Project Title: Academic Mentoring Programme Evaluation Study Plan
Evaluator (institution): NFER, University of Westminster and Department for Education
Principal investigator(s): Pippa Lord

Table of contents

National Tutoring Programme: Academic Mentors – summary

The NTP is a key part of the Government's COVID-19 catch-up response for schools and forms part of the wider Catch-Up Funding. The overarching vision of the NTP is to improve academic outcomes of the most disadvantaged young people. In year 1 of the programme, the AM programme was one of two tutoring services offered to schools as part of the NTP. The NTP offered schools access to tutoring via two independent services: Tuition Partners (TP) and the Academic Mentoring Programme (AM). With TP, schools selected a tuition partner to work with from a list of approved tuition providers in their area. The tuition partner then provided tuition services for the school. AM was structured differently, but essentially provided a similar tutoring service. Teach First wholly managed the provision of tutors (referred to as 'academic mentors') to schools; recruiting, training and placing them in schools. Unlike TP, the mentor then worked in the school setting as a full time employee of the school. Despite the difference in name – 'tutor' for TP and 'academic mentor' for AM – both services aimed to provide a similar tutoring service, focused on 1-1 and small group intervention.

PROJECT TITLE	Academic Mentoring (AM) Programme Evaluation
EVALUATOR (INSTITUTION)	NFER , University of Westminster and the Department for Education (referred to as the Evaluator)
PRINCIPAL INVESTIGATOR(S)	Pippa Lord, Roland Marden
STUDY PLAN AUTHOR(S)	Veruska Oppedisano, Richard Dorsett, Rachel Hayes, Ben Styles, Roland Marden, Luke Bocock, Alice Phillips, Pippa Lord
STUDY DESIGN	Programme evaluation involving a quasi-experimental design (QED)
PUPIL AGE RANGE AND KEY STAGE	16 years – teacher assessed grades replacing KS4
NUMBER OF SCHOOLS	School level data: 4,190 secondary schools in England, of which 366 are AM schools and the others are non-AM schools: ○ 366 schools, all Y11 pupils per school, in AM schools ○ all Y11 pupils in 3824 comparison schools
NUMBER OF PUPILS	Pupils: ○ 4,190 secondary schools in England (366 intervention schools and 3,824 comparison schools), for English and math. Assuming 10 pupil premium pupils per year group, 45,560 pupil premium pupils (3,660 from

	intervention schools and 41,900 from comparison schools)
INTERVENTION	Academic Mentoring Programme
PRIMARY OUTCOME MEASURE AND SOURCE	Attainment in Maths and English in Year 11, based on the teacher assessed grades that replaced GCSEs in 2021.
SECONDARY OUTCOME MEASURE AND SOURCE	Not applicable
SCHOOL ELIGIBILITY	Schools met one of the following criteria to be eligible for AM: • an Income Deprivation Affecting Children Index (IDACI) of 40 or greater (40% of pupils live in the three most deprived deciles) • an IDACI of 35 to 40 and an Achieving Excellence Areas (AEA) score of 4 to 6 • an IDACI of 30 to 35 and an AEA score of 5 or 6 • an IDACI of 25 to 30 and an AEA score of 6 • Schools with an above average pupil premium rate (average set at 23.5) if there is an available local academic mentor There was also a flexibility rule that allowed up to 30% of schools to be recruited even if they didn't meet the national eligibility criteria above. This flexibility rule was only applied where a school met certain conditions that suggested it was 'disadvantaged' despite it not meeting the IDACI criteria. This enabled the programme to respond to other dimensions of disadvantage, e.g. rurality.

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.0 **[*original*]** 2.0	13/12/2021 13/12/2021	Publication of version 1.0 was delayed due to ongoing changes to the proposed methodology caused by cancellation of statutory assessments over academic year 2020/21.

Table 1: TIDieR framework – programme summary

Aspect	Description
Programme	National Tutoring Programme: Academic Mentoring Programme (referred to in the evaluation and this study plan as 'AM').
Why (rationale)	Research shows that pupils' learning has been affected by school closures due to Covid-19 (Cullinane and Montacute, 2020; Coe, 2020). The programme was designed to provide additional support to schools to help disadvantaged pupils whose education had 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 (Dietrichson et al., 2017; Education Endowment Foundation, 2018b; a; Nickow, Oreopoulos and Quan, 2020; Torgerson et al., 2018).
Who (recipients)	State-maintained primary and secondary schools serving disadvantaged populations in England were eligible to apply to the programme. Schools had to meet one of the following 'disadvantaged' criteria: • an Income Deprivation Affecting Children Index (IDACI) of 40 or greater (40% of pupils live in the three most deprived deciles) • an IDACI of 35 to 40 and an Achieving Excellence Areas (AEA) score of 4 to 6 • an IDACI of 30 to 35 and an AEA score of 5 or 6 • an IDACI of 25 to 30 and an AEA score of 6 • Schools with an above average (23.5) pupil premium rate if there is an available local academic mentor The programme encouraged schools participating in the programme to select pupils to receive mentoring from disadvantaged households or those whose education had been disproportionately impacted by Covid-19. The programme was expected to reach a minimum of 900 schools and 50,000 children.
What (materials)	AM was one two tutoring pillars of the National Tutoring Programme (NTP). The NTP was a key part of the Government's COVID-19 catch-up response for schools. The overarching vision of the NTP was to improve academic outcomes of the most disadvantaged pupils.
What (procedures)	AM was intended to provide 1-1 and small group tutoring outside of normal lessons.
Who (provider)	AMs were mostly graduates with some experience in education or working with pupils. They were not all qualified teachers but could be working towards an initial teacher training qualification or considering a career in education. They received a package of ongoing training delivered by Teach First and were managed by their school.

