secondary*	*PP*	*TP/non-TP*	*Sample**	
RQ1a.es	*English*	*secondary*	*PP*	*TP/non-TP*	*Sample**	
RQ1b: What is the impact of TP on the attainment of pupils participating due to encouragement to do so? TBC						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ1b.ms	Maths	secondary	Participants, pp	TP	Population, Year 11	We draw on the planned Reach & Engagement trials[15] within TP schools. Tutors and pupils are randomised to interventions that aim to improve TP attendance. If the trials will induce sufficient additional TP session attendance to be effective instruments (as opposed to 'weak' instruments) by being significant in predicting participation, we will use the RCT treatment group as instrument in the instrumental variable regression of outcomes on participation. The RCT foundation of this estimator gives it a strong causal
RQ1b.es	English	secondary	participants, PP	TP	Population, Year 11	

[15] Links to the study plans of these trials in the hyperlinks: Leveraging Similarity to Improve Pupil Attendance, Prioritising Tutoring Relationships to Improve Pupil Attendance, Engagement-Boosting Reminders to Improve Pupil Attendance and Engagement.

						interpretation as a local average treatment effect, if the trial is properly conducted, i.e. with balanced treatment and control groups. This analysis will only use TP schools where eligibility can be observed; we will also estimate impacts of the subgroup of PP pupils. This analysis will use as outcomes the teacher assessed grades that are replacing GCSEs in 2021 if available on NPD. It will use KS2 scores as baseline covariates for Years 8-11, standardised tests or KS1 for Year 7.
RQ1c: What is the impact of the intensity of TP (dosage) on the attainment of PP pupils?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ1c.mp	maths	primary	PP	Schools that signed an MOU	Sample	We exploit the fact that some schools enrolled later in the programme and have not delivered it fully or partially by the time of the assessment. The analysis focuses on TP schools and a sample of comparison schools that signed up with a TP (via an MOU) but have not delivered or have only partially delivered TP at the time of the assessment. We estimate a regression with intensity of treatment (dosage) as the independent variable and we instrument dosage with the time when the school signed the MoU, the assumption being that early signed-up are more likely to have greater dosage than later signed-up. This analysis will use standardised assessments administered at baseline as covariates for primary schools and standardised assessments for the outcome measure at primary.
RQ1c.ep	English	primary	PP	Schools that signed an MOU	Sample	
RQ1c.ms	*maths*	*secondary*	*PP*	*Schools that signed an MOU*	*Sample**	
RQ1c.es	*English*	*secondary*	*PP*	*Schools that signed an MOU*	*Sample**	
RQ2: What is the impact of TP availability on the attainment of pupils predicted to participate?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ2.mp	maths	primary	predicted participants	TP/non-TP	sample	We estimate a predictive model of pupil participation within TP schools and use that to predict which pupils participate in both TP and [would participate in] comparison schools. By doing this, we aim to move closer to an impact on TP participants as a whole rather than PP pupils.
RQ2.ep	English	primary	predicted participants	TP/non-TP	sample	

RQ2.ms	*maths*	*secondary*	*predicted participants*	*TP/non-TP*	*Sample**	This analysis will use standardised assessments administered at baseline as covariates for primary schools and standardised assessments for the outcome measure at primary.
RQ2.es	*English*	*secondary*	*predicted participants*	*TP/non-TP*	*Sample**	
RQ3: What is the impact of the availability of TP on all pupils' attainment?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ3.mp	maths	primary	All pupils in years with TP	TP/non-TP	sample	We estimate impacts for all pupils in year groups receiving TP, regardless of whether they participate in TP. These estimates are likely to be smaller than RQ1a and RQ2 estimates since there is no attempt to identify pupils more likely to participate in TP and so its impact will be more diluted. Should TP be spread between a smaller number of schools extending eligibility to a larger proportion of their pupils, this dilution may be reduced. This estimator also captures the effect of spillover (peer) effects. This analysis will use standardised assessments administered at baseline as covariates for primary schools and standardised assessments for the outcome measure at primary.
RQ3.ep	English	primary	All pupils in years with TP	TP/non-TP	sample	
RQ3.ms	*maths*	*secondary*	*All pupils in years with TP*	*TP/non-TP*	*Sample**	
RQ3.es	*English*	*secondary*	*all pupils in years with TP*	*TP/non-TP*	*Sample**	
RQ4a1: What is the impact of TP availability on PP pupils' attainment in the population of schools?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ4a1.my11	maths	Year 11	PP pupils	TP/non-TP	population, Year 11	We estimate impacts for PP pupils in all schools where Year 11 is receiving TP in order to provide a check on the corresponding impacts for sampled schools (mirroring RQ1a). The population refers to all secondary schools in England (so all TP schools compared to all non-TP schools). No population analysis is possible for primary schools as KS2 tests in 2021 have been cancelled. This analysis will use maths and English teacher assessed grades that are replacing GCSEs in 2021 if available on NPD as outcome data. It will use KS2 scores as baseline covariates for Year 11.
RQ4a1.ey11	English	Year 11	PP pupils	TP/non-TP	population, Year 11	

RQ4a2: What is the impact of TP availability on predicted participants' attainment in the population of schools?						
RQ4a2.my11	maths	Year 11	predicted participants	TP/non-TP	population, Year 11	We estimate impacts for predicted participants in all schools where Year 11 is receiving TP in order to provide a check on the corresponding impacts for sampled schools (mirroring RQ2). The population refers to all secondary schools in England (so all TP schools compared to all non-TP schools). No population analysis is possible for primary schools as KS2 tests in 2021 have been cancelled. This analysis will use the maths and English teacher assessed grades that are replacing GCSEs in 2021 if available on NPD as outcome data. It will use KS2 scores as baseline covariates for Year 11.
RQ4a2.ey11	English	Year 11	predicted participants	TP/non-TP	population, Year 11	
RQ4a3: What is the impact of TP availability on all pupils' attainment in the population of schools?						
RQ4a3.my11	maths	Year 11	all pupils	TP/non-TP	population, Year 11	We estimate impacts for all pupils in all schools where Year 11 is receiving TP in order to provide a check on the corresponding impacts for sampled schools (mirroring RQ3). The population refers to all secondary schools in England (so all TP schools compared to all non-TP schools). No population analysis is possible for primary schools as KS2 tests in 2021 have been cancelled. This analysis will use the maths and English teacher assessed grades that are replacing GCSEs in 2021 if available on NPD as outcome data. It will use KS2 scores as baseline covariates for Year 11.
RQ4a3.ey11	English	Year 11	all pupils	TP/non-TP	population, Year 11	
RQ4b: What is the impact of the intensity of TP (dosage) on PP pupils' attainment in the population of schools?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ4b.my11	maths	Year 11	PP pupils	Schools that signed an MOU	population, Year 11	We exploit the fact that some schools enrolled later in the programme and have not delivered it fully or partially by the time of the assessment. The analysis focuses on TP schools and a sample of comparison schools that signed up with a TP (via an MOU) but have not delivered or have only partially delivered TP at the time of the assessment. We estimate a regression with intensity of treatment (dosage) as the independent variable and we instrument dosage with the time when the school signed the MoU,
RQ4b.ey11	English	Year 11	PP pupils	Schools that signed an MOU	population, Year 11	

						the assumption being that early signed-up are more likely to have full dosage than later signed-up. This RQ mirrors RQ1c. This analysis will use the maths and English teacher assessed grades that are replacing GCSEs in 2021 if available on NPD as outcome data. It will use KS2 scores as baseline covariates for Year11.
RQ5: How does the impact of TP availability vary among PP pupils, by school and pupil characteristics?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ5.mp	maths	primary	PP	TP/non-TP	sample	We revisit RQ1a to explore the extent to which impacts for PP pupils vary according to a number of school and pupil characteristics. This analysis will use standardised assessments administered at baseline as covariates for primary schools and standardised assessments for the outcome measure at primary. Population (Year 11) analysis will use KS2 for the baseline covariate and teacher assessed grades (GCSE) for the outcome measure. [This analysis was previously planned for other year groups in secondary schools. In that analysis KS2 SATs from NPD would have be used for Years 8-10, standardised tests or KS1 for Year 7. Outcomes would also have been from standardised assessments.]
RQ5.ep	English	primary	PP	TP/non-TP	sample	
RQ5.my11	maths	secondary	PP	TP/non-TP	population, Year 11	
RQ5.ey11	English	secondary	PP	TP/non-TP	population, Year 11	
RQ6: How do outcomes vary among TP pupils, by model of tutoring?						
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Sample/ population**	**Description**
RQ6.mp	maths	primary	participants	TP	Sample	We examine how outcomes vary according to a number of aspects of TP-related factors using a regression of attainment on tutoring characteristics for the sample of TP only. Since these are only observed among TP schools we present descriptive statistics (i.e. not causal) rather than impact estimates. This analysis will use standardised assessments administered at baseline as covariates for primary schools. For secondary schools, KS2 SATs from
RQ6.ep	English	primary	participants	TP	Sample	
RQ6.ms	maths	secondary	participants	TP	Sample	
RQ6.es	English	secondary	participants	TP	Sample	

						NPD will be used for Years 8-11, standardised tests or KS1 for Year 7 Outcomes will also be from standardised assessments except among Year 11 for whom NPD data (teacher assessed GCSE) will be used.

*some of the RQs were originally planned for the secondary school sample however due to the achieved numbers of comparison schools they will no longer be run. In most cases there is an equivalent RQ on the population data, which is planned on the Year 11 Teacher Assessed Grades, subjects to pre-analysis checks (see section below).

The secondary school evaluation sample analysis would have used the following data: Year 11 analysis would have used KS2 SAT results for the baseline covariate and teacher assessed grades (GCSE) for the outcome measure. For other year groups in secondary schools, KS2 SATs from NPD would have been used for Years 8-10, and standardised tests or KS1 for Year 7. Outcomes would also have been from standardised assessments. Year 7 are the cohort which did not take KS2 statutory tests in 2020 (when in Year 6), so other data sources for a baseline had been identified.

Design overview

Table 3: Design overview – evaluation sample (planned and revised targets)

Design	Matching/weighting Instrumental variables
Unit of analysis (school, pupils)	Pupils qualifying for Pupil Premium
Number of Units to be included in analysis (Intervention, Comparison)[16]	**Primary schools:** Original targets: 191 primary (95 intervention schools and 96 comparison schools), for each of English and maths (382 total schools, although some intervention primary schools may be eligible for both subjects) For each subject, primary: 3629 pupil premium pupils (1805 from intervention schools and 1824 from comparison schools) Revised targets at the end of June: 132 English primary schools (107 intervention, 25 comparison schools); 121 maths primary schools (95 intervention, 26 comparison schools). For primary English: 3168 pupil premium pupils (2568 from intervention schools and 600 from comparison schools) For primary maths: 2904 pupil premium pupils (2280 from intervention schools and 624 from comparison schools) *Achieved numbers:* *174 English primary schools (120 intervention, 54 comparison schools); 154 maths primary schools (105 intervention, 49 comparison schools).* *For primary English: 4176 pupil premium pupils (2880 from intervention schools and 1296 from comparison schools)* *For primary maths: 3706 pupil premium pupils (2520 from intervention schools and 1186 from comparison schools)*

[16] Depending on the method used, the number of units included in the analysis can differ from the pool of potential comparison units. For example, when using matching/weighting the pool of comparisons units could represent all schools in England, but only a certain number of units will be included in the analysis after a suitable match is found. Identifying the precise number of units included might not be possible at the design stage. In these cases Evaluators can speculate on the number of units that are expected depending on the method used.

		Secondary schools (note that analysis no longer being conducted on this sample, only on the population of Y11 pupils participating in TP): Original targets: 151 secondary (75 intervention schools and 76 comparison schools), for each of English and maths (302 total schools, although some intervention secondary schools may be eligible for both subjects) For each subject, secondary: 10570 pupil premium pupils (5250 from intervention schools and 5320 from comparison schools) Reviewed targets at the end of June: 105 English secondary schools (85 intervention, 20 comparison schools); 104 maths secondary schools (82 intervention, 22 comparison schools). For secondary English: 6720 pupil premium pupils (5440 from intervention schools and 1280 from comparison schools) For secondary maths: 6656 pupil premium pupils (5248 from intervention schools and 1408 from comparison schools) *Achieved numbers:* *96 English secondary schools (78 intervention, 18 comparison schools); 94 maths secondary schools (76 intervention, 18 comparison schools).* *For secondary English: 6144 pupil premium pupils (4992 from intervention schools and 1152 from comparison schools)* *For secondary maths: 6016 pupil premium pupils (4864 from intervention schools and 1152 from comparison schools)*
Outcomes	Variable	Educational attainment
	measure (instrument, scale, source)	Attainment in English and maths. Within primary schools this will be based on standardised assessments. Within secondary schools this will be based on assessment, standardised and combined into a single measure.
	Variable	Educational attainment

Baseline for outcome	measure (instrument, scale, source)	• Standardised assessments in English and maths at baseline for primary school pupils. • KS2 SATs from NPD will be used for Years 8-11, standardised tests or KS1 for Year 7.

For information about the population (Year 11) analysis see the section on Sample size calculations: revised calculations.

Participants and selection mechanism

How schools join the TP programme: TP schools

All state-maintained schools in England are eligible to participate in the NTP TP programme. There are three main routes for schools to sign up to receive tutoring:

- schools can contact TPs directly (from 2nd November 2020)
- schools can search for TPs on the NTP website and email them directly through the NTP website
- TPs can approach schools with their tutoring offer (this may include contacting schools they are already working with) (from 2nd November 2020).

Around 6,000 schools are expected to take part in 2020-21. At this stage there is little evidence about which schools will be more likely to take up the NTP TP programme. Once schools have signed a Memorandum of Understanding (MoU) with a TP, they are seen to have joined the programme, and are known in this evaluation as TP schools.

How pupils are selected for the TP programme: TP pupils

The TP programme[17] is focused on supporting disadvantaged pupils in years 1-11, including those eligible for Pupil Premium. As noted earlier in the study plan, schools have the discretion to decide which of their pupils should receive the tutoring, although they are being asked to focus on *'disadvantaged pupils, including pupils eligible for Pupil Premium funding, Free School Meals or those identified by schools as having an equivalent need for support'* (NTP guidance on signing up schools to Tuition Partners [unpublished])[18]. The impact evaluation will collect headline information, at a pupil level, on eligibility for pupil premium and SEND status (yes/no). Note that the IPE surveys will capture school leads' reasons for pupil selection, and interviews will be used to find out more detail (although these reasons will not be matched at a pupil level).

Funding has been announced for 235,000 pupils to receive 15 hours of tutoring each (either one-to-one or small group tutoring). TPs will agree the numbers of pupils to be tutored with each school they are working with. Once schools have selected pupils for tuition, and informed their Tuition Partner of these pupils, these pupils are known in this evaluation as TP pupils.

[17] https://nationaltutoring.org.uk/tuition-partners-funding-round

18 In the EEF online tuition pilot during summer 2020, schools were provided with similar guidance and the proportion of pupils receiving tutoring ranged between 60 % and 79% by tutoring provider (Marshall et al., 2021).

TP Pupil Premium Pupils: TP PP pupils

As noted above, within a school, the guidance stipulates that the tutoring should be targeted at disadvantaged pupils. Since the group of eligible pupils cannot be identified within non-TP schools (i.e. comparison schools – see below), much of the analysis in this evaluation focuses on those eligible for the Pupil Premium since these represent the core of the eligible group and are identifiable in both TP and non-TP schools. We refer to these pupils as TP PP pupils.

PP pupils may not coincide with the group of children who will receive the intervention. RQ2 involves an alternative approach to approximating the eligible group. We intend to estimate a predictive model of pupil participation within TP schools and use that to predict which pupils participate in both TP and [would participate in] comparison schools.

Population descriptive statistics

Part of the evaluation will involve descriptive statistics of reach and spread of the provision overall – both in terms of the TP schools and TP pupils in the TP population. TP population data will be used to describe the number of Tuition Partners (TPs), schools and pupils involved in the TP programme, and their characteristics (such as Pupil Premium), as well as an overview of attendance at sessions. The population data is collected and recorded on an ongoing basis by TPs, and provided to the Evaluator once a term. We expect this population data to be incomplete (as it is being provided by 33 different organisations, for 6,000 schools and over 200,000 pupils). The evaluator will collate, but not quality assure, the data that is provided. The Evaluator will match the population data to the NPD.

Recruitment to the evaluation sample ('Research Champions')

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 must sign a memorandum of understanding (MoU). Once signed up, schools will be contacted by the Evaluator to invite them to take part in the impact evaluation as a sampled school. Initially all schools that signed an MoU up to 22nd January 2021 were contacted with information about how to assist the impact evaluation further and inviting them to sign up to participate in the impact evaluation with details of financial incentives available (via the Evaluation Online Form) – for details on the incentives please see Appendix A. This involved finding out which standardised tests (provider, subject and timing/version) the school is using in the 2020/21 financial year and for which year groups, if any. In response to the national lockdown period and changes to delivery, we expanded the recruitment and contacted all schools participating in TP to take part in the evaluation as Research Champions. In part this was related to difficulties recruiting schools during a challenging period of disruption to education and 'bubble' closures; and also because a substantial amount of delivery was being pushed towards the end of the academic year, it became more important to consider dosage in the study design. This meant more schools would be eligible for the analysis as we would take account of dosage.

Schools in the impact sample were asked to provide access to their standardised assessment data (for more detail see the section on 'Outcome measures and other data'), and provide additional pupil details and tutoring attendance data. Recruitment of schools to the impact sample was closely monitored by key variables 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[19]. However challenges to recruitment meant that if schools met the key testing and delivery criteria (below) they were accepted into the evaluation sample. Schools signing up to the evaluation sample will be able to call themselves TP School Research Champions, to emphasise their additional role and commitment to research.

In order for a school to take part in the intervention evaluation sample (and receive the incentives), the school must have at least one school year in which pupils are receiving tutoring and in which standardised assessments are being conducted in the tutoring subject (maths/English); and agree to provide the pupil data needed for the evaluation.

The testing requirements differed by phase, reflecting the patterns of testing observed in schools[20]. In order to be eligible:

- primary schools needed to have conducted a standardised assessment at baseline, prior to tutoring (i.e. in the autumn 2020 term) and be planning to or willing to conduct another assessment in the summer 2021 term, after tutoring is completed.
- secondary schools needed to be planning to or willing to conduct a standardised assessment in the summer 2021 term, after tutoring is completed. We found that only around half of secondary schools had already completed a baseline assessment in the autumn term, and the lockdown in January 2021 removed the opportunity to administer a baseline with these schools prior to tutoring. Therefore we had to remove the requirement for a within-year baseline. This will be replaced with KS2 tests for Years 8-11, and standardised assessments or key stage 1 data for Year 7 pupils, as the key stage 2 tests were cancelled when they were in Year 6 in 2020.

When a secondary school was keen to participate and did not have a standardised testing regime in place, they were invited to administer RL or GL tests during the summer term[21].

A number of school-level characteristics informed the recruitment of schools to the evaluation sample. These are informed by the research questions, specifically the phase and subject of tutoring, which have minimum numbers required for the power calculations. In order to

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

20 The coverage of assessment providers is quite high in primary schools, where our discussions with assessment providers indicate that about 75% of schools have a standardised testing regime in place, but less so in secondary schools, where about 50% of the schools have a testing regime. This was been reflected in the school responses to our invitation to take part

21 Initially the offer was also available for primary schools however there were enough primary schools with the required testing pattern already in place so this was not needed

minimise burden on schools, it will be necessary to prioritise schools that have an existing testing regime in place that covers year groups in which tuition is happening.

The key characteristics monitored for high/low coverage were:

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

We were also aiming to ensure a good spread of other key characteristics, however 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. Therefore, due to the sample size and the changing circumstances that schools were facing compared to when we started the project due to the January 2021 lockdown, we limited the key characteristics to the above list. Where possible we also monitored the following, but these did not direct the recruitment process to the same extent:

- Tuition Partner
    - Mode of tutoring (face to face/online/both; individual/small group tutoring)
    - Pattern of year groups tutoring
- % eligible for free school meals (FSM)
- Region
- Attainment
- School type.

Selection of the comparison group and identification assumptions

We aimed to recruit a sample of non-TP schools that shares similar characteristics to the TP schools in the evaluation sample. To this, we selected all primary and secondary schools in England that were either Academies, Colleges, Free Schools, or Local authority maintained schools. Schools were matched on the characteristics in Table 5.

The eligibility requirements around testing patterns for the comparison schools mirrors that of the Research Champion/ evaluation sample:

- only primary schools with a testing regime in place will be recruited in the comparison school sample. This means that comparison primary schools will need to have done the baseline assessment at the time of recruitment and be planning to or willing to conduct a standardised assessment in the summer 2021 term.
- secondary schools needed to be planning to or willing to conduct a standardised assessment in the summer 2021 term. We will use KS2 data from NPD as the baseline for Year 8-11. For Year 7, the cohort that did not sit KS2 SATs in 2020, we will use standardised tests if available. If no standardised test data is available then we will use KS1 tests.

For each TP school signed up to the evaluation we identified eight similar-looking non-TP schools (based on characteristics listed in Table 5), with the aim of recruiting one of them as a comparison school. Before we contacted the schools in the comparison school sample, we shared the list of schools with our assessment partners, and asked them to tell us, where possible, which of the schools on the list are using their tests to inform recruitment approaches.

As outlined in the outcomes section below, the comparison schools were asked for permission to access their pupils' test results directly and securely from the test provider. In schools not using the online marking tools, or schools not using those tests, we worked with the schools to establish what standardised tests data they have available, and provide access to tests if necessary.

The challenges and additional burden faced by schools during the spring 2021 term have affected the recruitment of control schools, making it difficult to recruit all required schools from the lists of matched ones. Schools in both the intervention and comparison samples were offered a financial incentive to take part (see Appendix A). However comparison schools have 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.

Therefore in addition to the matched comparison sample (also see section on PSM, below) we also approached two further groups, similar to TP schools in terms of motivation to participate in the NTP:

- schools that completed an Expression of Interest (EoI) form on the NTP website but who did not go on to sign up (via a Memorandum of Understanding, MoU) to the programme
- schools that signed an MoU but who did not go on to start tuition.

We approached these groups due to a very slow response from the matched comparison group. These schools can act as control as they showed a similar interest in TP as 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 (see Section *Primary outcome analysis)*. The data requested from schools in these groups is the same as that described for the matched comparison group. Depending on the final achieved numbers in each group, schools from these 'TP interested' groups may need to be included in the final analysis.

We asked comparison schools to provide information about their PP pupils[22], at a pupil level about whether they have received any formal tutoring during the 2020/21 academic year (arranged by the school). This will be reported as descriptive analysis of the extent of tutoring among the comparison group.

A note on the sample

As outlined in the recruitment sections for the impact sample (Research Champions) and the comparison sample, decisions around selection and recruitment have been informed by

22 For PP pupils in year groups that have completed the standardised assessments.

practical considerations, as well as the methodological design. We recognise that this study is being conducted in atypical circumstances (the ongoing Covid-19 pandemic) and we are conscious of the pressures and challenges that schools are facing. This is one of the main drivers of our decision to use standardised test data that is already being administered by schools, so as not to place additional burden on pupils and schools.
Together with EEF we also recognise that for ethical reasons relating to the loss of teaching time due to Covid-19 it would not be appropriate to randomise, withhold or restrict access to the tutoring for schools and pupils that wish to receive it.
The planned population analysis (Year 11 teacher assessed grades) is intended to serve as an important check against the analysis on the impact sample, as outlined below. The advantage of the population analysis over the evaluation sample analysis is that there is no selection involved other than choosing to become a TP school. The relative disadvantage is that is limited to Year 11 (i.e. and not Year 6 as their national assessments were cancelled this year) and does not use recent baseline assessments.

Sample size calculations: original assumptions

We use cluster randomised trial power calculations to provide a lower bound indication of the MDES. We allow for clustering of pupils within schools. The evaluation is not based on a randomised intervention but instead relies on a non-experimental approach. The true MDES is therefore likely to be somewhat higher given the non-random nature of selection.

We assumed that some 'non-TP' (comparison) schools may decide to sign-up to NTP during the course of the year and during the evaluation, thereby moving from the comparison group into the intervention group. To account for this, we measure the intervention not only as a 0/1 dummy, but also as a categorical variable indicating the fraction of hours completed by the time of the assessment. We have allowed for the same level of attrition in both groups. Assumed sample size (Table 4) is based on a conservative level of take-up (in our original calculations we assumed 19 pupil premium pupils per school at primary and 70 secondary[23]); this allows for schools with incomplete coverage of year groups, either in terms of tutoring or testing. It also reflects our intention to focus primarily on Pupil Premium pupils.

Table 4a contains the sample sizes we planned to achieve, i.e. pupils with: (1) baseline [primary only] and (2) end-point assessment, and for whom we have (3) the required pupil data (for matching to NPD, tutoring attendance data). Based on the three data points, we originally intended to over-recruit schools by 20 per cent, 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
- 378 secondary schools:
  - 189 secondary schools (95 intervention and 94 comparison) for English
  - 189 secondary schools (95 intervention and 94 comparison) for maths

[23] These figures were updated in our revised calculations, below.

These numbers assumed no overlap between English and maths so the total number of schools was likely to be lower given overlap.

For both primary and secondary schools, we assume 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. We assume similar parameters for primary and secondary schools, noting that the EEF Catch-up Literacy and Catch-up Numeracy evaluations likewise assumed ICCs that were quite similar.[24]

Note that since our analysis focuses on disadvantaged pupils, we do not produce separate estimates for the FSM subgroup.

For both primary and secondary school pupils, we originally estimated a MDES of 0.13. We note that, among comparable EEF trials, the ICC among the achieved sample is sometimes higher than that assumed when designing the trial. For instance, with the Tutor Trust re-grant, the actual ICCs were 0.29, 0.17 and 0.23 for maths, reading and GPS (Grammar, Punctuation and Spelling), respectively rather than the assumed 0.19. Increasing the assumed ICC in our case to 0.30 increases the MDES to 0.17 and 0.18 for primary and secondary schools respectively. For context, the EEF toolkit pages on one-to-one tuition (EEF, 2018a) suggest a weighted mean effect size of 0.37 and the pages on small group tuition (EEF, 2018b) suggest a weighted mean effect size of 0.31. It would however be expected that any potential impact of doing TP would be smaller than this given the large scale of the roll-out and the variation in implementation that was expected upfront across the different Tuition Partners.

This calculation used the R package 'PowerUpR'.

Table 4a: Original sample size calculations for estimation sample (post-attrition) from the original study design

		Study Plan			
		Primary English	Primary Maths	Secondary English	Secondary maths
Minimum Detectable Effect Size (MDES)		0.13	0.13	0.13	0.13
Pre-test/ post-test correlations	level 1 (pupil)	0.70	0.70	0.70	0.70
	level 2 (school)	0.70	0.70	0.70	0.70

[24] Our assumptions regarding the ICCs are, if anything, conservative. Allen et al. (2018) suggest ICCs of 0.07 at the end of KS1, 0.10 at the end of KS2 and around 0.15 at the end of KS4.

Intracluster correlations (ICCs)		0.15	0.15	0.15	0.15
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		19	19	70	70
Number of schools	Intervention	95	95	75	75
	comparison	96	96	76	76
	Total	191	191	151	151
Number of pupils	Intervention	1805	1805	5250	5250
	comparison	1824	1824	5320	5320
	Total	3629	3629	10570	10570

Sample size calculations: revised calculations

The recruitment of the comparison sample was slower than required. We reviewed the MDES calculations using the sample of schools recruited to the Research Champions that shared pupil data by the end of August. The other research questions are based on Instrumental Variables design and it is not possible to reliably compute power calculations in these contexts. All the other assumptions reflected those made originally. Updated MDES calculations are provided in the Table 4b below. The figures of PP pupils were updated to 24 in primary and 64 in secondary as indicated from data of schools that signed an MoU in March[25]. The **achieved** sample sizes in both primary and secondary 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
- 190 secondary schools:
  - 96 secondary schools (78 intervention and 18 comparison) for English
  - 94 secondary schools (76 intervention and 18 comparison) for maths

Once the data has been cleaned and matched to the assessment data, these figures may still change (e.g. if assessment data is missing, or if in TP/intervention schools the timing of tuition is later than the assessment, some schools may move from intervention to comparison).

[25] These figures are based on the pupil data submitted by TPs at the end of March 2021. More schools and pupils signed up during the summer term so these figures will be checked again using the data at the end of Year 1.

Table 4b: Sample size calculations for estimation sample (post-attrition) revised summer 2021, based on achieved numbers of schools recruited

		Study Plan			
		Primary English	Primary maths	Secondary English	Secondary maths
Minimum Detectable Effect Size (MDES)		0.14	0.15	0.21	0.21
Pre-test/ post-test correlations	level 1 (pupil)	0.70	0.70	0.70	0.70
	level 2 (school)	0.70	0.70	0.70	0.70
Intracluster correlations (ICCs)		0.15	0.15	0.15	0.15
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		24	24	64	64
Number of schools	Intervention	120	105	78	76
	comparison	54	49	18	18
	Total	174	154	96	94
Number of pupils	Intervention	2880	2520	4992	4864
	comparison	1296	1186	1152	1152
	total	4176	3706	6144	6016

The number of comparison secondary schools recruited was considerably lower than planned. As the MDES rose above 0.20 in the secondary sample (the original design had an MDES of 0.13 and the revised MDES after recruitment was 0.21) it was decided not to pursue analysis for RQs 1a, 1c, 2 and 3 in the secondary school sample (instead see the mirror RQs 4a and 4b for the population analysis), and for RQs 1b and 5 the analysis will be carried out on the Year 11 population data rather than the secondary school evaluation sample. All the other RQs will be retained.

For the year 11 **population analysis**, there are 1554 unique secondary schools doing TP. We assume 80% of secondary schools have usable pupil data, which leaves us with 1243 schools (this assumption of 80% may be revised as the data is cleaned). Power calculations suggest that with 1243 TP secondary schools and 1243 non-TP schools the MDES would be 0.03.

Outcome measures and other data

Key principles

The QED aims to look at the impact of tutoring on attainment, as the purpose of the NTP TP is to support pupils to 'catch-up' and reduce the amount of learning lost due to the COVID-19 pandemic and the restrictions on schools in 2020 and 2021.

We are aiming to reduce burden by, wherever possible, using assessment data that schools would be collecting anyway in 2020/21. Furthermore, we recognise that although there might be a focus on tutoring Years 6 and 11, tutoring is taking place in *all* year groups from Year 1 to Year 11. Consequently, we will collect assessment data for all year groups receiving tutoring and not just the year groups expected to take national curriculum tests. A third reason we planned to utilise existing standardised test data is because Key Stage 2 tests and GCSEs were at risk of disruption due to the pandemic. Since the original design, the Key Stage 2 tests were cancelled for 2021, and GCSEs grades will instead be awarded based on teacher assessments.

We will use tests that have been standardised using a nationally representative sample of students in the recent past. All the tests developed by our consortium members that we propose for this study (see below) have been standardised in this way and many schools use these tests, and standardised tests from other testing companies[26], on a regular basis. Standardised tests yield standardised scores, which have a fixed mean and standard deviation. This, in turn, means that if we align their scales using their national standardisation parameters, standardised scores from different tests can be analysed together thus allowing impact to be measured across all year groups in each phase simultaneously.

All optional test standardised scores will have a published mean and standard deviation from when they were standardised with a nationally representative sample. We will determine these parameters for Key Stage 2 tests (for the baseline covariate in secondary schools) and the teacher assessed GCSEs from the population. In each case, the following formula will be 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).