Aspect	Description
How (format)	Academic mentoring was delivered either online or face-to-face. Each eligible school could employ up to two mentors in the following subject areas: Primary • Numeracy • Literacy Secondary • Maths • English • Science • Humanities • Modern Foreign Languages
Where (location)	Academic mentoring was expected to be delivered in schools (before, during or after school). In certain circumstances, mentoring could be delivered at home (e.g. for pupils who were shielding/medically vulnerable and were 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 mentoring to be available for pupils at home. This led to a greater proportion of online tutoring.
When and how much (dosage)	AMs were recruited in three waves and started work in schools in either October 2020, January 2021 or February 2021. It was expected that AMs would work with at least 50 pupils between September 2020 and July 2021, pro-rata for AMs that started during the academic year.
Tailoring (adaptation)	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 a greater proportion of academic mentoring to be available for pupils online at home during that time.

About the National Tutoring Programme:

This section of the study plan sets out the current context in schools, the government-funding response, and the rationale for the NTP programme including evidence for academic mentoring. It then outlines the range of stakeholders involved in the programme (providers and participants). It highlights the key mechanisms and features of the programme and of mentoring 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 reopen from September 2020, and although schools had a duty to provide remote learning where needed, provision was still affected by the challenges of Covid-19 (e.g. local closures, year groups or bubbles unable to be in school, teachers' teaching online at the same time as in class).

Research highlighted that pupils were behind in their curriculum learning. In their review of the impact of school closures on attainment the Education Endowment Foundation (EEF) reported projections that school closures would 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 (Coe, 2020). According to the report of a national survey (weighted sample) of school leaders and teachers published in September 2020, teachers estimated that 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; Sharp et al., 2020; Hodgen 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 during this period: 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 (Education Endowment Foundation, 2018a) and on small group tuition (Education Endowment Foundation, 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 of tutoring to be 0.3 SD for small group tuition and 0.37 SD for one-to-one tuition; Nickow et al. (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.

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 mentoring should take into account wider context and 'moderators' that might affect the implementation of that mentoring. 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 AM 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).

Particular benefits of tutoring for disadvantaged students

There is evidence to suggest that the advantages of 1-1/small group tuition may be particularly relevant for disadvantaged students (Dietrichson et al., 2017; Torgerson et al., 2018). These students may suffer in the classroom due to comparison to their peers. A perceived sense of failure may result in low motivation and low self-efficacy leading to poor learning outcomes from classroom learning. In contrast, teaching these students 1-1 or in homogenous small groups, allows favourable comparisons between students and allows teachers to readily communicate student improvements (Mischo and Haag, 2002). These incentives, in turn, help maintain high levels of motivation (Pintrich and Schunk, 2002).

Who is involved in the AM programme? Providers and participants

The following stakeholders were involved in the AM programme:

- Programme Managers at Teach First were leading the design and development of the AM programme in the academic year 2020–2021[1]. Programme Managers at Teach Frist were responsible for ensuring there was a rigorous recruitment and selection process of AMs, along with support and training throughout the academic year. Teach First continued to coach the AM through the year via Curriculum and Training Leads.
- Academic Mentors (AMs): AMs were responsible for delivering academic mentoring to participating pupils in schools with a focus on providing small group and one to one tutoring. AMs were recruited in three waves and could join a school in either October 2020, January 2021 or February 2021. It was expected that AMs would work with at least 50 pupils between September 2020 and July 2021, pro-rata for AMs that started during the academic year.
- Schools: State-maintained primary and secondary schools that served disadvantaged communities throughout England were eligible to apply for an AM. In this context, disadvantaged communities was defined by the Income Deprivation Affecting Children Index (IDACI) and Achieving Excellence Areas (AEA) measures. Schools that had an above average pupil premium rate could also apply for an AM.
- Pupils: Participating schools were able to identify which of their pupils they felt would most benefit from academic mentoring. Pupils could be in Year 1 – Year 11. AM did not have prescribed conditions on the characteristics of pupils who received the intervention, although schools were encouraged to support pupils from disadvantaged

[1] Note, the design and delivery of the whole of the National Tutoring Programme in its first year was led by a collaboration of five charities - the Education Endowment Foundation, Sutton Trust, Impetus, Nesta and Teach First - supported by the KPMG Foundation.

households or those whose education has been disproportionately impacted by Covid-19. The programme was expected to reach 50,000 pupils in the academic year 2020-21.

Context in schools: Spring 2021

In January 2021, part-way through the AM programme, another national lockdown was announced, and schools were told to close to pupils other than those whose parents 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 academic mentoring, which had been planned to expand in earnest in January 2021. A number of key changes were made to delivery at this point by the NTP:

- most delivery moved online.
- the NTP agreed that during the lockdown online mentoring could be delivered to pupils learning from home.
- face to face mentoring was often only allowed within a defined 'bubble' i.e. a set group of pupils.

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.

About the Academic Mentors Pillar of the National Tutoring Programme: programme theory and design

This section of the study plan identifies the importance of certain delivery features and structures for effective learning. Academic mentoring was designed to address the effects of the loss of teaching time from school closures due to COVID-19. The clear message from the research is that academic mentoring needs to be sufficiently high quality with sessions having the right duration and frequency to achieve optimal results. It is recommended that tutors are knowledgeable in their subject area and trained in pedagogy, and that they deliver at least weekly sessions to students for a term or longer. Tutor subject knowledge and pedagogic expertise are commonly identified as important delivery elements for tuturing as well as the following structural characteristics, relationship with classroom learning, duration and frequency.

Tutor subject knowledge

The literature suggests tutor subject knowledge is beneficial for learning outcomes. Skilled teaching requires a complex interrelationship between knowledge of lesson structure and subject matter (Leinhardt and Greeno, 1986). Tutors with strong subject knowledge are more likely to be able to communicate that knowledge effectively to students. But learning can still occur where it is less strong, for example, when tutors are peers or volunteers (Fantuzzo, King and Heller, 1992; Rogoff, 1990). Therefore, although tutor subject knowledge should not be considered a prerequisite for tutorial learning it is clearly advantageous and preferable to it not being present at all.

Pedagogic expertise

The techniques that tutors use to facilitate learning is widely acknowledged in the literature as important. In particular, academic mentoring that exploits the intimate environment offered by 1-1/small group tutorials is likely to be highly effective (Collins and Stevens, n.d.). In this sense, tutorials should be an interactive rather than a didactic experience between tutor and student (Lepper, Drake and O'Donnell-Johnson, 1997; Lepper and Woolverton, 2002). Tutors should make the tutorial a learning conversation in which students contribute much of the dialogue and the tutor intervenes appropriately to guide learning (Education Endowment Foundation, 2018a; McArthur, Stasz and Zmuidzinas, 1990; Merrill et al., 1992). Among the most important pedagogic principles identified is the idea of tutors managing conversations that encourage active learning from students (Chi et al., 2001). Ideally, students should be at the centre of these learning conversations, encouraged to explain their answers and ask questions and with tutors holding back from giving detailed explanations. Tutors should also use this conversational style to probe students' understanding of content. For example, this could include tutors using comprehension-gauging questions rather than accepting student's own assessment of their understanding.

Structured format

Aside from delivery, there is considerable focus in the literature on the most effective format for sessions. This relates to the frequency and duration of sessions as well as, when interventions take place in school, how sessions are coordinated with classroom learning. This latter issue is particularly relevant to AM as all mentoring is directly managed by the school. The clear message in the literature is that the format and coordination of sessions with

classroom teaching has an important impact on the effectiveness of academic mentoring. In terms of format, short, regular sessions (30-40 minutes, three to five times a week) over a term or more appear to result in optimum impact. In terms of coordination, the close alignment of teaching to the classroom curriculum is strongly recommended.

Relationship with classroom learning

An issue of concern in the literature is how targeted school interventions such as tutoring relate to wider school learning. Research suggests that learning is more effective when academic mentoring is linked with regular classroom teaching (Education Endowment Foundation, 2018a). However, experience has shown that academic mentoring in school can often be quite separate from classroom activities with relatively little connection between what students experience in and away from the classroom. In practice, this means it can be left to the student to make links between the coverage of the intervention and the wider curriculum coverage back in the classroom. Given that supported students are usually those who find accessing learning difficult in the first place, this presents an additional challenge.

The academic mentoring students receive should therefore be closely aligned with what is being taught in regular classes, for example, by providing remedial support on difficult topics. The coordination of academic mentoring and classroom teaching should be fostered by a close and supportive relationship between tutor and teacher.

Duration and Frequency

Most studies demonstrate higher learning gains from extended periods of academic mentoring. For example, one study found that students receiving less than 20 hours tutoring scored 1 grade point higher than non-participants and those who had received more than 20 hours tuition scored 1.8 points higher than those who had no tuition (Smyth, 2008). Also the 20 week programmes Every Child a Reader and Every Child a Writer both showed larger achievement gains than the 10 hours of tuition provided through the Making Good Progress (Every Child a Chance Trust, 2009a; b). Studies suggest that intensive tutoring, where sessions are held several times a week tend to have greater impact (Elbaum et al., 2000)

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 but also in terms of how the programme as a whole is working, and what improvements needed to inform future tuition programmes.

The evaluation aims to quantify the impact of the AM programme on pupil attainment/learning outcomes.

The evidence generated through this evaluation will be used to help the AM programme and tutors design and deliver future high-quality tuition that makes a difference to pupils' attainment. It will also help schools in future academic years to better target and engage the pupils who would benefit from tutoring. Given the unprecedented pandemic-related circumstances in which the AM programme was implemented and the continuing Covid-19 related disruptions in schools throughout the academic year 2020/21, the findings from the evaluation will need to be interpreted in light of this 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

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

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 population data for reporting on the monitoring data provided by AMs about participants[2]); burden on schools; use of NPD data; and the timescales of the programme and how best to provide formative feedback throughout the evaluation. These issues are discussed further where relevant.

We had originally planned to conduct analysis on an evaluation sample of primary (years 1-6) and secondary (years 7-10) schools in addition to the year 11 analysis. For this analysis we were planning to use test results for Renaissance Learning (RL) standardised assessments. Unfortunately the sample of schools we managed to secure was small and it was decided not to proceed with this option.[3] The analysis will now focus only on year 11 pupils using as outcomes the teacher assessed GCSE grades awarded in 2021.

[2] Data received from 366 Secondary schools and 531 Primary schools for a total of 38,300 pupils.
[3] The RL school sample size produced an MDES of 0.43 (Primary) and 0.46 (Secondary) which was significantly above the accepted threshold of 0.20.

Research questions

The research question for the evaluation is as follows (note further sub-research questions are detailed in Table 1):

Impact RQ: What is the impact of AM on learning outcomes for pupils? (this will be investigated through a number of estimators of impact, in maths and English in Year 11).

The primary objective of the impact evaluation is to determine what difference, if any, is made by AM to attainment outcomes (maths and English). The impact evaluation uses a quasi-experimental design, due to the need to maximise reach to as many schools and pupils as possible, as soon as possible (an objective that a randomised trial would have been likely to constrain).

We are concerned that the process for grade determination for the teacher assessed grades (the attainment outcomes) may mean that it is difficult to detect any potential impact of the AM programme for the reasons outlined in the analysis section. 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 will include analysis of attainment outcomes controlling for i) pupil characteristics, including gender, ethnicity, English as an additional language (EAL), prior attainment, pupil premium (eligible vs. not), SEND vs. not, NTP tutoring received (vs. not); ii) school characteristics such as Ofsted rating (using the four-point scale), proportion of FSM (high vs. low), and iii) and geography (London, Government Office Region, and regions).

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

We will also run descriptive analyses to compare outcomes associated with different mentoring models among AM schools in the evaluation sample. The differences considered will include mode of delivery (online vs. face to face) and mentor:pupil ratio (1:1 to 1:10). We will also examine variation in outcomes by tutor characteristics (Qualified Teacher Status; shared gender with pupil/mentor).