For any such amalgamation of tests into a single analysis (in this case by phase and subject), the statistical challenge is one of variance and the interpretational challenge one of validity. It should be noted that GCSEs themselves are made up of several exam boards (and therefore papers) and it is common practice to analyse these together in RCTs and other evaluations. By including different measures, we are increasing measurement error and muddying what domain of learning we are measuring. We have deliberately chosen respected test providers who make it their business to write tests that are aligned with the present National Curriculum taught in English schools. It will be important to report information such as how recently each test was standardised, what domains their individual components are purported to measure and their reliability. This detail will give context around their amalgamation for analysis and aid in its interpretation. Given the interaction between tutors, tutees and their test performance is likely to be quite different between school phases and subjects, we are not amalgamating beyond this level. Indeed, even within a phase it might be problematic to analyse tests from

[26] When inviting schools to participate we asked which other tests they were using in case we could include more providers, however none of the other examples used by responding schools were standardised tests.

different year groups together for the same reasons. This is part of the reason why we originally planned a population analysis since, at least at KS2, it would have involved everyone sitting exactly the same test. Relying on this analysis exclusively, however, meant only two of 11 year groups would have been covered. Furthermore, we had identified this analysis to be at risk anyway due to possible Covid-19 induced exam cancellations and this was borne out in January 2021 when the KS2 tests were cancelled. We are concerned that the process for grade determination may mean that it is difficult to detect any potential impact of the TP programme for the reasons outlined in the analysis section.

Test choice will be driven by the schools rather than the evaluation. This approach will ensure minimal disruption to schools' existing testing regime where they are already using standardised tests before and after TP tuition. As outlined above, where possible we will work with schools with a testing regime already in place for the academic year. However if this does not yield enough schools for the sample, and if there are schools wishing to take part in the impact evaluation who have not conducted or booked standardised tests with their pupils, we will work with the school to schedule end-point tests after tuition has finished[27]. In such cases, involvement in the evaluation provides rich attainment data on pupils in the chosen year groups during their catch-up journey.

For the population analysis, it will not be necessary to standardise GCSE [teacher assessed] grades as we are not planning on amalgamating measures for this analysis.

In order to match the attainment data to the pupils, we will collect data about pupils in TP schools and comparison schools. Box 1 below outlines the data to be collected about different groups of pupils covered by the evaluation.

Box 1: Data to be collected:

Named/NPD data collected for the population of TP pupils (RQ4):

Collect pupil data (name, DOB, UPN) from all TP pupils via Tuition Partners.

This named data will be matched to NPD to collect the longer list of background variables for the population analysis

For the impact evaluation sample (RQ1a, RQ1b, RQ1c, RQ2, RQ3-RQ5, RQ6 [intervention only]) - Named/NPD data collected from the intervention and comparison schools:

TP (intervention) schools: collect pupil data/assessment data/tutoring attendance data* IF there is at least one pupil in the year group receiving tuition and the year has done/is doing standardised assessments.

[27] Originally we also planned to offer baseline tests to schools not already using them, however as explained elsewhere, the lockdown in January 2021 meant the window for new baseline tests closed and we had to proceed without them.

Comparison schools: collect pupil data/assessment data from all year groups with assessment data

* the assessment data mentioned here includes NPD data for Year 11. Note that if a school does not already use standardised assessments but was willing to administer tests as part of the evaluation, we would collect assessment data for PP pupil only (as the analysis is on PP pupils, not [exclusively] TP although there will be overlap).

Use of de-identified data in the population analysis:

We will access de-identified attainment data from NPD for all Y11 pupils in all English schools as part of the population analyses (RQ1b, RQ4a, RQ4b, RQ5). This will use the prior national curriculum assessment (KS2) as the baseline and the teacher assessed GCSE grade in 2021 as the endpoint. We will need it from all schools to establish a comparison.

We would then add in the named data about the TP pupils so we know which pupils in Y11 received tutoring. Therefore the population analysis in the robustness check would be a hybrid of named (covered above) and de-identified data.

Baseline measures

For all year groups, the baseline measures used will be the standardised test taken by pupils in the autumn term 2020 (or early spring term 2021, so long as it is before the start of TP tuition in the school) for all primary schools. This will include tests in maths and literacy, analysed separately by phase.

Agreements have been made with four providers of standardised tests for them to share the test results (overall standardised score) with the Evaluator directly and securely, for schools that give permission to do so. The assessment providers will share data for the entire year group in participating intervention schools if at least one pupil in that year group is receiving tutoring. The four providers (and phase of education covered) are:

- NFER (primary)
- GL Assessment (primary and secondary)
- Renaissance Learning (primary and secondary)
- Rising Stars (primary).

If a school uses tests supplied by the providers listed above, but does not use their online marking system, they are able to participate by providing their test results (overall score) with the Evaluator. These will then be converted to standardised scores using a look-up table.

If a school uses tests from another supplier not listed above we will check that the test is standardised and request its reliability from the provider, and these schools will also be able to participate by providing their test results (overall score) with the Evaluator.

As noted above, the requirement for a within-academic year baseline has been dropped for secondary schools due to the lower incidence of testing reported by schools and the January 2021 lockdown which prevented the administration of baseline tests in schools. Instead, for Years 8-11 we will use KS2 SATs as baseline from NPD; for Y7 (the cohort that did not sit KS2 SATs in 2020) we will use standardised test data if available, or KS1 results (from NPD) if no standardised test data is available.

Outcome data

For Year 11 we will use the teacher assessed GCSE grades, which we understand will be available on NPD (without school level variables).

For all other year groups, in each school we will use standardised test data. In primary schools this will be from the same provider used at baseline where possible. We will work with each school to support them to schedule the assessment for after the TP tuition has been completed where possible[28]. This will include tests in maths and literacy.

NPD data for evaluation sample

As our current understanding (September 2021) is that the teacher assessed GCSE grades will be available in the NPD, we will ask the NPD team to match in named data to our population and evaluation sample pupil list via school Unique Reference Number (URN) [29], Unique Pupil Number (UPN), name and date of birth (DOB). In addition to the results of teacher assessed GCSE grades, we will request the variables listed below.

In order to establish comparison groups, we will request de-identified data for all pupils born between 1st September 2003 and 31st August 2015 including their final KS2 and/or KS4 pupil data from 2014/15 until 2020/21 (including those with endks=0 or schres=0).

We will collect the following variables:

- Basic data for matching to NPD (name, date of birth, Unique Pupil Number - UPN)
- Background characteristics such as gender, ethnicity, socio-economic status
- Information on pupil performance / attainment
- Special educational needs
- Attendance / exclusion
- Interactions with social services

We will adjust for multiple testing using the simulation approach of Clarke et al. (2020), as implemented by the Stata program wyoung.ado. Testing many hypotheses may give rise to concerns about multiple inference. In this context, the probability that we incorrectly reject at least one null hypothesis is greater than the significance level used for each individual hypothesis test. We address this multiple inference concern by controlling for the family error rate, i.e. the probability of incorrectly rejecting one or more null hypotheses belonging to a family of hypotheses. When testing multiple hypotheses, we will calculate family adjusted p-values based on 1,000 bootstraps of the procedure of Romano-Wolf (Clarke et al., 2020).

[28] We recognise that this might not always be possible as delivery of TP has shifted to later in the academic year due to the lockdown in the spring term 2021. We are aiming to collect the start and end dates of the tutoring and dosage data from TPs where this exists, as well as the date the standardised test was taken/uploaded to the assessment provider's system. We have less control over the timing of the year 11 teacher assessed GCSEs which have to be submitted by teachers by 18th June 2021.

[29] Our understanding is also that URN and other school level identifiers will need to be removed from the dataset once the match has been completed, due to restrictions on school level analysis of the 2021 teacher assessed grades.

Propensity Score Matching

Our sampling approach used matching to control for school selection into TP by constructing a matched comparison group of non-TP schools that was similar in important regards to the TP schools in the evaluation sample (see above and Table 5, below). 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.

Table 5: Variables used for matching

Factor	Indicator(s)	Hypothesised Justification	Influence decision to take up intervention ?	Impact on outcomes?	Influenced by outcomes?	Used for matching?
School level achievement	% of Students (disadvantaged and not) who achieved expected standard in KS2 in 4 years before	Low previous attainment explains why schools would use the intervention	Yes	Yes	No	Yes, only the average score regardless disadvantage, and only the year before
	Reading, Maths and Writing progress measure in 4 years before (for disadvantaged and not)	Low previous attainment explains why schools would use the intervention	Yes	Yes	No	No
	Average Maths and English attainment (for disadvantaged and not), 4 years before	Low previous attainment explains why schools would use the intervention	Yes	Yes	No	No
	Average attainment at KS1, at Local Authority level	Low previous attainment explains why schools would use the intervention	Yes	Yes	No	Yes
School type and size	Management primary – Community, Academies, Foundation, Free schools, Sponsored Academies, Voluntary school	Type of school is correlated with choice of using the intervention	Yes	Yes	No	Yes

Factor	Indicator(s)	Hypothesised Justification	Influence decision to take up intervention ?	Impact on outcomes?	Influenced by outcomes?	Used for matching?
	Management secondary – Community, Academies, Foundation, Free schools, Sponsored Academies, Voluntary school, Studio schools, University Technical college.	Type of school is correlated with choice of using the intervention	Yes	Yes	No	Yes
	School type, primary, secondary or both	Part of design		No	No	Yes
	School size, total number of students in pre-year	The school size is negatively correlated with the probability of receiving the intervention		Yes	No	Yes
	Total income per-pupil	Correlated with disadvantage	Yes	Yes	No	No
	Teacher-student ratio	Correlated with schools wanting to use the intervention	Yes	Yes	No	Yes
	Ofsted, overall effectiveness	Low ranking correlated with schools wanting to use the intervention	Yes	Yes	No	Yes
Location	School in urban/rural area	Ensure coverage of similar school types.	Yes	Yes	No	Yes
	School in London, GRO and regional dummies	Ensure coverage of similar school types and of geographical representation.	Yes	Yes	No	Yes
	IDACI quintile, 2018 and interaction of IDACI tertiles with avg attainment in the previous year	Neighbourhood Neighbour deprivation correlated with schools wanting to use the intervention	Yes	Yes	No	Yes
	Index of Crime, LSOA	Neighbourhood crime rate correlated with schools wanting to use the intervention	Yes	Yes	No	No

Factor	Indicator(s)	Hypothesised Justification	Influence decision to take up intervention ?	Impact on outcomes?	Influenced by outcomes?	Used for matching?
	Index of housing, LSOA	Neighbourhood deprivation correlated with schools wanting to use the intervention	Yes	Yes	No	No
	Areas previously under tier 3 lockdown in 2020/21	Lockdown may affect attendance and participation to the programme	Yes	Yes	No	No
Students' demographics	Free School Meals – % eligible in pre-year	Correlated with schools wanting to use the intervention	Yes	Yes	No	Yes
	EAL - %, in pre-year	Correlated with schools wanting to use the intervention	Yes	Yes	No	Yes
	SEN - % in pre-year	Correlated with schools wanting to use the intervention	Yes	Yes	No	Yes

We matched on variables listed in Table 5 as they differ between TP and non-TP schools. The variables listed in table 5 slightly differ from those published in the first version of the study plan. The list does not include variables that were not significant in affecting TP status (staffing variables) and the school management category has been edited to avoid several missing values. The decision not to include not significant variables is also supported by the need to have a parsimonious set of school level controls to reflect the smaller sample size of the RC sample.

Some variables were listed in the first version of the study plan but not included in the matching. We included achievement from the previous academic year, rather than from the previous four academic years as these values, while jointly significant, are highly correlated. The per-pupil income is not exactly comparable between academies and local authority schools (as academy trusts have higher expenditures/income due to they do administration which would ordinarily be done by a local authority) and therefore we dropped this variable from the analysis. The index of Crime and the index of Housing were not significant in predicting TP participation, once we controlled for IDACI index, and were therefore not included from the analysis. Finally, the dummy for the area having been under tier 3 lockdown is correlated with the regional dummies and not used.[30]

In our case the matching approach was complicated since it forms part of the recruitment strategy. It is set out in detail in the steps below:

1. Identify the TP schools and the non-TP schools. This will be on the basis of a cut-off; those not participating in TP by end-March 2021 will be viewed as non-TP schools.
2. Pooling the TP and non-TP samples, estimate a probit regression of TP participation on the variables listed in the table below.
3. Using the estimation results, predict the probability of being a TP school (the propensity score)
4. Drop those TP schools whose propensity score is more than a specified proportion of a propensity score standard deviation above a specified percentile of the propensity score standard deviation among the non-TP schools. These proportions and percentiles will be identified through preliminary analysis described below. The purpose of this step is to ensure the selected non-TP schools are sufficiently similar to the TP schools to as suitable comparators (that is, to identify non-TP schools on the common support).
5. Working through the TP schools in reverse order of their propensity score, match each one to its nearest non-TP school without replacement, subject to the propensity scores being sufficiently close. This can be done in Stata using the psmatch2 ado file by issuing the command: "psmatch2 treatment_variable xvariables, descending noreplacement caliper(caliper_width)". The preferred caliper_width will be identified

[30] We may need to review our approach to the school-level variables in NPD, once we know which/how many school-level variables we are able to match into the NPD. We understand that there are likely to be restrictions on school level data in NPD for the 2020/21 cohort due to the move to teacher assessed GCSEs and the removal of school accountability linked to the 2021 grades. See: https://www.gov.uk/government/news/extra-measures-to-support-students-ahead-of-next-summers-exams

through preliminary analysis. We choose to match without replacement since the aim of matching is to identify schools to recruit into the evaluation sample.

6. Save the matched non-TP schools to a file; call this file NonTP1.
7. Repeat steps 4-6 m times, each time saving a file of matched schools and then removing those selected from the pool of potential comparator schools (i.e. if a school in NonTPi is matched then this will be removed for groups m>i). This results in m files NonTP1, ,,,, NonTPm.
8. Send the list of the NonTP1, ,,,, NonTPm schools to the assessment providers for them to identify which of them have a testing regime in place in the current academic year, for which year groups and subjects.
9. Schools in the matched samples that are testing for the relevant year groups and subject will be contacted and recruited starting with NonTP1 and working down the list.

Matching used the NFER Register of Schools, which includes all primary and secondary schools in England. It was augmented with additional variables on the educational attainment of disadvantaged children from the DfE school performance tables. We emphasise that matching will be carried out as a means of identifying the sample and therefore is prior to outcomes being known.

Preliminary analysis

To test whether propensity score matching controls for school selection, we will conduct placebo tests for the TP schools in earlier years using the data from the NFER Register of Schools. If the selection of the control sample controls adequately for unobserved factors, we do not expect to find any significant difference in attainment between TP and control schools prior to the intervention. We will perform a placebo test at the outset (before the recruitment of comparison schools) as a way to inform our approach and sampling. We will then perform a second placebo test after constructing the sample of matched comparison schools to check the similarity of the two groups of schools before the intervention. The placebo testing will be done for each of the four preceding years, using results for KS2 and GCSE to demonstrate similarity of the achieved match in the past. The placebo test on the three years before was conducted in March 2021, when we drew the lists of potential comparison schools. They showed similarity between matched and comparison samples for both primary and secondary schools.

We will finalise our preferred estimation approach (e.g. random effects or fixed effects in ebalance) once we have done the placebo testing and once we have the data, which has two main elements:

1. Placebo testing to identify target comparison schools for recruitment. We will use the variables listed in Table 5 to match TP schools to non-TP schools. We will assess the performance of the match by a) comparing observed characteristics of TP schools and their matched comparators and b) comparing outcomes of TP schools and their matched comparators. This will be a school-level analysis but we will be able to consider outcomes for PP pupils.
2. Placebo testing to assess the performance of potential estimation approaches. We will compare observed characteristics of TP and matched comparator schools. We will apply the estimation approach (e.g. random effects, diff-in-diffs, including fixed effects in ebalance) in pre-TP years as an additional form of placebo test. We will then repeat this after simulating

impacts in order to understand the sensitivity of the estimators. As part of this exercise, we can assess whether approach is likely to be sufficiently sensitive.

We will report ordinary least squares (OLS) regressions in the Appendix and show a sensitivity analysis on the use of weights (with and without) from the PSM.

Analyses

Population (Year 11) analysis considerations

As outlined above, the secondary school analysis will use teacher assessed grades awarded in 2021 instead of GCSE scores. There are several concerns on the appropriateness of using teacher assessed grades as an outcome measure that we have considered:

> Concern 1: That teacher-assessed GCSE grades may not be an accurate representation of pupil performance. Schools may have 'bumped up' GCSE grades in 2021, particularly around the grade 3/4 boundary, during the teacher assessment process[31].
>
> Concern 2: Knowledge/selection of pupils doing TP led to bias (conscious or unconscious) in the teacher assessed grades. This could lead to positive bias (as they know these students have had additional support), or negative bias (as these pupils have been previously identified as struggling). Pupils eligible for Pupil Premium funding were also disproportionately represented in the group of pupils who received tutoring via NTP and teacher unconscious bias may lead to lower awarded grades for these pupils.
>
> Concern 3: There are uncertainties around whether the teacher assessed grades will reflect pupils' performance after the tutoring. Schools may have used work produced over the year to reach their final teacher assessed grade, rather than performance in a test at a fixed time point. This may lead to grades not reflecting a pupil's latest performance.
>
> Concern 4: Whether the assessments are sensitive enough to change. This concern is linked to the three prior concerns, with all of these potentially affecting the measure's sensitivity to change.

Therefore we plan to conduct some checks (ex ante) before we start the impact analysis and checks (ex post) while performing the impact analysis to inform the presence of any of the above concerns. The checks will inform the approach to the analysis and our interpretation of the results. While these checks will be helpful, it is important to note that they will not be able to detect with certainty whether there is any systematic bias (i.e. if the tests fail to detect systematic bias, that will not mean that there is no systematic bias) therefore the findings will need to be treated with caution.

---

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

The tests that we will carry out are as follows.

Ex ante tests (across all schools):

i) To address the concern that GCSE teacher-assessed grades may not be an accurate representation of pupil performance and that schools may have bumped up grades this year (concern 1), we will compare the distributions of GCSEs awarded 2 and 3 years prior (2019 and 2018) and this year (2021) for all pupils and for PP pupils (as a group in itself) across all schools. If the distribution of grades across the years is significantly different for both groups of pupils, this is a potential concern we will account for in the interpretation of results. However we do not view this as a reason to not run the analysis as we already know that 2021 results are different to prior years, and as Ofqual points out it is not possible to disentangle whether this is due to the change in assessment process, or due to the disruption to education during the pandemic[32]. If teachers do bump-up grades we would expect the distribution to become more bimodal, with the ones at the grade 3/4 ('fail'/'pass') threshold moving up the distribution. If the distribution of grades across the years is significantly different for both groups of pupils, we plan to conduct the analysis anyway as the test is not indicative of systematic bias across TP and non-TP schools, but may provide an explanation for assessments not being very sensitive to change.

ii) To address the concern that schools select the pupils who undertake TP and thus may apply some conscious or unconscious bias in the assessment (concern 2), we plan to use across-subjects variation to help identify if any bias is subject specific and not pupil specific in the sample of pupils tutored. As long as any bias is a teacher bias and each teacher teaches a different subject, the cross-subject comparison should be fine: if the bias is across subjects, then the cross-subject comparison cannot reveal any systematic difference. We also note that we are concerned here only with secondary schools, and we suspect that it is unlikely that all of their teachers would be aware that a particular pupil is receiving tutoring in another subject. If it was at primary, of course we believe this would be an important consideration as a single teacher would usually be teaching and assessing all subjects, however at secondary the pupils are likely to have as many teachers as subjects. If we find evidence of teacher bias at pupil level, this would represent a serious concern to the validity of the analysis, as it may point to systematic bias in TP schools versus non-TP schools. However, it is worth caveating that we may not be able to detect this bias: because of the pandemic pupils with different characteristics showed heterogeneous performance, which may show up across pupil characteristics and across subjects.

[32] see previous footnote

If this test reveals the possibility of systematic bias, we will proceed to carry out tests ii) and iii) described in the 'Ex post' subsection below, before undertaking the analysis. If the results is confirmed by both tests, they will clearly indicate the existence of systematic bias across TP versus non_TP schools, and therefore suggest the unreliability of the estimates. If they are not confirmed, the analysis will be undertaken and caveats will be discussed in the interpretation of results. It has to be noted that the test will not be able to disentangle the effect of TP on a tutored subject from the possibility that a teacher has a specific bias versus the pupil that can only be reflected in the subject he/she teaches him. It should also be noted that the tutoring program might potentially help students in other subjects if there are spillovers.

Evidence from the tests listed below will inform the interpretation of the results.

Ex post (comparing TP and non-TP schools):

i) To address uncertainties around whether the teacher assessed grades will reflect pupils' performance after the tutoring and whether the assessments would be sensitive enough to change (concerns 3 and 4), we plan to exploit dosage of tutoring. Assessment is an average of performance change, we can use dosage (as it exogenously depends on time of enrolling) on the sample of TP schools only. The assumption would be that time of enrolling is exogenous to performance. However, we point out that schools that enrol earlier may be more enthusiastic about the programme and have larger dosages. We will expect to see a bigger effect among those with larger dosage. This can be due to: a) dosage matters in improving ability; b) larger dosages reduces the dilution in teacher assessed grades who take a holistic approach. The strategy will at least point to the fact that there is some impact of TP.

ii) To address the concern that schools bumped up grades (concern 1), we can also test if the distribution of tests across the years (i.e. 2021 vs 2017/2018/2019) is different across TP/non TP schools for all pupils and PP pupils. Evidence of significant difference in distribution across TP and non TP schools may suggest the presence of measurement errors (i.e. bump-ups).
If TP and non-TP both bump up scores equally, then we would expect TP schools to have a slightly more positive impact because of TP (if TP is effective). If TP and non-TP schools bump up scores differently, then we would need to investigate whether the difference is related to bias (see checks (ii) in the ex ante section and (ii) and (iii) in the ex post section) or indication of a positive (or indeed negative) effect of TP.

iii) To address the concern that schools bumped up grades (concern 1), we will perform the analysis on Y11 pupils predicted to do TP. If the effect is positive, this can be due to the positive effect of TP or to bumping up of grades. We will perform the analysis on Y11 children predicted NOT to participate in TP. If there is an impact also on children not predicted to participate in TP, then it could be interpreted as evidence of bumping-up grades. However, we caveat for the fact

that this could also be due to the presence of spillovers or because of non-random selection of schools into treatment that are not fully controlled for in the methodological approach. If TP is effective, predicted TP should always have higher scores than predicted NO-TP even if there is bump up. The reliability of this test depends on how well we can predict participation to TP. The test cannot disentangle the increase in grade due to TP from a systematic teacher bias towards TP pupils only.

While we already know that that Concern 1 (bumping up grades) is very likely to be an issue, it is not indicative of systematic bias between TP and non-TP schools. The risk that the assessments are not sensitive enough to change will inform the interpretation of the results in case of no significant effect found.

We will be more concerned about the validity of these measures if some of the tests outlined above addressing concerns 2, 3 and 4 point toward the presence of systematic bias between TP and non-TP schools (specifically: ex ante test (ii); and ex post tests (ii) and (iii), see above). We highlight the fact that there could be more than one interpretation to these checks, that may not allow us to detect bias: i) the possibility of between subject spillovers; (ii) the heterogeneous effects of the pandemic itself across pupils and subjects. However, if all three of these tests point towards the presence of systematic bias, we would consider whether not to proceed with the empirical analysis.

We should note that in ex ante ii) and ex post ii) it is difficult to disentangle the impact of TP from the effect of bias as they both go in the same direction and they both affect the same population. We don't have any prior information about the distribution of bias and impact of TP that would help discriminate between the two. This is only possible in ex post iii) as we make use of pupils not predicted to participate. We unfortunately do not have the same type of control group in the other two tests.

Below we outline all of the originally planned analysis. However as explained above, the number of secondary schools recruited to the evaluation sample was lower than planned and therefore analysis for RQs 1a, 1c, 2 and 3 in the secondary school sample will no longer be pursued (instead see the mirror RQs 4a and 4b for the population analysis), and for RQs 1b and 5 the analysis will be carried out on the Year 11 population data rather than the secondary school evaluation sample. All the other RQs will be retained.

Outcome analysis (RQ1)

We will use two approaches: Weighting/regression and instrumental variable (IV) regression. We describe these in turn below.

WEIGHTING/REGRESSION (RQ1A) **What is the impact of TP availability on PP pupils' attainment?**

The sample of non-TP (comparison) schools identified using the matching-based sampling approach described above has proven difficult to recruit.

As outlined above in the 'Selection of the comparison group' section, the control group is therefore composed of schools selected according to different criteria: i) matched schools; ii)

schools that signed an Expression of Interest (EoI) in TP and iii) schools that delivered TP after the ex-post assessment.

To address this, we will construct weights to bring the achieved non-TP sample into line with the TP sample. We will show regression results with and without using the weights. Furthermore, working with the achieved sample, we can construct these weights using both pupil- and school-level variables. This allows us to more tightly control for differences between the TP and non-TP schools. In addition to the variables listed in Table 5 above, we will also include English as an additional language, special educational needs and interaction with social services (all pupil-level variables taken from the NPD). The procedure will be implemented once we have access to all data, including pupil level data, that can therefore be used in constructing the weights to adjust for imbalances in the sample.

We choose a weighting approach rather than a second matching stage because it helps addressing the fact the comparison sample has been selected according to three different criteria and we want to avoid any sample loss. We will construct weights using entropy balancing. This can be implemented in Stata using the ebalance ado file and the inverse probability weighting approach as a robustness check. It has the advantage of automating the process of balance checking and thereby reducing the scope for researcher bias. We emphasise that this weighting step will be carried out before outcomes are known. Since RQ1a focuses on PP pupils, the weights will be calculated only for the PP pupils.

We will report a comparison of the characteristics of TP schools in the evaluation sample with:

1) all non-TP schools in the evaluation sample;
2) those effectively recruited for the study as comparison schools;
3) those effectively recruited for the study as comparison schools, weighted.

To estimate impacts, we will regress the outcome on two measures of TP: i) a 0/1 indicator for TP being available and ii) a categorical variable measuring the fraction of hours completed by the time of the assessment (dosage). We will control for the baseline measure of the outcome of interest and year group indicator. All school level variables listed in Table 5 will be used as controls[33], including the test assessment provider. Pupil-level controls will include background variables, such as gender, ethnicity, EAL and special educational needs. Residuals will be clustered at the school level to account for any common school-specific unobservable component. Regression will be based on pupils in the TP schools and their matched comparators, using the weights derived using entropy balancing. The software used to run the model is Stata.

[33] We may need to review our approach to the school-level variables in NPD, once we know which/how many school-level variables we are able to match into the NPD. We understand that there are likely to be restrictions on school level data in NPD for the 2020/21 cohort due to the move to teacher assessed GCSEs and the removal of school accountability linked to the 2021 grades. See: https://www.gov.uk/government/news/extra-measures-to-support-students-ahead-of-next-summers-exams

The coefficient on the TP indicator will represent the estimated treatment effect, on an 'intention to treat basis'. For RQ2 and RQ3, the regression will be estimated on the sample of predicted TP pupils and all pupils, respectively.

The final analysis has moved from what we originally planned and, as a result, it is only a practical consideration that justifies the inclusion of schools in the comparison sample selected through different criteria. We will complement this analysis with the mitigations already planned as part of the analysis: RQ1b described below and RQ4 on Year 11 data. We also added mitigation analysis described below (RQ1c and RQ4b).

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

We will use instrumental variables (IV) techniques to provide estimates of TP that do not rely on the selection on observables assumption. This approach builds on the 'reach and engagement' RCTs which aim to test methods of increasing take up among pupils.

Two trials randomised tutors to interventions that aim to improve the relationship with the students and to improve pupil attendance: the first one leverages similarities between pupils and tutors; the second one improves tutors' relational self-efficacy. The third trial randomized pupils to an intervention that consists in weekly motivational messages. It targets only 5,000 pupils and, if effective, could be used in conjunction with the other two trials to achieve power.[34]

The trials will be delivered by the Summer with results shared with the EEF by October/November 2021 (TBC). Evidence from the trials will inform whether any or all of the randomised encouragements will induce sufficient additional take-up of TP for them to be effective instruments (as opposed to 'weak' instruments). If any of the tutor-randomised or pupil-randomised trials has a statistically significant impact on take-up of TP we will use it as an instrument for participation.

If any or all the trials are effective in boosting TP session attendance, we will use the RCT groups (treatment or control) as instruments in the instrumental variable regression of outcomes on participation. The resulting treatment-control difference in take-up is an experimental estimate of the impact on participation of the additional encouragement. We will use the randomised encouragement as the instrument for our IV estimator. In its simplest form, the IV estimator is simply the treatment-control difference in attainment divided by the treatment-control difference in participation. The resulting estimate is interpreted as the impact of TP participation among those induced to participate by the additional encouragement.

In practice, we will operationalise this by estimating a 2-stage least squares regression of the outcome on an indicator of whether the pupil participated in TP (instrumented using the reach randomised allocation) and the same variables used in the linear regression. Estimation will be based on pupils in TP schools only. The coefficient on the TP participation will be the impact

[34] Links to the study plans of these trials in the hyperlinks: Leveraging Similarity to Improve Pupil Attendance, Prioritising Tutoring Relationships to Improve Pupil Attendance, Engagement-Boosting Reminders to Improve Pupil Attendance and Engagement.

estimate and constitutes a local average treatment effect; the average impact among pupils induced to participate by the randomised encouragements. Note that the IV approach does not require that we focus just on Pupil Premium children. This follows from the fact that we perform the analysis using only TP schools, where eligibility can be observed. However, we will also estimate impacts of the subgroup of PP pupils.

There are two points to highlight about the IV estimator. First, it estimates the impact of participation for the subgroup of pupils who only participated because of the additional encouragement to do so. These estimates may be different from those for the eligible population as a whole. Nevertheless, they are of interest to the extent that they provide an insight of what the effect of extending the reach of TP might be. Second, the approach relies on the encouragements having an impact on participation. If this is not the case, IV estimates can be unreliable. This is the problem of weak instruments and we will test for it using the Montiel Olea-Pflueger (2013) approach, as implemented in the Stata program weakivtest.ado. If the null hypothesis for weak instrument is not rejected, we will not proceed with this analysis.

INSTRUMENTAL VARIABLES (RQ1C; RQ4B): WHAT IS THE IMPACT OF THE INTENSITY (DOSAGE) OF TP ON THE ATTAINMENT OF PP PUPILS**?**

The dosage analysis will be conducted on the sample of RC schools (RQ1c) and on Year 11 pupils (RQ4B);

We will use a second instrumental variables (IV) technique to provide estimates of TP that do not rely on the selection on observables assumption. This approach exploits the fact that some schools signed up to TP but have not yet delivered TP or have delivered it only partially at the time of the assessment. These schools are similar to TP Research Champion schools in terms of interest in the programme. We note that the timing of engagement in TP is non-random. We therefore provide supplementary evidence on this point by checking that prior characteristics of schools are not related to the timing of adoption amongst participants.

If a sufficiently high number of comparison schools are recruited from the sample of late TP sign up, the date of signing up the MoU may be positively associated with dosage and it can be used as instrument in the instrumental variable regression of outcomes on TP completion. In practice, we will use the number of days passed between the MoU and the time of assessment as an instrument for dosage.

The treatment-control difference in TP completion is an estimate of the impact on the intervention delivery of early sign up to the programme. The assumptions for the instrument to be valid are that i) the instrument is a significant predictor of the treatment (TP completion), 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 up to the MoU. As long as a mix of both high and low achieving schools signed up to the MoU earlier than others, the assumption could be plausible. As a check, we will estimate the two-stage least-squares regression (2SLS) with a placebo outcome such as KS2 scores.

We will use heterogeneity across schools on the date they sign up the MoU as instrument for our IV estimator.

Similarly to the previous IV analysis, we will estimate a 2SLS regression of the outcome on a categorical variable of the dosage of TP received (instrumented using the date the MoU was signed with the TP) and the same variables used in the linear regression. Estimation will be based on pupils in TP schools only, exploiting the fact that some TP schools will have not yet started or completed the delivery of TP at the time of the assessment. The coefficient on the dosage will be 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. We will estimate this using all pupils in year groups doing TP and look at the impact on PP pupils only. The reliability of this analysis is subject to the quality and completeness of the data received by TP in relation to the time they delivered the sessions. If we cannot assess with precision when the sessions were delivered in relation the time of the assessment for the majority of TP, it will not be possible to define the dosage and therefore to run this analysis.