The research questions are summarised in Table 1. There are five research questions. In all cases, there are two outcomes: maths, year11 and English, year11. For each research question, the outcome/phase combinations are identified by a suffix: ey11 (English, year 11) and my11 (Maths, year 11). For each specific research question, Table 1 gives the outcome, the phase, the type of pupil for which impacts are estimated and the type of school used in the analysis. The final column provides a brief description.

Table 1 - Research questions

Research Question**	Outcome	Phase	Pupil type	School type	Description
RQ1: What is the impact of AM availability on PP* pupils' attainment?					
RQ1.ey11	English	secondary	PP	AM/non-AM	The estimation method is based on matching/weighting and instrumental variables. This approach builds on the eligibility criteria. We will use two instruments in our estimator: the combination of the IDACI and AEA thresholds and the fraction of PP pupils above the average. We focus on PP pupils since they are expected to form a large proportion of the eligible group and can be identified within both the AM 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 who received AM as a whole. The analysis will be on PP pupils in year 11 in AM and non-AM schools. This analysis will use the teacher assessed data that are replacing GCSEs in 2021. It will use KS2 scores as baseline covariates.
RQ1.ey11	Maths	secondary	PP	AM/non-AM	
RQ2: What is the impact of AM availability on the attainment of pupils predicted to participate?					
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Description**
RQ2.ey11	English	secondary	predicted participants	AM/non-AM	The estimation method is based on matching/weighting and instrumental variables. This approach builds on the eligibility criteria. We will use two instruments in our estimator: the combination of the IDACI and AEA thresholds and the fraction of PP pupils above the average. Using NPD data and AM attendance data, we estimate a predictive model of pupil participation within AM schools. We use the results of that estimation to predict which pupils participate in both AM and non-AM schools (had AM been available to them). By doing this, we aim to move closer to an impact on AM participants as a whole rather than PP pupils.
RQ2.my11	Maths	secondary	predicted participants	AM/non-AM	

					This analysis will use the teacher assessed data that are replacing GCSEs in 2021. It will use KS2 scores as baseline covariates.
RQ3: What is the impact of the availability of AM on all pupils' attainment?					
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Description**
RQ3.ey11	English	secondary	All Y11 pupils	AM/non-AM	The estimation method is based on matching/weighting and instrumental variables. This approach builds on the eligibility criteria. We will use two instruments in our estimator: the combination of the IDACI and AEA thresholds and the fraction of PP pupils above the average. We estimate impacts for all pupils in year 11 in schools receiving AM, regardless of whether they participate in AM. These estimates are likely to be smaller than RQ1 and RQ2 estimates since there is no attempt to identify pupils more likely to participate in AM and so its impact will be more diluted. Should AM 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 spill-over (peer) effects. This analysis will use the teacher assessed data that are replacing GCSEs in 2021. It will use KS2 scores as baseline covariates.
RQ3.my11	Maths	secondary	all Y11 pupils	AM/non-AM	
RQ4: How does the impact of AM availability vary among PP pupils, by school and pupil characteristics?					
Research Question	**Outcome**	**Phase**	**Pupil type**	**School type**	**Description**
RQ4.ey11	English	secondary	PP	AM/non-AM	We revisit RQ1 to explore the extent to which impacts for PP pupils vary according to a number of school and pupil characteristics. The analysis will use KS2 for the baseline covariate and teacher assessed grades (GCSE) for the outcome measure.
RQ4.ey11	Maths	secondary	PP	AM/non-AM	
RQ5: How do outcomes vary among AM pupils, by model of tutoring?					

Research Question	Outcome	Phase	Pupil type	School type	Description
RQ5.ey11	English	secondary	Pupils receiving AM	AM	We examine how outcomes vary according to a number of aspects of AM-related factors using a regression of attainment on mentoring characteristics for the sample of AM schools, using AM pupil level attendance data. Since these are only observed among AM schools, we present descriptive statistics (i.e. not causal) rather than impact estimates. The analysis will use KS2 for the baseline covariate and teacher assessed grades (GCSE) for the outcome measure.
RQ5.ey11	Maths	secondary	Pupils receiving AM	AM	

** ey11=English, year 11, my11=Maths, year 11

Design overview

Table 2: Design overview

Design		Matching/weighting Instrumental Variable
Unit of analysis (school, pupils)		Year 11 pupils

Number of Units to be included in analysis (Intervention, Comparison)[4]		**Secondary:**
		• Schools:
		  ○ 366 AM schools
		  ○ 3824 non-AM schools
		• Pupils:
		  ○ 4,190 secondary schools in England (366 intervention schools and 3,824 comparison schools), for English and maths. 45,560 pupil premium pupils (3,660 from intervention schools and 41,900 from comparison schools)

Outcomes	Variable	Educational attainment
	measure (instrument, scale, source)	Teacher assessed grades in maths and English.
Baseline for outcome	**Variable**	Educational attainment
	measure (instrument, scale, source)	KS2 SATs from NPD

[4] 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.

Participants and selection mechanism

The impact analysis will focus on the population of Y11 children in schools in England. Compared to the previous proposed analysis using the sample of primary and secondary RL schools, the proposed analysis has the advantage of no selection involved in taking part in the evaluation other than choosing to become an AM school. The relative disadvantage is that it is limited to year 11, it does not use recent baseline assessments and the outcome measure is potentially less reliable this academic year than standardised assessments.

How schools join the AM programme: AM schools

Schools are eligible to apply for an Academic Mentor if they have either:

an Income Deprivation Affecting Children Index (IDACI) of 40 or greater (40% of pupils live in the three most deprived deciles)

an IDACI of 35 to 40 and an Achieving Excellence Areas (AEA) score of 4 to 6

an IDACI of 30 to 35 and an AEA score of 5 or 6

an IDACI of 25 to 30 and an AEA score of 6

or if a school does not meet the criteria above, but has an above average pupil premium, set at 23.5%.