As above, the local average treatment effect may be different from the impact on the eligible population as a whole, as estimates will identify the impact of the programme on schools that signed an MoU early. Second, the risk of weak instrument will be tested using the Montiel Olea-Pflueger (2013) approach. If the null hypothesis for weak instrument is not rejected, we will not proceed with this analysis. Third, the analysis assumes we can observe the precise number of sessions delivered before the teacher assessed grades are submitted. We will use a cut-off date for the teacher assessed grades submission, with the risk of some measurement errors in the exact amount of TP delivered before the assessment.

Inference

Uncertainty will be conveyed using confidence intervals. We note that with the linear regression impact analysis, no account is taken of uncertainty arising from the matching and weighting, which we regard as pre-processing steps. We will adjust for multiple testing using the simulation approach of Westfall-Young (1993). Full estimation results (including standard errors) will be given in an appendix.

Further analyses

These analyses can be performed with any of the estimation methods proposed in the previous section.

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

Our approach to the primary analysis provides an estimate of the impact on a subgroup of the eligible population, PP pupils, which may not coincide with the group of children who will receive the intervention. RQ2 involves an alternative approach to approximating the eligible group. It involves modelling the probability of pupil participation in TP 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). The results will be used to predict participation in both TP and non-TP schools in the evaluation sample. Having done this, we will follow a similar approach to

that of RQ1a but instead of selecting PP pupils, we will instead select predicted TP pupils (where, in non-TP schools, predicted TP pupils are those who would be predicted to participate were TP available). Impacts will be estimated as (weighted regression-adjusted) comparison of outcomes among predicted eligible pupils in TP and matched non-TP schools. Like the approach for RQ1a, we will regress the outcome on the two measures of TP, the 0/1 indicator for TP being available and the categorical variable measuring the fraction of hours completed by the time of the assessment (dosage).

The two strategies we proposed - the estimation on PP only (RQ1a) and the predicted TP estimation (RQ2) - counterbalance themselves. If all pupils receiving TP are PP pupils, then the PP selection would perfectly allow to estimate the impact of TP. If not, the predicted TP estimate is a natural stabilizer.

There is a risk that the proportion of PP pupils that receive TP is low and we are unable to predict who receive TP well. According to the population of TP data collected up to March 2021, 27% of PP pupils are receiving TP. As we discuss in the preliminary analysis section, we will finalise the estimation approach (random effects, including fixed effects in ebalance) to be sufficiently sensitive to a low fraction of PP pupils being targeted in the intervention. In the population sample, we will provide a sensitivity analysis that restricts the analysis to the sample of TP schools that targeted a majority of PP pupils for tuition.

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

As another means of understanding the overall effect of TP, a third analysis will focus on attainment of all pupils (rather than PP pupils or predicted TP pupils) in year groups with TP in TP schools compared with all pupils in comparison year groups in non-TP schools in the evaluation sample. These estimates are likely to be smaller than RQ1a and RQ2 estimates, as the TP impact will be more diluted. This estimator also captures the effect of spill over (peer) effects. The purpose of this is to capture the overall impact of TP. The regression analysis will control for the same school level and pupil level characteristics mentioned in the primary analysis. Like the approach for RQ1a and RQ2, we will regress the outcome on the two measures of TP, the 0/1 indicator for TP being available and the categorical variable measuring the fraction of hours completed by the time of the assessment (dosage).

What is the impact of TP availability on pupils' attainment in the population of schools? (RQ4a)

RQ4a is an additional analysis that uses the pupils in the full population of secondary schools as observed in the NPD. It will be limited to Year 11 if, as we anticipate, the results of teacher assessed GCSEs will be available on the NPD. We would use NPD baselines (i.e. KS2 data). Other than this, it would apply a similar overall matching/ weighting/ regression approach to that described above (RQ1a), except matching would only need to be carried out once since the issue on failure to recruit comparison schools would not arise. We will perform this analysis for PP pupils (RQ4a1), pupils predicted to participate in TP (RQ4a2) and all pupils (RQ4a3), thereby mirroring RQ1a, RQ2 and RQ3 (although in secondary schools only, due to only Year 11 and not Year 6 data being available in the NPD). The analysis is complementary to the evaluation sample analysis and it limits forms of selection that may take place in the evaluation

sample, where Research Champion schools may not be representative of the sample of TP schools.

An additional dimension of this analysis is that it becomes possible to bootstrap the entire matching/weighting/regression process in order to achieve standard errors that capture the full uncertainty of the estimation approach. Comparing these standard errors to those that ignore the uncertainty arising from the matching/weighting steps provides a means of assessing whether the level of uncertainty in our main estimates is adequately captured.

For RQ4b, see the earlier section on instrumental variables (RQ1c; RQ4b).

Moderator analysis: How does the impact of TP availability vary among PP pupils, by school and pupil characteristics? (RQ5)

Moderator analysis will be conducted through interaction terms on the following categories[35] of:

1. School characteristics[36]: Ofsted rating (high vs. low); proportion of FSM (high vs. low); type of school (academy/maintained); school size (by quartile). These variables identify the context where TP is delivered and allow to analyse whether TP has been more effective in disadvantaged contexts. They refer to the 'mobilise' (engaging schools) phase of the logic model.
2. Pupil characteristics: prior attainment; pupil premium (eligible vs. not); SEND vs. not; KS or age; school attendance; English as an additional language, ethnicity and gender. These variables identify participants and allow to analyse whether TP has been more effective on disadvantaged pupils or children with specific demographics. They refer to the 'mobilise' (engaging pupils) phase of the logic model.
3. Other: geography (urban/rural; low/high IDACI): different geographical areas may have different provision of TP in terms of quantity and quality. If schools in more deprived areas have lower attainment outcomes, this may be correlated with the supply of TP in those areas, and the risk that low-quality tutors reach more disadvantaged schools (development and mobilise – activating tutors phase of the logic model).

Estimates are based on the schools in the evaluation sample for primary schools, and the Year 11 population for secondary schools.

Since our estimation approach focuses on PP pupils and predicted TP pupils, the issue of compliance does not arise.

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

A descriptive analysis (using the data collected via templates for the above impact analysis) will compare outcomes associated with different tutoring models and moderators among TP

[35] In the first version of the study plan (prior to the national lockdown) we also proposed a Covid-19 hotspot moderator thinking in terms of the Tier 3/Tier 4 restrictions being brought in and out by area, however then all areas entered the same restrictions in January 2021. Furthermore we have not been able to identify a dataset that would allow us to define hotspot easily.

[36] As highlighted in a previous footnote, we may need to review/amend our approach to the school-level variables in NPD, once we know which/how many school-level variables we are able to match into the NPD.

schools in the evaluation sample. We do not propose any impact analysis within RQ6 since we cannot observe the counterfactual treatment model among non-TP schools. Instead, this element of the analysis will summarise mean attainment among participating pupils in TP schools according to the model of tutoring they experience. We will regress attainment on the variables listed below for the sample of TP schools only to assess heterogeneity. In particular, we will look at:

1. The intervention: mode of delivery of completed sessions (online vs. face to face); timing of the session (during vs. after lessons); tutor:pupil ratio (1:1 vs 1:2 vs 1:3); number of blocks schools choose (low/high buy-in schools); intensity of delivery (determined by sessions attended/number of weeks tutoring is spread over); completed versus scheduled sessions.
2. Tutors: Experience/qualifications; TP tutor training; shared characteristics with pupil/tutee (gender, ethnicity).
3. Other: early/late delivery.

Missing data

The section describes how we deal with missing values at follow up. The key issue is whether missing data differs systematically between TP and control schools. We would not expect to find missing values at school level, as we are recruiting schools with a testing regime in place. If a school that drops from the programme is using the online repository of an assessment provider or is providing tutoring to Year 6/Year 11, then we could still explore assessment outcomes, unless the school also withdraws from the evaluation and requests that the data is not used. If the pupil drops out from TP, we would observe their assessment anyway. We would miss the observation for students missing the test, and it would be important to understand why the child missed the test.

We expect 15-20% of students missing the test, taking into account the possibility of children being sick or isolated on the day of the test. We will explore the extent of missingness by counting the observations for which the assessment variables are missing, and the pattern of missingness in the outcome variables and by TP participation status.

To explore the pattern of missingness, we will run a logistic regression on the probability of dropping out at follow up which includes TP participation status, individual observable background characteristics and the baseline outcomes.

The logistic model will provide evidence on whether dropping out is correlated with TP participation, observable characteristics and baseline testing. The possibility that there are statistically significant differences across TP and control schools is limited by the fact that all schools recruited in the study have a testing regime in place with one of the assessment providers.

The student could miss the test because of sickness or isolation, in which case baseline testing would not be correlated with the probability of drop out. We would then safely assume that the observation is missing completely at random. Focusing only on the sample of complete cases would not bias the estimates. If missingness is correlated with having achieved a low score at

baseline and this factor is associated with substantive model outcome, then controlling for baseline attainment should address the issue. Missing data will be 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 are replaced by the estimated plausible values by the estimation of multiple datasets. The results obtained from each dataset are combined using Rubin's rules to create a "complete" dataset (Schafer, 1999). Results with MI will be reported in addition to the headline impact estimates.

Effect size calculation

Estimates will be presented as effect sizes, calculated using the Hedges' g formula. Formally, the effect sizes are calculated as follows:

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

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

Implementation and process evaluation (IPE) [37]

The IPE has been designed to cover the distinct phases of programme implementation, multiple stakeholder groups, and scale and complexity of the programme. The TP programme involves multiple phases and participants, considerable scale, and at least some of the ways of operating are not prescribed[38], meaning potential for variation in approach and success. This is an opportunity for the IPE to examine what works, and why, in multiple contexts.

It has also been designed to dovetail with programme delivery, providing iterative, responsive, and formative insights. The IPE will examine the implementation of Tuition Partners and the implications of this for the programme theory, design, and effects. The IPE will complement the impact evaluation through analysis to help contextualise the impact findings and through informing the moderators analysis. The IPE includes 'review and next steps' points for the evidence to be fed back into the programme and evaluation.

[37] Designed to follow the principles detailed in the Implementation and Process Evaluation Guidance (2019).
[38] And some processes – especially for later phases – are under development at the point of writing.

As part of this, the IPE adopts a flexible approach to fit against the staggered school recruitment. The IPE approach will be flexible to emerging programme roll-out and practices, and reflect findings from initial scoping work with schools to assess feasibility of classroom teacher and pupil IPE data collection.

The IPE is also designed with the implications of Covid-19 in mind. We anticipate that the evaluation of TP will be subject to restrictions and challenges from Covid-19, and as such we have built in from the start use of remote methods for qualitative research (such as video and telephone interviews) and online quantitative methods that are naturally resilient, and ensure the safety of our staff and participants, and the integrity of the data we collect.

The IPE activities are outlined under *Research Methods* below.

Research questions

The primary objective of the IPE is to examine the implementation of the programme against the programme design to help understand what happened, why, and the implications of this for the programme effects.

Implementation RQ: How was TP implemented and what are the implications for the programme theory, design, and effects? (this will be investigated through a number of qualitative and quantitative research activities with programme participants and beneficiaries across the five phases of the programme).

RQ7: How well has the programme been implemented? *[Implementation]*

- What approaches have those delivering the programme adopted at each phase, and why?
- For each phase, how have actual activities matched to or differed from the programme design?
- What are the experiences of those delivering the programme at each phase?
- What were the key barriers and facilitators of successful implementation at each phase? How can/are these barriers being overcome?
- What implications do these experiences have for the:
    - Programme theory?
    - Programme design?
    - Programme effects?

RQ8: To what extent has the programme both reached and engaged disadvantaged schools and pupils? Why/why not? *[Reach]*

- What is the profile of schools and pupils receiving tutoring as part of the programme?
    - How many schools/pupils has it reached?
    - To what extent is the programme reaching the target disadvantaged schools/pupils?
    - What proportion of TP schools/pupils have high PP?
    - To what extent is the programme reaching pupils with SEND?
- To what extent are pupils completing their allotted tuition?
- How has the design of the programme supported or hindered reaching and engaging with disadvantaged schools and pupils?

- What were the key barriers and facilitators of reaching the target numbers and profile of schools and pupils? How can/are these barriers being overcome?
- What implications are there for the intended effects of the programme if reach aspirations have not been met?

RQ9: How well has the programme delivered high quality tutoring? *[High Quality Tutoring]*

- What are the programme processes and activities designed to help achieve High Quality Tutoring?
- Have those processes been implemented as expected? (Why/why not?)
- How well are the core elements of high quality tutoring (dosage, focus, experience) being delivered?
- What are the key barriers and facilitators of delivering high quality tutoring? How can/are barriers being overcome?
- How has the design of the programme supported or hindered delivering high quality tutoring?
- What implications are there for the intended effects of the programme if high quality tutoring has not been (fully) delivered?

RQ10: What is the perceived impact of the programme? *[Impact]*

- How has the programme performed against original expectations?
- What have been the barriers and facilitators of success?
- What recommendations would programme participants make for future iterations of the programme?

RQ11: What factors (moderators) influence (or are perceived to influence) outcomes? *[Moderators]*

- Which of the predetermined moderators are most important, and why?
- What other moderators are there?
- What implications are there for the intended effects of the programme?

These research questions are supplemented by a set of sub-research questions – specific lines of enquiry - for each phase of implementation and research audience. These are documented in the *IPE Research Questions Matrix* below (Table 6), though this matrix is iterative and additional lines of enquiry may be added as the IPE progresses.

Table 6: IPE research questions matrix

Programme phase	**Design**	**Develop**	**Mobilise**	**Deliver**	**Legacy**
IPE analytical/learning objectives covered in each phase	Process (fidelity / quality)	Process (fidelity / quality)	Process (fidelity / quality); Reach and engagement (incl. responsiveness); Moderators; Barriers / facilitators	Process (fidelity / quality); Reach and engagement (incl. responsiveness); Moderators; Barriers / facilitators; Programme differentiation; Perceived impact	Process (fidelity / quality); Reach and engagement (incl. responsiveness); Moderators; Barriers / facilitators; Perceived impact
Cross cutting dimensions	Logic / theory review; Cost evaluation; Formative findings / improvement recommendations				
Common RQs for each phase	What was expected (focus on key expectations/risks)? What happened and why (focus on facilitators / barriers to implementation)?				
Programme Managers lines of enquiry	Experiences of establishing the evidential basis for the TP programme Experiences of collaborating with DfE to agree funding settlement Establishing governance structures	Activities to develop intervention supply (sector engagement, grant agreement process) Activities to develop intervention demand (sector engagement) Activities to establish key concepts and tools, including scoring criteria for high quality and best practice guidance.	Experiences of activating TPs (incl. using best practice guidance, capacity building and support approaches, tools, and delivery by Nesta/Impetus and experienced otherwise) Activities to engage schools (reach and engagement research, information events) Activities to match delivery and need	Ongoing capacity building activities Experiences of monitoring TP delivery (gathering data)	Perceptions of TP programme (incl. extent to which it meets PM expectations) Views about the programme's contribution to the evidence base on the effectiveness of tutoring Perceptions of long-term sustainability of systems and effects
Tuition Partners lines of enquiry	n/a	Experiences of applying to the TP programme (incl. motivation, expectations, facilitators/barriers) Expectations of the TP programme (incl. perceptions, understanding)	Activities to reach quality / scale requirements specified in grant agreement – including tutor recruitment, briefing and training	Experiences of delivery (school take up, school facilitation) Experiences of monitoring attendance and quality	Perceived sustainability of TP offer with disadvantaged schools and pupils - plans for the future Perceptions of change to capacity/quality of tutoring

Programme phase	**Design**	**Develop**	**Mobilise**	**Deliver**	**Legacy**
			Perceptions of capacity building support (Nesta/Impetus) Perceptions of EEF support to reach scale requirements and deliver high quality tutoring Experiences of engaging schools Experiences of tutor-pupil matching	Perceptions of ongoing capacity building support (Nesta/Impetus) Perceptions of delivery – what is working well / less well; whether meeting expectations; suggestions for improvements	Extent to which TPs have built networks / peer support Perceptions of TP programme (incl. extent to which it meets TP expectations)
School leads lines of enquiry	n/a	n/a	Expectations of the TP programme (incl. perceptions, understanding) Reasons for taking part Experiences of engaging TPs Experiences of identifying (and potentially matching) pupils	Experiences of delivery (pupil take up, scheduling, equipment, supervision) Role in monitoring pupil / parent feedback Perceptions of quality Perceptions of delivery – what is working well / less well; whether meeting expectations; suggestions for improvements	Perceived impact on pupil outcomes (cognitive attainment, other) Perceptions of tutoring, and likelihood of future use Perceptions of TP programme (incl. extent to which it meets school expectations)
Tutors lines of enquiry	n/a	n/a	Expectations of the TP programme (incl. perceptions, understanding) Reasons for taking part Experiences of working with TP(s) (incl. recruitment, training/briefing)	Experiences of delivery (incl. school facilitation, pupil attendance, channel and format of delivery, session quality) Experiences of monitoring quality and attendance (and	Perceived impact on pupil outcomes (cognitive attainment, other) Perceived impact of TP programme on future plans

Programme phase	**Design**	**Develop**	**Mobilise**	**Deliver**	**Legacy**
			Experiences and views of quality processes Role (where relevant) in TP-school engagement and tutor-pupil matching	ongoing engagement with TP in this area) Perceptions of pupil (and parent) engagement with tutoring Perceptions of how tutoring aligns with classroom teaching	Perceptions of TP programme (incl. extent to which it meets tutor expectations)
Classroom teachers lines of enquiry	n/a	n/a	Expectations of the TP programme (incl. perceptions, understanding) Experiences (where relevant) of identifying (and potentially matching) pupils	Experiences of delivery (pupil take up, scheduling, equipment, supervision) Role in monitoring pupil / parent feedback Perceptions of quality (incl. tutor matching, subject, monitoring / feedback, alignment with classroom teaching) Perceptions of delivery – what is working well / less well; whether meeting expectations; suggestions for improvements	Perceived impact on pupil outcomes (cognitive attainment, other) Perceptions of tutoring, and likelihood of future use Perceptions of TP programme (incl. extent to which it meets teacher expectations)
Pupils lines of enquiry	n/a	n/a	Expectations of the TP programme (incl. perceptions, understanding) Reasons for taking part Experiences of signing up to the programme	Experiences of delivery (incl. relationship with tutor, quality, monitoring / feedback, alignment with classroom teaching) Perceptions of delivery – what is working well / less well;	Perceived impact on pupil outcomes (cognitive attainment, other) Perceptions of tutoring, and likelihood of future use

Programme phase	Design	Develop	Mobilise	Deliver	Legacy
				whether meeting expectations; suggestions for improvements	Perceptions of TP programme (incl. extent to which it meets pupil expectations)

Research methods

The IPE design allows for an iterative approach and regular feedback loops. It will involve research activities with the following research audiences (programme stakeholders & participants)[39]:

- EEF/Programme Managers
- Tuition Partners
- Schools (head teachers or nominated school leader/leads)
- Classroom teachers
- Tutors
- Pupils

All research design and delivery will be carried out by Kantar – as part of the Evaluator consortium. (Surveys will be administered through NFER's online system, Questback.) In deciding the sample selection for each wave of research activities, we will take account of a potentially significant degree of variation and number of important characteristics, including the moderators identified in the ITT.

Surveys will be conducted on a census basis – the online approach is cost effective and so facilitates broad coverage of each stakeholder group. We will encourage good response rates by using a series of reminders; and a multi-pronged approach involving tranches/waves so that participants have several opportunities to participate/share their views; as well as publicising the surveys via headteachers, the TP newsletters etc.

Qualitative sampling will be purposive, with an intention to be reflective of the profile of the programme stakeholders across the course of the IPE.

Development of data collection instruments will be led by Kantar, with input and review from NFER, University of Westminster, and EEF. Qualitative materials will be reviewed following the initial depth interviews or groups to explore if changes are needed. Data from the initial first few days of the online surveys will be reviewed as part of standard quality assurance processes, with a review of responses conducted to identify if any changes are needed for the remaining fieldwork period.

Table 7 below provides a detailed breakdown of the research audiences, methods, sampling criteria, timings, and data uses across the IPE.

As noted earlier, the IPE, in particular the qualitative research elements, has been designed to be iterative in response to emerging findings and priorities. The three waves of research activities allow for coverage of each of the five phases of programme delivery and, combined with the intended sample sizes for each audience, coverage of the broad range of characteristics and factors of interest for both the IPE and impact evaluation. The design and scale of the IPE also provides some ability to overcome the challenges that the current COVID-19 situation may present. The evaluation must be mindful of the issues schools and pupils face, and the burden on them, plus the practical impacts of COVID-19 restrictions or school closures. The design of the IPE (and evaluation generally) must, therefore, remain

[39] Note that the IPE will not involve research with comparison schools who are not participating in the TP programme.

flexible and respond to these circumstances. This may, for example, involve reconfiguring the timing and numbers of interviews within a term, shifting resources to focus on specific audiences (or not), or giving schools flexibility in relation to their participation.

Version 2 note: the IPE design was resilient to the January 2021 lockdown. The three wave approach, which is aligned to the programme roll out, was adjusted to the changing context – for example, to capture experiences of online at-home delivery. The rate of completion of interviews with schools slowed during January/February 2021 as fewer tuition sessions took place. Wave 1 fieldwork was extended. And as a result of the reduced timeframe for delivery, there was some overlap in timing between waves.

Table 7: IPE methods overview (the numbers quoted are as per the original plan) (achieved numbers will be reported)

Research audience (programme stakeholders & participants)	Data collection methods	Sampling, timings and volumes	Research purpose
EEF/Programme Managers	▪ In-depth interviews ▪ Meetings & workshops ▪ Half-termly IPE review workshops	Ongoing flexible engagement with PMs (EEF, Impetus, Nesta) across the course of the evaluation.	To build our understanding of the programme, refine the logic model and research questions, finalise the evaluation plan, and evaluate the early implementation phases of the logic model. These activities are also part of processes to share programme updates and evaluation findings.
Tuition Partners	▪ In-depth interviews	We plan to conduct 60 qualitative interviews (c.20 in each in wave) We will aim to speak to all TPs at least once, plus repeated interviews sampled to cover a diversity of implementation fidelity according to MI data.	To gather feedback on all activities and implementation progress, what is working and not, what makes for successful tutoring, and suggestions for improvements.
Tutors	▪ In-depth interviews ▪ Online focus groups ▪ Online survey	We plan to conduct 90 qualitative interviews (c.30 per wave), plus 10 online focus groups (c.5 each in wave 1 and 3). Sampling will be purposive at each wave to cover key areas or dynamics of interest, based on MI data; for example TP ranking, highest qualification, employment status, number of pupils. We will also conduct two online surveys (in waves 1 and 3) to capture earlier programme processes, and perceptions of barriers, benefits and enablers to impact respectively.	We will capture tutor experiences of each of the programme phases and processes (e.g. training, matching, delivery); their perspectives on the role of providers (TPs), schools and teachers, in helping tutors fulfil their role, and their views on perceived benefits to pupils' learning and other outcomes. In addition, tutors' views on optimum group sizes, how they have aligned tutoring to pupils' needs, and their own plans/interest in pursuing a teaching career, will also be explored.

Research audience (programme stakeholders & participants)	Data collection methods	Sampling, timings and volumes	Research purpose
Schools (headteacher or other nominated leaders)	▪ In-depth interviews ▪ Online survey	We plan to conduct c.105 qualitative interviews with school leaders. This includes an initial round of 15 scoping interviews with headteachers early in the project and three waves of c.30 interviews per wave. Sampling will be purposive at each wave to cover key areas or dynamics of interest, based on MI data; for example, education stage, type of provision, school PP/FSM profile, Ofsted rating, number of participating pupils, location. We will also invite all schools participating in the programme to complete two online surveys (in waves 1 and 3 – each with multiple batches) to capture earlier programme processes, and perceptions of barriers, benefits and enablers to impact respectively.	Initial interviews with schools will explore the feasibility of other research activities, especially the participation of classroom teachers and pupils. The main waves of in-depth interviews will capture school views on each phase of the programme, including how successfully each has been implemented, their views on barriers/facilitators, and their suggestions for improvements. The survey will cover feedback on early programme processes and, later in the programme, capture views on ongoing implementation and perceived benefits.
Classroom teachers	▪ In-depth interviews ▪ Online survey	We plan to conduct 90 qualitative interviews (c.45 each in waves 2 and 3). Sampling will be purposive at each wave to cover key areas or dynamics of interest, based on MI data; for example,	To understand perspective on the implementation and success of the programme – both to perceived quality of tutoring and how tutors are integrating the classroom curriculum and benefiting pupils.

Research audience (programme stakeholders & participants)	Data collection methods	Sampling, timings and volumes	Research purpose
		subject specialism, education stage, type of provision, school PP/FSM profile, Ofsted rating, number of participating pupils, location. We will also invite all classroom teachers with participating pupils (contacted via school leaders) to complete an online survey in the summer term (wave 3).	Given the potential variance in teacher contextual factors (e.g. type of school and pupils, prior experiences of tutoring, location), the online survey has been included to capture experiences and views on implementation and programme effects at a broader level (open to all teachers eligible to participate).
Pupils	▪ Online discussion groups[40] ▪ MI data analysis	We plan to conduct 60 online focus groups with pupils in receipt of tuition (c.30 each in waves 2 and 3) to ensure coverage of a variety of characteristics and dynamics. Sampling will be purposive at each wave to cover key areas or dynamics of interest, based on MI data; for example, education stage, gender, subject, specialism, type of provision,	We will capture pupils' experiences of tutoring and their views and how well this approach has helped them with their learning. Pupils will provide important feedback on the quality of tutoring, preferences for approach/mode (e.g. 3:1 vs 1:1, online vs face to face), practical issues such as access to technology and responses to COVID-9 (e.g. self-isolation), and their suggestions for how to maximise the suitability and effectiveness of provision.

[40] Given the restrictions posed by Covid-19 we anticipate the usual routes to pupil participation will not be available to us. There are also important safeguarding considerations when conducting research with children online. We do not envisage conducting a survey of pupils to avoid overburdening schools. We will consult with headteachers to scope the feasibility and fieldwork protocols for online pupil discussion groups. We will also use provider (TP) data collection channels to capture their experiences and feedback.

Research audience (programme stakeholders & participants)	Data collection methods	Sampling, timings and volumes	Research purpose
		school PP/FSM profile, Ofsted rating, number of participating pupils, location. We do not envisage further primary research with pupils not in receipt of tuition. However, we can analyse any MI data provided by schools pertaining to this group.	

IPE Analysis

The IPE will predominantly involve qualitative research but will also incorporate analysis of quantitative data from surveys and programme management information too. To answer the IPE research questions, we will triangulate the evidence from the qualitative research, surveys, and MI across the programme phases and stakeholders, alongside the impact evaluation findings.

One of the challenges for the IPE will be controlling and focussing the analysis. Given the scale and complexity of the programme, including the five phases of delivery and multiple stakeholder groups, and the number of overarching research questions, analysis could very quickly become unwieldy.

There are also numerous objectives and potentially competing frameworks to conduct analysis within, meaning the IPE analysis could become overly complicated and potentially ungoverned as analysis seeks to fit within multiple structures at the same time.

To address these challenges, the IPE will take the following system/approaches:

- **The primary framework for all analysis will be the overarching research questions –** ensuring that it links directly to the main objectives of the IPE.
- **The secondary analysis framework will be the IPE research question matrix** – to further focus analysis on the predetermined lines of enquiry, researchers will map their analysis to the research question matrix as part of populating our analysis database.
- **The IPE analysis will be guided by the following approaches:**
    - **Realist evaluation[41] -** for understanding *what works*, *for whom*, *why*, and *in what circumstances*.
    - **Contribution analysis[42]** - for assessing and inferring causality in programme evaluations. It explores a wide range of data sources to interrogate the underlying assumptions of the logic model and to trace observed outcomes back to interventions in a step-by-step process.
- **The IPE analysis will cover a further set of analytical/learning objectives –** which are linked to the overarching research questions:
    - Process (fidelity/quality): monitoring actions, processes, and systems (including assessing against the programme design/intended approaches), identifying best practice, and understanding how well implementation is working.
    - Reach and engagement (including responsiveness): exploring take up (including barriers and facilitators).
    - Moderators and other contextual factors: exploring the role and influence of a range of moderators and factors that might be expected to affect take-up, engagement and pupil outcomes.

[41] Based on the work of Tilley and Pawson (1997), Realist evaluation emphasises the contextual conditions, the precise mechanisms of change, and the desired / observed outcome patterns of a programme to be evaluated (C+M=O formula).
[42] https://www.betterevaluation.org/en/plan/approach/contribution_analysis

- Barriers and facilitators: identifying barriers to implementation, and to programme outcomes and impacts.
- Perceived benefits: perspectives on enablers and barriers to pupils' learning and wider outcomes.
- Mechanisms: identifying and helping to understand causal effects (or lack of) within the programme, including implications for mid- to long-term effectiveness.
- Formative findings and suggested improvements: asking participants to identify key events, issues, or actors within the programme that should be followed up on or drawn out in our analysis, and any recommendations for the design and implementation of TP.
- Cost evaluation: helping to collect data and information to assess the cost of delivery (including unintended costs) and value of the programme.

- **Analysis will be conducted using Kantar's systematic framework analysis approach –** to give structure not just to how analysis is broken down, but also the process followed (see below)
- **Analysis will also include open, iterative phases –** involving group brainstorms and findings mapping, where we identify features and patterns within the data, mapping the range and nature of data, finding associations, defining concepts, creating typologies, and undertaking sub-group analysis.

While the period post-fieldwork is often where we will have the greatest emphasis on analysis, our researchers will conduct analysis on an ongoing basis through independent and joint review of all material collected. This way we can examine existing hypotheses and assumptions, and develop, test and refine new hypotheses over the course of the project. This also means we are better able to provide early sight of thinking and evidence.

Our analytical process is systematic and includes a content analysis method known as framework analysis – an analytical process that is both flexible and systematic. It involves constructing a thematic framework against which data is synthesised and then mapped to identify features and patterns: defining concepts, mapping the range and nature of phenomena, creating typologies, finding associations, and providing explanations. This is followed by a process of weighing up the salience and dynamics of issues and searching for structures within the data that have explanatory power, rather than simply seeking a weight of evidence.

We will analyse the IPE data thematically and inductively, building up our analysis to address the main research questions. We will conduct "cell" level analysis of lines of enquiry[43] within phases of activity by stakeholder group, for example looking at the evidence from Tutors in relation to early programme processes, or classroom teachers on moderators[44]. We will also conduct thematic analysis across phases and stakeholders (for example looking across the stakeholder groups at the role communication from programme partners and other organisations has played in facilitating delivery).

We will also analyse the data deductively in relation to the logic model, including factors that may influence the strength of the relationship between the intervention and the outcome. This may include adaptability, tailoring/alignment with the school curriculum, and any

[43] As mapped within the Research Question Matrix.
[44] Or where necessary drilling down into sub-cells.

mediators for engaging disadvantaged children. The contribution analysis approach will be particularly useful here, given its emphasis on the reasoned interpretation of evidence from multiple sources and use of multiple perspectives, including external experts and those involved in the programme, in this process.

In order to provide formative feedback on the implementation of the TP programme as it progresses – what happened and why, barriers and facilitators to achieving the intended programme outcomes, and implications for ongoing design and effectiveness – we will take an iterative approach to analysis. This reflects our approach to fieldwork and involves our researchers conducting analysis on an ongoing basis through independent and joint review of all material collected (survey data, qualitative interview notes, audio files, video, documentary evidence).

To support the IPE (and wider evaluation), a programme Theory of Change, phase and sub-phase process depictions, and phase level logic models have been created to capture the programme objectives, processes, intended effects, and mechanisms for change.[45] Extracts from these have been presented earlier in this study plan, with the full "Logic Model" presentation accompanying this plan. The IPE evidence will be used to refine the Theory of Change and Logic Model as the evaluation progresses.