Descriptive statistics

Part of the evaluation will involve descriptive statistics of reach and spread of the provision overall – both in terms of the AM schools and AM pupils in the AM population. AM population data will be used to describe the number of schools and pupils involved in the AM programme, and their characteristics (such as Pupil Premium eligibility), as well as an overview of attendance at sessions. AM data include the list of secondary schools where at least one pupil receives AM mentoring. This includes 366 schools that will be used to define AM availability and address RQ1-RQ4. For 188 secondary schools, we have detailed pupil level AM attendance data that will be used to run some of the checks on the suitability of the teacher assessed grades, provide descriptive statistics on pupils involved in AM, predict AM participation and address RQ5. We will check that the school level characteristics of the 188 schools are similar to the 366 schools. The population data is collected and recorded on an ongoing basis by AMs. The evaluator will collate, but not quality assure, the data that is provided. The evaluator will match the population data and the pupil level attendance data to the NPD.

Sample size calculations

We use cluster randomised trial power calculations to provide an indication of the MDES. We allow for clustering of pupils within schools. We note that the evaluation is not based on a randomised intervention but instead relies on a non-experimental approach. Hence, these power calculations should be viewed as approximations.

Our sample size calculations refer to the population of 366 secondary schools doing AM and are computed in relation to RQ1. Assumed sample size is based on a conservative level of take-up within schools (we have assumed 10 pupils per school[5]). It also reflects our intention

[5] The number is based on the data on pupils receiving AM. On average there are 20 pupils per school doing AM. About 50% of them are PP pupils which leave us with 10 pupils.

to focus primarily on Pupil Premium pupils as the intervention targets disadvantaged children that can therefore be easily identified in control schools. Note that since our analysis focuses on disadvantaged pupils, we do not produce separate estimates for the FSM subgroup.

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. Our assumptions regarding the ICCs are, if anything, conservative. Allen et al., 2018 suggest ICCs of 0.15 at the end of KS4.

Power calculations indicate that with 366 AM secondary schools and a minimum of 366 non-AM schools the MDES would be 0.07. This calculation used the R package 'PowerUpR'.

As noted above, these power calculations are based on a randomised experiment design. The analysis will use matching and IV regression. Assuming the identifying assumption for matching holds, the reported MDES should be good approximations for the matching case. With IV, it is less clear that the reported power calculations will provide good approximations.

Outcome measures and other data

Key principles

The QED aims to look at the impact of mentoring on attainment, as the purpose of the NTP AM programme 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 will access de-identified data from NPD for all y11 pupils in all English schools. 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 AM schools and the AM pupils so we know which school and which pupils in Y11 received mentoring.

Primary outcome

For Year 11 we will use teacher assessed grades. We are concerned that the process for grade determination may mean that it is difficult to detect any potential impact of the AM programme for the reasons outlined in the preliminary analysis section. 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.

NPD data

As our current understanding (November 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 sample pupil list via school Unique Reference Number (URN), 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 in year 11 in the academic years 2014/15-2020/21, including their final KS2 and/or KS4 pupil data from 2014/15 until 2020/21 (including those with endks=0 or schres=0)[6].

We will request the following variables from the NPD:

- 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

Preliminary analysis – identifying preferred approach

The impact analysis will use matching/reweighting and IV. Within these two approaches, there are several analytical choices. To help optimise this choice, we will conduct a preliminary assessment of the effectiveness of various approaches in earlier years.

Matching and reweighting both assume that selection can be adequately captured by appropriately controlling for observed characteristics such as:

- AEA score;
- IDACi score;
- Proportion of FSM,
- GPS (grammar punctuation spelling) average grades in year 2007;
- Region;
- School type;
- KS1 and KS2 value added (at district level).

The objective is to control for school selection into AM. Under propensity score matching, we would estimate the propensity score (the probability of being an AM school) and construct the comparison group as the subset of non-AM schools that have a similar distribution of propensity scores. Weidmann and Miratrix, 2021 provide evidence that simple matching approaches may work well for this purpose. Under reweighting, we might use the estimated propensity score to re-weight the sample (inverse probability weighting, IPW) or an alternative approach such as entropy balancing (implemented in Stata using the ebalance ado file) which has the advantage of automating the process of balance checking and thereby reducing the scope for researcher bias.

In both cases, we end up with a weighted sample, where the weights are intended to control for school-level selection. Estimated impacts are likely to be more precise (and, arguably, less likely to be biased) if they control also for pupil characteristics and we will include regressors capturing gender, age, FSM, English as an additional language, special educational needs and interaction with social services and KS2 results (all variables taken from the NPD).

Broadly, the options are

[6] endks indicates whether a pupil is at the end of a specific KS and schres indicates whether a pupil should be included in the results calculations for their school taking into consideration any amendments, refugee status and age group where relevant.

a) expand the set of variables included in the estimation of the propensity score to include pupil characteristics (thereby moving from a school-level to a pupil-level exercise)
b) control separately for pupil characteristics (and potentially school-level characteristics) in a second stage. This second stage could either be a further stage of matching or a linear regression.
c) A combination of a) and b) which, in the case of IPW, is known as "doubly robust" estimation.

We will experiment with these approaches, conducting placebo tests in earlier years using the previous cohorts of NPD data. If the selection of the control sample controls adequately for unobserved factors, we do not expect to find any significant difference in attainment between AM and control schools in the years prior to the intervention. The placebo testing will be done for each of the three preceding years, 2017-2019 (except 2020 for which there are no KS2 and GCSE available), using results for KS2 and GCSE to demonstrate similarity of the achieved match in the past. Estimators will be assessed according to:

- How closely the matched comparator schools resemble the AM schools.
- The size and precision of the placebo impact estimates in prior years.
- The ability of the estimators to detect simulated impacts in prior years.

The **IV estimators** will similarly be assessed using data from prior years. Here, there are fewer analytical decisions but it is still useful to conduct placebo tests to assess the likely performance and sensitivity of the approach. Specifically, we will assess:

- The size and precision of the placebo impact estimates in prior years.
- The ability of the estimators to detect simulated impacts in prior years.