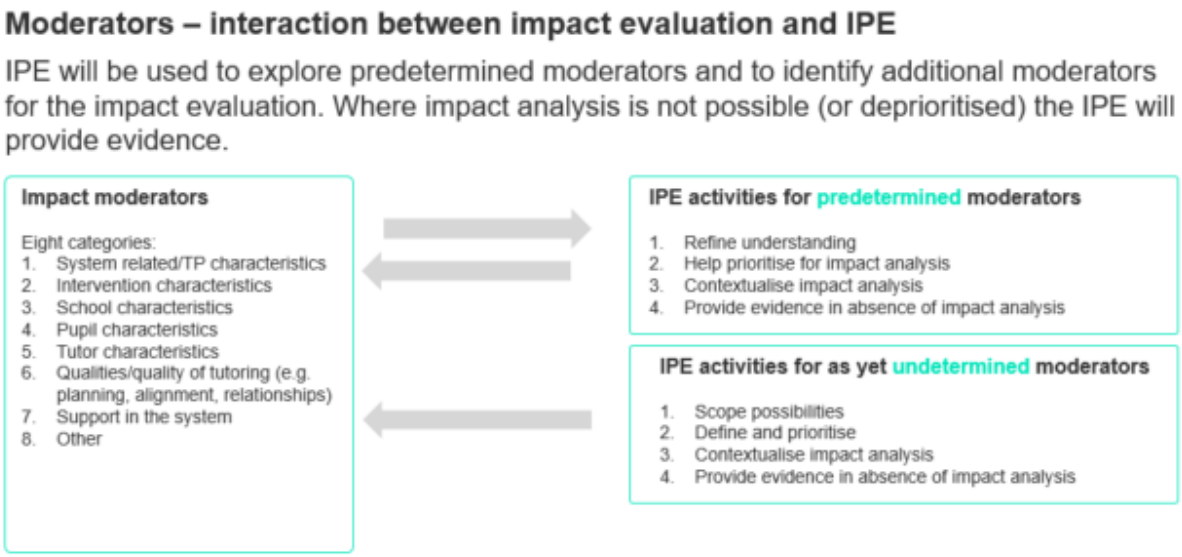
The analysis will also help refine the moderators to be used in the impact evaluation. There are a number of potential factors that may materially affect the extent to which the intended programme impacts are achieved, either directly or through influencing the implementation of the programme. These moderators can be organised into several categories:

1. **System-related TP level – e.g. TP quality ranking, TP specialism (e.g. SEND), other organisational characteristics**
2. **Intervention characteristics –**the mode of tuition (online, in-person, mixed), group size (one-to-one, small group), the subject
3. **School characteristics –** phase (primary/secondary), proportion of pupils eligible for free school meals, Ofsted rating
4. **Pupil characteristics –** ethnicity, gender, prior attainment, attendance, SEND
5. **Tutor characteristics –** similarly, demographic characteristics, but also professional, such as prior experience tutoring, highest qualification
6. **Qualities/quality of tutoring –** including elements such as planning and delivery of tutoring, alignment with pupils' needs, and also dynamics between tutors and pupils
7. **Support in the system –** guidance, training, ongoing support for Tuition Partners and tutors, plus schools
8. **Other –** including effects from COVID (e.g. localised lockdowns causing disruption)

The impact evaluation will conduct analysis of the extent to which these factors influence the outcomes observed. The IPE will be used to explore predetermined moderators and to identify additional moderators for the impact evaluation. Where impact analysis is not possible (or deprioritised) the IPE will provide evidence.

[45] These have been informed by existing programme documentation and scoping work with EEF, Nesta, and Impetus.

Figure 5: interaction between impact evaluation and IPE to explore and analyse moderators



Overall, the IPE will provide an in-depth account and assessment of the implementation of the intervention in addition to perceptions of its impact. We will produce a headline presentation for each half-termly review workshop with Programme Managers that clearly articulates the programme, research to date, emerging (formative) findings, and recommendations for programme and evaluation. These findings will be incorporated into formal reporting outputs at the end of each term and integrated into the overall evaluation outputs. Results from the IPE will be organised by the five research question topics (implementation, reach, high quality tutoring, impact, and moderators).

Cost evaluation

Approach to cost evaluation

Our general approach is to gather more data where costs are likely to be variable. As NTP TP brings together various providers each delivering their own model of tuition, we may expect more variability in the costs. The primary cost estimates will be on the per pupil cost to schools for participating in TP (noting that there will be variation, depending on the nature of the TP provision, so we may need to provide some ranges in cost data – see cost sensitivity analysis below). The market value of the programme will also be estimated. The main source for cost data for a school is the school survey which will ask head teachers/TP school leads to report the amount of time and money spent on different aspects of TP.

We will collect data relating to any monetary costs paid by schools particularly in the 'mobilise' and 'delivery' phases of TP (see IPE section above), as well as opportunity costs (as measured by time spent on TP activities). Costs associated with the mobilise phase are staff time spent researching TPs, engaging with the programme and TP organisation(s), and identifying pupils to participate. The delivery phase costs can be separated into set-up costs (such as purchasing additional equipment needed, arranging rooms, writing communication to parents) and ongoing costs (organising and supervising sessions, incentives for pupils to attend, IT support). These are summarised in Table 8.

In terms of considering cost data on activities undertaken in schools that are not part of the NTP TP, we plan to collect data from comparison schools on what they are spending their catch up funding on and any spend on Academic Mentors (AM) or other tutoring. The data collected from comparison schools will be financial only. Note, for context, we will also ask TP schools what they are spending their catch up funding on (for example AM).

Cost information data collection

Table 8 below details the costs we hope to collect, what cost estimates they will feed into and how they will be collected. The school survey will provide much of this information and will be triangulated with school interview data. Other interview data will also provide context about costs.

Table 8: Cost data collection overview

Category	Item	Needed to calculate…	Pre-requisite, start up, ongoing cost?	Evaluation data source(s)
Programme management	Scoping time/ programme set up time/ management time	NTP TP - market cost	Start up/ ongoing	Communication with programme managers
TP management	Setting up/managing of the programme	NTP TP - market cost	Start up/ ongoing	Interviews with TPs/ accounts submitted to EEF
Personnel for training	Teacher/programme lead training on interface	NTP TP - time cost	Start up	School survey/ Head/ Programme Lead interviews
Personnel for preparation and delivery	Programme lead time to run/coord	NTP TP - time cost	Ongoing	School survey/ Head/ Programme Lead interviews
	Refreshments/pupil incentives to attend?/ rewards?	NTP TP - monetary cost	Ongoing	School survey/ Head/ Programme Lead interviews/ TP Interviews
	Supervision time for sessions	NTP - time cost	Ongoing	School survey/ Head/ Programme Lead interviews
	Programme lead time to set up/ launch session	NTP - time cost	Start up	School survey/ Head/ Programme Lead interviews
	Programme lead time for technical set up IT	NTP - time cost	Start up	School survey/ Head/ Programme Lead interviews
Training and programme-level costs	Programme fee (25%)	NTP - monetary cost	Ongoing	Invoices submitted to EEF
Facilities, equipment and materials	Headsets/ microphones/ pc/laptop/tablet	NTP - monetary cost	Pre-requisite	School survey/ Head/ Programme Lead interviews

BAU cost data collection

Data will be collected from comparison schools through a short online survey during Summer 2021.

Data on the market value of the different interventions will be collected direct from EEF (such as how much the programme cost, including the 25% contribution from schools and the 75% that was subsidised), and/or by reviewing payments made to TPs by EEF (if this data is available to be shared). This could be triangulated via interviews with TP if this is considered useful.

Cost evaluation analysis

The overall cost per pupil of tutoring sessions (or per block of tuition sessions) will be calculated. Providing data allows, to provide more meaningful estimates, the per pupil cost will also be calculated separately for different types of provision. For example for online versus face-to-face tuition and primary versus secondary. We envisage using the cost data to produce sensitivity analysis to reflect different moderators as outlined in the Evaluation overview and the Impact section on moderators where there is likely to be significant variation and where data allows.

These different costs will be calculated as the costs incurred by the schools as part of participation in TP (in terms of monetary cost and time committed) but also the market value of delivery as not all the costs will be borne by the schools. Costs will be calculated and presented following the most recent EEF cost evaluation guidelines. Average costs (for all financial costs) over a three-year time period (discounting for inflation) will be calculated as this allows comparison with other interventions and will provide more accurate estimates for the cost of TP going forwards.

Ethics

The evaluation went through ethical approval at project start up on 29th September 2020 – at a meeting where all members of the Evaluator team were present. This ethics checklist is a key process within NFER's Code of Practice (CoP), and any issues raised are escalated to CoP group. All items on the checklist met with approval and did not need to be raised. A copy of the checklist is in Appendix B. Note, at the time of writing this study plan, the University of Westminster are seeking ethical approval from their ethics committee. We will update the study plan with the outcome of this as soon as this is known, and address any issues accordingly.

All participants take part in the evaluation activities with informed consent.

Schools agree to take part in the programme, via the headteacher signing a Memorandum of Understanding (MoU). The headteacher confirms that they: have read and understood the information provided about the project; understand that their school's participation is voluntary and that they can withdraw their school from delivery or the evaluation at any time; will share the information sheet with the designated school lead, and the privacy notice with parents/pupils; know who to contact about complaints or concerns about the programme or the evaluation; and agree to facilitate where possible the activities for delivery and the evaluation (as set out in the MoU).

All participants (parents, and KS4 pupils, tutors, school staff and TP staff) are provided with a privacy notice relevant to processing their (or their child's) data. Participants can withdraw from data processing at any time during the evaluation – and instructions are provided in the privacy notice and withdrawal forms for how to inform their school, the TP and/or Evaluator that they do not with their data to be processed.

All interviewee and survey participants are provided with information about the purpose of the data collection and how their data will be used, prior to taking part in that data collection/giving their views. As above, they can withdraw from data processing at any time.

Data protection

Data protection statement and GDPR compliance

The Evaluator will be 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. Other members of the consortium have equivalent accreditations to demonstrate their compliance with DPA and GDPR.

To carry out the evaluation, it will be necessary to use and share personal data about pupils (both those who take up the offer and those who do not), as well as key staff members at participating schools and TP staff and tutors delivering the catch-up tuition, so that they can be asked about delivery.

The Evaluator has put in place appropriate measures to prevent pupils' personal information from being accidentally lost, used or accessed in an unauthorised way, altered or disclosed. In addition, each organisation involved will limit access to pupils' personal information to their staff members who have a business need to see it. Any data shared between the school, the Tuition Partners, EEF, the Evaluator and DfE will be via secure portal.

The online surveys will be administered using Questback. Questback's privacy statement can be found at https://www.questback.com/data-privacy/.

Legal bases

To make the use of pupils' data in the evaluation lawful, the Evaluator has identified specific grounds, known as a legal basis, for its processing. The legal basis available depends on the type of organisation, and these are outlined below.

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

The University of Westminster have 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 Tuition Partner is:

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.

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

NFER will securely submit the pupil data to the National Pupil Database (NPD) team to be matched to the pupil data held on NPD. The University of Westminster will access 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.

The project meets the Office for National Statistics "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 will be 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 cannot be done without processing personal data but processing does not override the data subject's interests. The research team and the EEF are committed to publishing the results of the study.

- Safe settings: all researchers working on the NPD data will only access the data via the SRS secure online system. Our organisations will apply for safe room connectivity to have SRS remote connectivity access.
- Safe outputs: All outputs will be checked by the ONS team to ensure that the outputs do not allow identification of individuals. Outputs will be 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 (PMR), as the DfE will match the data we collect with the NPD data. The PMR (meaningless identifier) replaces the 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.

The Parent (and KS4 Pupils with wording suitably adjusted) Privacy Notices contain the following information about personal data collection and linking to NPD:

- The Tuition Partner will collect some personal data about your son/daughter directly from their school, including name, date of birth, UPN, if your child is eligible for pupil premium and whether your child has special educational needs. They will also record any attendance at tutoring sessions.

- The Evaluator will also collect pupil background details, tutoring attendance, and assessment data from the school or the school's commercial test provider. The Evaluator will use your son/daughter's UPN to obtain further background information (for example their gender, ethnicity, household proximity to school and whether they are eligible for Free School Meals) from the NPD. The Evaluator will use short focus groups to gather pupils' views and perceptions of the programme.

- No individual will be named in any report for this project. Pupils' personal data will be shared between the organisations mentioned in this privacy notice. The school will provide their chosen Tuition Partner with information about your child. The Tuition Partner will share your child's data with the Evaluator. The Evaluator will be using a secure online portal to collect pupil data electronically. Your child's full name, date of birth and UPN will be shared with the NPD team to request their background characteristics.

- If data collected for the evaluation of the TP programme is to be used in other COVID-19 related research, it will be shared with the research organisations appointed to carry out that research.

- Three months after the publication of the 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 EEF and hosted by the 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 National Pupil Database. This data may also be linked to other research datasets for the purpose of Covid-19 related educational research. Further information about the EEF archive is available from: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluatingprojects/evaluator-resources/archiving-evaluation-data/

Rights and retention periods

Parents (and KS4 pupils) can withdraw their child from the programme and/or from their data being processed, until it is added to the EEF archive. Should they withdraw from the programme or evaluation (i.e. decide not to engage with Tuition Partners or the evaluation), the Evaluator will still use the evaluation data that the school has provided up to that point and link it to NPD unless the parent/KS4 pupil indicates otherwise. If at any time, parents/KS4 pupils wish to withdraw their data or have errors corrected in it, contact details are provided in the Privacy Notices for who to contact about this.

As noted above, three months after the publication of the 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 EEF and hosted by the 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 National Pupil Database. This data may also be linked to other research datasets for the purpose of Covid-19 related educational research.

The Evaluator will securely delete any personal data relating to the evaluation one year after the publication of the final report, currently expected to be December 2021.

The Tuition Partner will securely delete any personal data collected for the evaluation alone at the end of the TP programme, when final grants have been paid (expected to be August 2021).

The Tuition Partner may keep personal data collected as part of the delivery of their tuition services for longer – this is covered in the privacy notice they provide. Once data has been archived, it is held in the EEF archive until it is no longer needed for research purposes.

Data controller and processing roles

The Department for Education (DfE), the EEF and the Evaluator are joint data controllers for the evaluation. They decide how and what data will be collected and used. The Evaluator is also a data processor, as are Tuition Partners. (Note Tuition Partners are also a joint data controller in regard to data associated with the programme. This study plan is concerned with the evaluation.)

Personnel

Table 9: Personnel

Name	Institute	Roles and responsibilities
Pippa Lord	NFER	Project Director and Consortium Lead – responsible for directing the Consortium and quality of delivery.
Data management and operations workstream		
Kathryn Hurd	NFER	Workstream lead – responsible for overseeing data management, evaluation and comparison school recruitment, school contacting and testing

Jishi Jose	NFER	Project manager – responsible for overseeing the day-to-day running of the operations of the project
Guido Miani	NFER	Operations lead on data collection and impact evaluation activities
Impact workstream		
Helen Poet	NFER	Impact workstream lead – responsible for overseeing the impact workstream
Veruska Oppedisano	University of Westminster	Statistician and impact evaluation design
Richard Dorsett	University of Westminster	Overseeing impact evaluation design
Ben Styles	NFER	Impact QA
IPE workstream		
Ben Collins	Kantar	Director, responsible for IPE quality
Alice Coulter	Kantar	Research Director, responsible for IPE delivery
Rosie Giles	Kantar	IPE day-to-day lead
Sarah Tang	NFER	Cost evaluation lead
Matt Walker	NFER	NFER IPE consultant

Risks

Table 10: Evaluation issues and risks

Risk	Assessment	Controls, countermeasures and contingencies
Covid-19 restrictions lead to school or year group closures affecting the completeness of data	**Likelihood:** high **Impact:** high	Monitor lockdown procedures/Government announcements. Ensure sample of evaluation schools includes drop-out contingency. Ensure data can be submitted online (by schools and TPs). For evaluation schools, extend testing period(s) if restrictions are extensive – risk of delay to data feeds. Use NFER's Covid-19 research to understand the potential biases caused by children being kept off school and consider this in analysis.
Covid-19 restrictions within schools leading to evaluation challenges (e.g. physical handling of test papers and mixing of pupils;	**Likelihood:** high **Impact:** moderate	Online test options available. Paper tests quarantined for a period if tie prior to use. New tests are administered by schools with test administrators only offered as an option.

Risk	Assessment	Controls, countermeasures and contingencies
unable to host visitors)		
The rate of roll-out of TP is slower than planned e.g. shortage of tutors	**Likelihood:** moderate **Impact:** high	The programme evaluation timeline will respond to programme roll out, and adjust as needed (in terms of data collection from stakeholders). Adjust timeline to recruit evaluation schools if required; but ensuring 15 sessions and assessment data from all evaluation schools before the end of term might be affected.
School attrition	**Likelihood:** moderate **Impact:** moderate	Clear initial and ongoing communication with schools explaining principles and expectations. All schools sign MoU with clear identification of population requirements. Evaluation schools provided with further information, and choice of tests is driven by schools. Minimise burden on schools by using assessment they already use, and where possible collection of assessment data via the assessment provider's online tool. One key contact per school with regular keep in touch about next steps, and dedicated evaluation email address. Incentives for evaluation schools.
Contamination: comparison group schools receive TP	**Likelihood:** moderate **Impact:** moderate	Over-recruit comparison schools in order to allow for some signing up to TP. Pupil-level data covering the activities of pupil premium children in comparison schools will be collected to mirror the MI data in TP schools and included in the analysis.
GCSE teacher assessed grades not available on NPD	**Likelihood:** moderate **Impact:** high	Reporting can be delayed. When recruiting the schools to the evaluation sample we are asking if they would be willing to share their teacher assessed grades with us directly, should this scenario arise. Standardised tests from other year groups will provide substantial attainment data for the primary outcomes.
Data quality is low for the population data feeds	**Likelihood:** moderate **Impact:** moderate	Termly checks on TP data submissions, with a follow-up check on any missing data by EEF Programme Managers to encourage complete data submissions. Data includes a number of identifier fields per pupil (e.g. UPN, name, DoB) and so matching to NPD may be possible where there is incomplete data. Impact evaluation does not rely on population data.
Population data is not able to be matched or is matched incorrectly (across datasets)	**Likelihood:** moderate **Impact:** high	The population data will be collected by TPs and provided to NFER. Data matching will be carried out by the Evaluator on the variables available, using macros and programming to reduce human error. Collation/cleaning will be conducted in line with a data specification. Any erroneous-looking UPNs (such as with the wrong number of digits) will be cleaned out/removed.
Evaluation data is not able to be matched or is matched incorrectly (across datasets)	**Likelihood:** low **Impact:** high	Evaluation schools will be required to provide UPNs for all pupils in the evaluation sample. Any data queries will be addressed directly with the school, to ensure data accuracy. Matching will conducted in line with a data specification. A number of matching variables will be used per pupil (e.g. name, DoB, UPN) to ensure correct match.

Risk	Assessment	Controls, countermeasures and contingencies
Research Champions (impact/intervention sample) are not representative of the population of TP schools	Likelihood: moderate Impact: moderate	Sample to be monitored as schools are recruited. The RC sample will be recruited in parallel to the population of TP schools, as there is no deadline for schools to sign up to TP (other than being able to fit in the tutoring before the end of the academic year), so the population will not be defined until after the RC sample is complete. Population analysis also planned.

Timeline

Table 11: Planned timeline*

Date	Activity	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 publish	Consortium
Early December 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 – April 2021	W1 IPE fieldwork Surveys with school leads and tutors Interviews with TPs, school leads and tutors	Kantar
Dec 2020	First population data uploads; compilation and checks	NFER
Jan 2021	First formative feedback presentation to EEF – focusing on programme design, development and mobilisation, and reach in term 1	Consortium
5th January – 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 Mar 2021	Cut-off date for evaluation sample (Research Champion) recruitment	NFER
March – June 2021	W2 IPE fieldwork Interviews with TPs, school leads, classroom teachers, tutors and pupils	Kantar
End March 2021	Second population data uploads; compilation and checks	NFER
Mar- April 2021	Draw comparison sample and placebo check	UoW

Mid-Apr 2021	Confirm schools to be contacted for the comparison group	NFER/UoW
Mid-late Apr 2021	Second formative feedback presentation to EEF – focusing on mobilisation, reach in term 2, and delivery including challenges and, facilitators, moderators; schools and tutors views	Consortium
Mid-April - May 2021	Recruit comparison schools.	NFER
June 2021	Run placebo check again on the recruited comparison sample (weighting will be applied if required)	UoW
May – July 2021	W3 IPE fieldwork Surveys with school leads, classroom teachers and tutors Interviews with TPs, school leads, classroom teachers, tutors and focus groups with tutors and pupils	Kantar
End May 2021	Feedback presentation to TPs	Consortium
18th June 2021	Deadline for schools to submit their teacher assessed grades for GCSEs (Year 11)	
June/July 2021	Testing window for end-point standardised assessments	
July - August 2021	Summer term data collection from Research Champion and Comparison schools	
End August 2021	Final population data uploads from TPs; compilation and checks MI data collection from comparison schools	NFER
Late July to mid August 2021	Access assessment data from providers	NFER
Aug – Oct 2021	IPE analysis	Kantar
Mid August to Nov 2021	Data cleaning (matching MI/pupil data to standardised assessment data for impact evaluation sample) Send data to NPD to match in.	NFER
September 2021	Emerging findings presentation/meeting to EEF and to TPs and other stakeholders – focusing on IPE	Consortium
Nov-Dec 2021 (TBC re Y11)	NPD (unamended) data available and matched into dataset	NPD team/ UoW
Dec 2021 (tbc) - March 2022	Impact analysis	UoW
October to March/April (tbc) 2021	Draft reporting	All
April – Aug 2022 (tbc)	Final reporting and revisions	All
Aug/Sept 2022 (tbc) Publication		

* Reporting deliverables are shaded

References

Allen, R., Jerrim, J., Parameshwaran, M. and Thomson, D., 2018. *Properties of commercial tests in the EEF Database. EEF Research Paper Series No 001.* [pdf] Available at: <https://educationendowmentfoundation.org.uk/public/files/Support/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf> [Accessed 20 September 2021].

Clarke, D., Romano, J. P., & Wolf, M. (2020). The Romano–Wolf multiple-hypothesis correction in Stata. The Stata Journal, 20(4), 812-843.

Cullinane, C. and Montacute, R., 2020. *COVID-19 and social mobility impact brief #1: school shutdown.* [pdf] Available at: < https://www.suttontrust.com/wp-content/uploads/2021/01/School-Shutdown-Covid-19.pdf > [Accessed 28 September 2021].

Dietrichson, J., Bog, M., Filges, T., 2017. Academic interventions for elementary and middle school students with low socioeconomic status: a system review and meta-analysis', *Review of Education Research,* 87,(2), pp.243-282. https://doi.org/10.3102/0034654316687036.

Education Endowment Foundation, 2018a. *One to one tuition*. [online] Available at: <https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/one-to-one-tuition/> [Accessed 20 September 2021].

Education Endowment Foundation, 2018b. *Small group tuition*. [online] Available at: <https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/small-group-tuition/> [Accessed 20 September 2021].

Education Endowment Foundation, 2020. *Impact of school closures on the attainment gap: rapid evidence assessment*. [pdf] Available at: <https://educationendowmentfoundation.org.uk/public/files/EEF_(2020)_-_Impact_of_School_Closures_on_the_Attainment_Gap.pdf> [Accessed 20 September 2021].

Hainmueller, J., 2012, Entropy Balancing: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis,* 20(1), pp.25-46.

Kuhfeld, M., Soland, J., Tarasawa, B., Johnson, A., Ruzek, E. and Liu, J., 2020. *Projecting the potential impacts of COVID-19 school closures on academic achievement* (EdWorkingPaper No. 20-226). [pdf] Available at: <https://www.edworkingpapers.com/sites/default/files/ai20-226-v2.pdf> [Accessed 20 September 2021].

Marshall, L., Bury, J., Wishart, R., Hammelsback, R. and Roberts, E., 2021*. Independent Evaluation of the National Online Tuition Pilot*. [online] Available at: <https://educationendowmentfoundation.org.uk/public/files/National_Online_TuitionPilot.pdf> [Available 20 September 2021].

Moss, G., Aug 2020. *5 reasons to be cautious about estimates of lockdown learning loss.* [online] Available at: < https://schoolsweek.co.uk/5-reasons-to-be-cautious-about-estimates-of-lockdown-learning-loss/ > [Accessed 28 September 2021].

National Tuition Partners (2021). Guidance on signing up schools to Tuition Partners [unpublished].

National Tuition Partners (n.d.) Subsidised high-quality tutoring for schools from an approved list of Tuition Partners. [online] Available at: <https://nationaltutoring.org.uk/tuition-partners-funding-round> [Accessed 6 January 2021].

Nickow, A., Oreopoulos, P., and Quan, V., 2020. The impressive effects of tutoring on Prek-12 learning: A systematic review and meta-analysis of the experimental evidence, NBER

Working Paper No. w27476. [online] Available at<https://ssrn.com/abstract=3644077> [Accessed 20 September 2021].

Olea, J. L. M. and Pflueger, C. M.,(2013) A robust test for weak instruments. *Journal of Business and Economic Statistics* 31, pp.358-369.

Ritter, G., Barnett, J. and Denny, G., 2009. The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Education Research*, [e-journal] 79(1), 3-38. <https://doi.org/10.3102/0034654308325690>

Romano J.P. and Wolf M., 2005a. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing, *Journal of the American Statistical Association* 100(469): 94-108.

Romano J.P. and Wolf M., 2005b. Stepwise Multiple Testing as Formalized Data Snooping, *Econometrica* 73(4): 1237-1282.

Sharp, C., Nelson, J., Lucas, M., Julius, J., McCrone, T. and Sims, D., 2020. *Schools responses to Covid 19: The challenges facing schools and pupils in September 2020*. [online] Available at: https://www.nfer.ac.uk/schools-responses-to-covid-19-the-challenges-facing-schools-and-pupils-in-september-2020/ [Accessed 20 September 2020].

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*. [pdf] Available at: <https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Tutor_Trust.pdf> [Accessed 20 September 2021].

University College London, 2020. *Briefing note: Inequalities in resources in the home learning environment.* [pdf] Available at: <https://repec-cepeo.ucl.ac.uk/cepeob/cepeobn2.pdf> [Accessed 20 September 2021].

APPENDIX A – Use of incentives for the Evaluation of Tuition Partners

This appendix provides an outline of where incentives are being used as part of the TP Programme evaluation.

Table 12: Incentives for evaluation schools*

Activity	Payment made to schools
Schools that share additional pupil data and baseline and endpoint assessment data through commercial test providers that we are working with.	£500
Schools that either provide us directly with their test data or sit additional tests	£500 plus £150
In-depth Interviews School Head or Classroom Teacher**	A school would receive a £50 payment for each in-depth interview completed by a staff member.
Focus Groups with pupils involved in the TP Programme	Each school who takes part in a focus group would receive a payment of £100 as a thank you for their time.

*As schools' data provision was sometimes disrupted by Covid-related circumstances, incentive payment amounts will be tiered according to the amount/nature of data provided.

**This could alternatively be paid directly to the class teacher (e.g. where discussed with the school).

Table 13: Incentives for comparison schools

Activity	Payment made to schools
Schools that share pupil data and allow access to their commercial assessments through provider	£500***
Schools that share pupil data and either provide commercial test data directly to us or sit additional assessment	£500*** plus £150

***This amount was increased after publishing the first version of the study plan, to £1,000 in order to engage comparison schools in the evaluation.

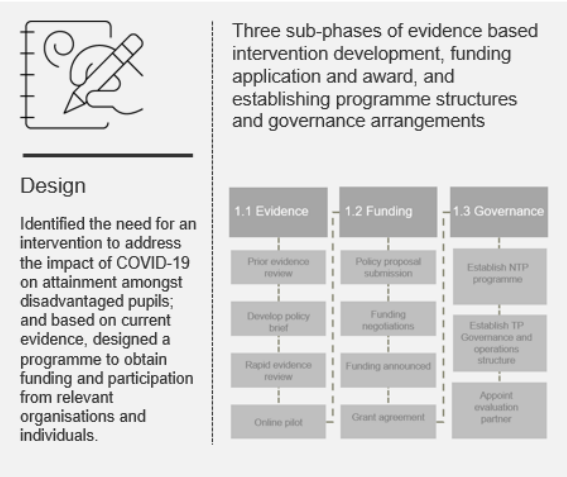
APPENDIX B: Code of practice and ethics approval checklist

Table 14: Code of practice and ethics approval checklist

Section of Code of Practice	Consideration of Code of Practice (CoP)	Yes	No	N/A
Ethics	Level of consent required – does the project allow for the level of consent required?	✓		
	Will research participants be provided with all the required information to enable them to make an informed choice?	✓		
	Have you looked at and do you intend to follow the guidance on selecting children/young people for interview?	✓		
	Will you follow the protection and safety guidelines?	✓		
	If the project involves children/young people have all those involved undergone disclosures/child protection training?	✓		
Data protection	Will the project follow the 8 principles of the data protection act?	✓		
	Will the project follow the rules for the processing of sensitive personal data?	✓		
Data security	Will the project allow for safe transfer of data into and out of our systems?	✓		
	Will the project include a secure coding system for recording participants' names?	✓		
	Have data transfer issues / protocols been discussed / confirmed with the client?	✓		
Caring for research participants	Will the project take into account designing research questions that make sense to children/young people?	✓		
	Will the project follow the guiding principles for the development of assessment instruments, methods and systems? (*Will only use standardised tests which we believe satisfy requirements)*	✓		
	Will the project involve taking, producing and using visual images? (Please refer to points to consider when taking photographs or video images, storing images, producing illustrations and using visual images)		✓	✓

APPENDIX C: Programme processes for each phase of activity

Figure 4a: programme process diagram – Design and Develop phases



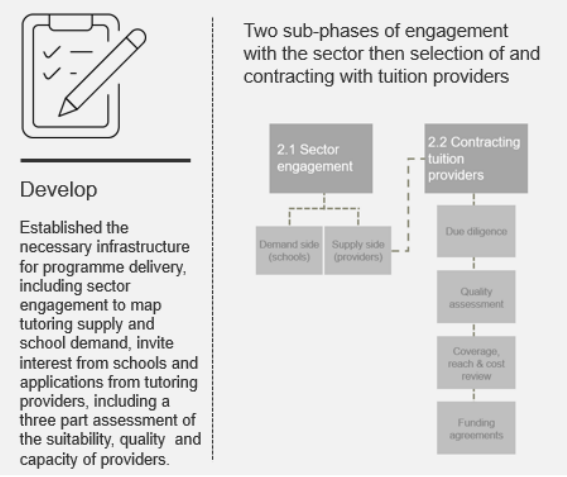


Figure 4b: programme process diagram – Mobilise phase

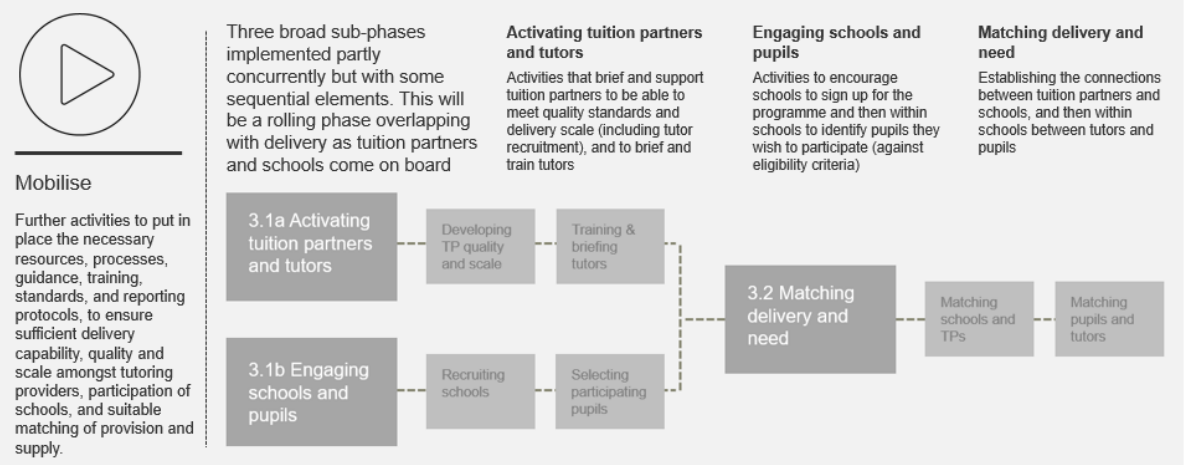


Figure 4c: programme process diagram – Delivery and Legacy phases