Sensitivity analyses – probing the nature of teacher-assessed grades

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[7].
>
> Concern 2: Knowledge/selection of pupils doing AM 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

[7] 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.

mentoring 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 mentoring. 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 these checks 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.

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 not to run the analysis as we already know that 2021 results are different to prior years. 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[8]. If teachers do inflate 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 AM and non-AM 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 AM 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 mentored. 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

[8] See previous footnote 7.

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 mentoring 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 AM schools versus non-AM 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.

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 are confirmed by both tests, they will clearly indicate the existence of systematic bias across AM versus non-AM 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 AM on a mentored 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 mentoring programme 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 AM and non-AM schools):

i) To address uncertainties around whether the teacher assessed grades will reflect pupils' performance after the mentoring and whether the assessments would be sensitive enough to change (concerns 3 and 4), we plan to exploit dosage of AM mentoring. Assessment is an average of performance change, we can use dosage (as it exogenously depends on time of enrolling) on the sample of AM 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 AM.

Dosage data are collected for the sample of 188 schools with detailed pupil-level data. They measure the number of sessions each year 11 pupil received.

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 AM/non AM schools for all pupils and PP pupils. Evidence of significant difference in distribution across AM and non-AM schools may suggest the presence of measurement errors (i.e. bump-ups).

If AM and non-AM both bump up scores equally, then we would expect AM schools to have a slightly more positive impact because of AM (if AM is

effective). If AM and non-AM 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 AM.

iii) To address the concern that schools bumped up grades (concern 1), we will perform the analysis on Y11 pupils predicted to do AM. If the effect is positive, this can be due to the positive effect of AM or to bumping up of grades. We will perform the analysis on Y11 children predicted NOT to participate in AM. If there is an impact also on children not predicted to participate in AM, 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 AM is effective, predicted AM should always have higher scores than predicted NO-AM even if there is bump up. The reliability of this test depends on how well we can predict participation to AM. The test cannot disentangle the increase in grade due to AM from a systematic teacher bias towards AM pupils only.

While we already know that that Concern 1 (bumping up grades) is very likely to be an issue, it is not indicative of systematic bias between AM and non-AM 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 could be taken to imply the presence of systematic bias between AM and non-AM schools (specifically: ex ante test (ii); and ex post tests (ii) and (iii), see above). We highlight the fact that there could be more than one interpretation to these checks, that may not allow us to detect bias: i) the possibility of between subject spillovers; (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 AM 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 AM 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.

Primary outcome analysis

We will use matching/weighting estimates and IV regression. The precise nature of the implementation of these estimators will be guided by the results of the preliminary analysis described above.

Matching/weighting: What is the impact of AM availability on PP pupils' attainment? (RQ1)

We will use weights to construct a comparison group of non-AM schools observationally similar to the AM sample. These weights will be derived using either a matching or a reweighting approach (see discussion above).

Impacts will be estimated using weighted regression of pupil outcome on a 0/1 indicator for AM being available. Other regressors may include: the school level variables used to estimate weights and pupil-level characteristics including baseline KS2 attainment, gender, ethnicity, EAL and special educational needs. Residuals will be clustered at the school level to account for any common school-specific unobservable component. The software used to run the model is Stata.

This analysis will be conducted on the PP pupils only and will estimate the impact of AM, on an 'intention to treat basis' using both pupil- and school-level variables. This allows us to more tightly control for differences between the AM and non-AM schools. In addition to the variables listed in the preliminary analysis subsection (AEA score, IDACi score, proportion of FSM, GPS (grammar punctuation spelling) average grades in year 2007, region, school type, KS1 and KS2 value added, at district level), we will also include KS2 attainment, gender, ethnicity English as an additional language and special educational needs (all variables taken from the NPD).

We will report a comparison of the characteristics of AM with:

1) all non-AM schools;
2) the matched sample of comparison schools;

Instrumental variable: What is the impact of AM availability on PP pupils' attainment? (RQ1)

IV estimates do not rely on the selection on observables assumption. Our approach builds on the AM eligibility criteria and is based on the intuition that, for schools close to the thresholds, eligibility is as good as randomly assigned. An implication of this is that, for such schools, eligibility influences AM participation but does not otherwise materially affect outcomes. This in turn suggests that eligibility can be used as an instrument in an IV analysis of the impact of AM on outcomes. Since we know some eligible schools will not participate in AM and some ineligible schools will participate, we will implement our estimator using IV (which we note is also robust to the possibility that some ineligible schools may participate). In fact, we can use two instruments in our estimator: the combination of the IDACI and AEA thresholds and the fraction of PP pupils above the average. 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 AM availability among pupils in schools close to the eligibility thresholds.

In practice, we will operationalise this by estimating a 2-stage least squares regression of the outcome on an indicator of whether the school is eligible for AM (instrumented using the two eligibility criteria) and the same variables used for the matching/reweighting estimate. The coefficient on the AM coefficient will be the impact estimate.

The approach relies on the eligibility criteria having an impact on participation. If this is not the case, IV estimates can be unreliable. This is the problem of weak instruments. A preliminary analysis on the data of schools recruited in the AM intervention until March shows that AM schools are more likely than non-AM schools to satisfy at least one of the two eligibility criteria.

Figure 1 shows eligibility on the grounds of IDACI (x-axis) and AEA (y-axis). Pink dots identify eligible schools and blue dots identify non-eligible schools. Figure 2 shows participation in AM

on the grounds of IDACI (x-axis) and AEA (y-axis). Pink dots identify AM schools who are not eligible according the IDACI/AEA and according to the PP criteria and blue dots identify AM schools who are eligible according to the IDACI/AEA criterion. The graph shows that the majority of the AM schools satisfy the eligibility as defined by the two indexes.

Figure 3 shows eligibility on the ground of the second criterion, by plotting the distribution of schools according to the fraction of PP pupils. The red line at 0.25 identifies the average value of fraction of PP in the sample. The majority of schools in the non-AM sample lie on the left of the threshold, while the majority of the schools in the AM sample lie on the right of the threshold. The two graphs provide evidence that the eligibility criteria actually can predict AM participation.

Having two instruments, we can perform the Sargan overidentification test to test for the exogeneity of the instruments. The test, whose null hypothesis is that the instruments are appropriately independent of the error process, allows to evaluate the validity of the instruments.

The nature of our instruments is such that the estimator has a regression discontinuity design. For this, we require that the variables conferring eligibility -IDACI/AEA and PP proportion - are continuous at the thresholds. We also require that schools cannot manipulate their eligibility status. This requirement is likely to be met since it is determined by IDACI, AEA and PP proportion, all of which are not under the control of the schools (certainly at the time of the programme implementation).

To estimate impacts, we will regress the outcome on an AM indicator and a baseline measure of the outcome of interest. AM will be instrumented with the eligibility criteria. Pupil-level controls will include background variables, such as gender, ethnicity, EAL, FSM and special educational needs. Standard errors will be clustered at the school level to account for any common school-specific unobservable component. Regression will be based on pupils in the AM eligible and ineligible schools. The software used to run the model is Stata. The coefficient on the AM indicator will represent the estimated treatment effect, on a 'local average treatment effect' (LATE) basis.

To check the robustness of estimates obtained parametrically, we will estimate the treatment effects via non-parametric method, making use of the threshold in this fuzzy Regression Discontinuity Design. The advantage of the non-parametric method is to provide estimates based on data closer to the thresholds cut-off, reducing bias that may otherwise result from using data farther away from the cut-off. Following Imbens and Gelman (2014), the specification will use local linear and quadratic polynomials.

Figure 1: IDACI and AEA scores for AM (pink) and non AM (blue) schools

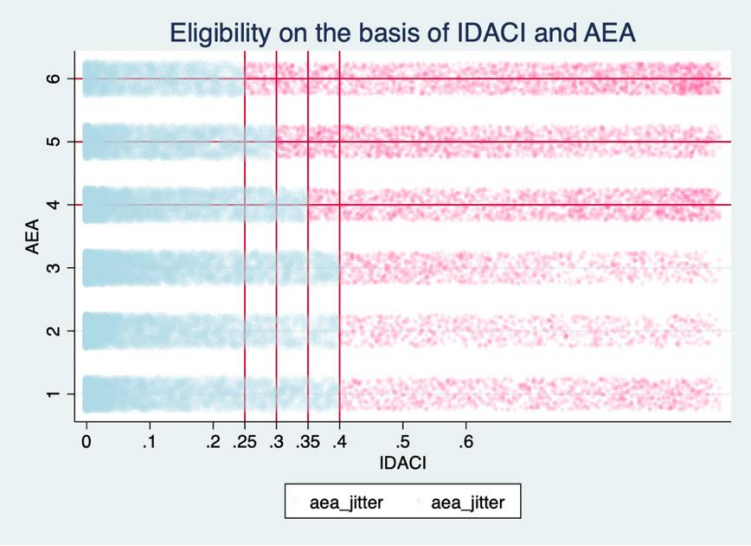


Figure 2: Participation in AM on the basis of the IDACI and AEA scores

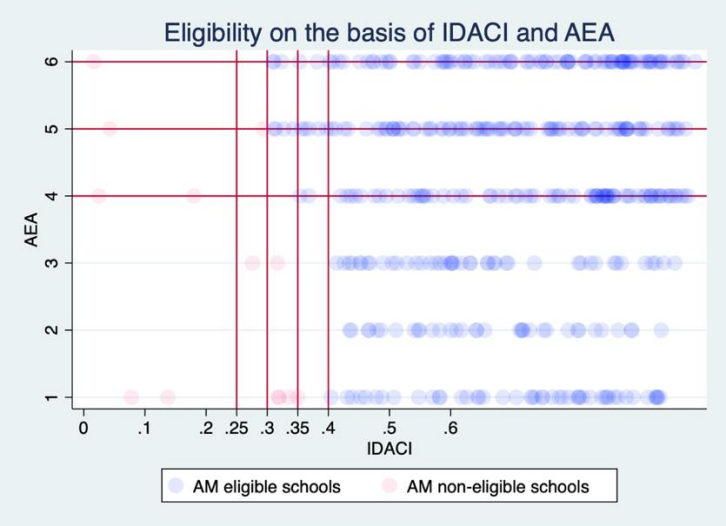
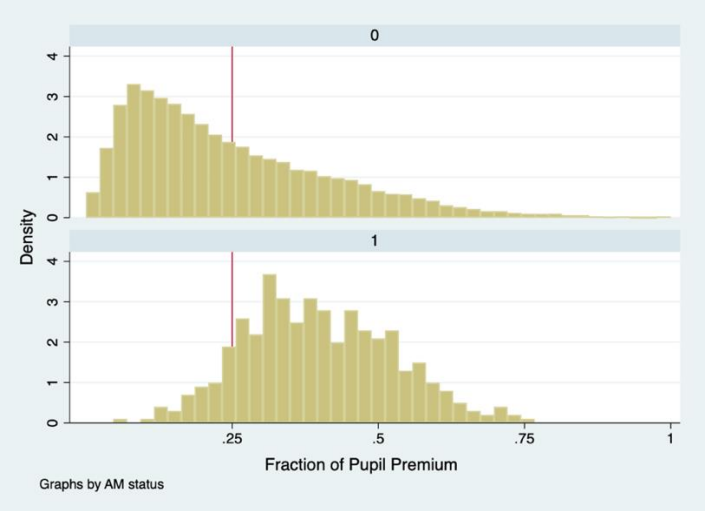


Figure 3: Distribution of PP across AM/non AM schools



Inference

Uncertainty will be conveyed using confidence intervals. We will adjust for multiple testing using the simulation approach of Clark et al. (2020), as implemented by the Stata program rwolf.ado.

Secondary analyses

What is the impact of AM 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 population, PP pupils, which may not coincide with the group of children who will receive the intervention. AM does not have prescribed conditions on the characteristics of pupils who receive the intervention. Preliminary data on the intervention show that about 50% of pupils receiving AM are PP pupils.

RQ2 involves an alternative approach to approximating the group of children who may receive AM. It involves modelling the probability of pupil participation in AM schools, using various markers of disadvantage recorded in the NPD and attainment (socio-economic status measured by FSM/PP, special educational needs, interaction with social service, prior attainment, English as additional language and ethnicity). To this, we will use pupil-level AM attendance data provided for a sample of 188 secondary schools delivering AM to Y11. The results will be used to predict participation in both AM and non-AM schools in the sample. Having done this, we will follow a similar approach to that of RQ1 but instead of selecting PP pupils, we will select predicted AM pupils (where, in non-AM schools, predicted AM pupils are those who would be predicted to participate were AM available).

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

As another means of understanding the overall effect of AM, a third analysis will focus on attainment of all Year 11 pupils (rather than PP pupils or predicted AM pupils) with AM in AM schools compared with all Y11 pupils in comparison non-AM schools in the evaluation sample. These estimates are likely to be smaller than RQ1 and RQ2 estimates, as the AM 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 AM. The regression analysis will control for the same school level and pupil level characteristics mentioned in the primary analysis.

Moderator analysis: How does the impact of AM availability vary among PP pupils, by school and pupil characteristics? (RQ4)

Moderator analysis will be conducted through interaction terms on both the matching/weighting and the instrumental variable specifications on the following categories individually of:

1. School characteristics: Ofsted rating (high vs. low, defined as above/below median); proportion of FSM (high vs. low, defined as above/below median). These variables identify the context where AM is delivered and allow to analyse whether AM has been more effective in disadvantaged contexts.
2. Pupil characteristics: prior attainment; SEND vs. not; English as an additional language vs. not, ethnicity (defined according to the seven major ethnic groups (Asian, Black, Chinese, White, Mixed and any other) and gender. These variables identify participants and allow to analyse whether AM has been more effective on disadvantaged pupils or children with specific demographics.
3. Other: geography (urban/rural; low/high IDACI, , defined as above/below median): different geographical areas may have different provision of AM in terms of quantity and quality. If schools in more deprived areas have lower attainment outcomes, this may be correlated with the supply of AM in those areas, and the risk that low-quality tutors reach more disadvantaged schools.

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

A descriptive analysis (using the data collected via templates for the above impact analysis) will compare outcomes associated with different mentoring models and moderators among AM schools in the evaluation sample. We do not propose any impact analysis within RQ5 since we cannot observe the counterfactual treatment model among non-AM schools. Instead, this element of the analysis will summarise mean attainment among participating pupils in AM schools according to the model of mentoring they experience. We will regress attainment on the variables listed below for the sample of AM schools only to assess heterogeneity. The analysis will be performed on the sample of 59 AM secondary schools that have pupil level attendance information. In particular, we will look at:

1. The intervention: mode of delivery of completed sessions (online vs. face to face); if the sample of pupils experiencing both online and face-to-face is sufficiently large we will use them as a third category; mentor:pupil ratio (1:1 up to 1:10).
2. Mentors: QTS qualification/not; shared gender with pupil/mentor.

Missing data

We do not expect to find missing values at school level, as we are using the population of schools in England, and all provided teacher assessed grades for Year 11 pupils. For missing data on variables used for matching and weighting and on covariates, we will define them as missing and control for them in the analysis.

Effect size calculation

Matching/reweighting 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.

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 A. The University of Westminster has obtained the approval from the Ethics Committee to conduct the research in June 2021.

All participants (parents, and KS4 pupils, tutors, school staff and AM 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 for how to inform their school, Teach First and/or Evaluator that they do not with their data to be processed.

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 AM staff delivering the mentoring, 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, Teach First, EEF, the Evaluator and DfE will be via secure portal.

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 and NFER 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 Teach First 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

- Teach First will collect some personal data about your son/daughter directly from their school, including name, date of birth, UPN, year group. They will also record any attendance at mentoring sessions.
- The Evaluator will also collect pupil background details and assessment data from 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, eligibility for pupil premium, free school meals, English as an additional language, has a social worker or is a looked after child, has special educational needs, a disability or has fallen behind or is at risk of falling behind in their school work) from the NPD.
- 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 Teach First with information about your child. Teach First 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 AM 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 Academic Mentors 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 April 2022.

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

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

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 Teach First. (Note Teach First are also a joint data controller in regard to data associated with the programme. This study plan is concerned with the evaluation.)

Personnel

Name	Institute	Roles and responsibilities
Pippa Lord	NFER	Project Director and Consortium Lead – responsible for directing the Consortium and quality of delivery.
Kathryn Hurd	NFER	Workstream lead – responsible for overseeing data management, evaluation and comparison school recruitment, school contacting and testing
Kinnery Koria	NFER	Project manager – responsible for overseeing the day-to-day running of the operations of the project
Roland Marden	NFER	Project Director for AM workstream
Palak Roy	NFER	Project Leader for AM workstream
Veruska Oppedisano	University of Westminster	Statistician and impact evaluation design
Richard Dorsett	University of Westminster	Overseeing impact evaluation design
Min Zhang	University of Westminster	Analyst on the impact evaluation
Alice Phillips	Department of Education	Impact evaluation design
Kim Williams	Department of Education	Impact evaluation design
Arnaud Vaganay	Department of Education	Impact evaluation design
Ben Styles	NFER	Impact QA

Risks

Table 3: Evaluation issues and risks

Risk	Assessment	Controls, countermeasures and contingencies
Evaluation data is not able to be matched or is matched incorrectly (across datasets)	**Likelihood:** High **Impact:** High	Data sharing agreement between TF and RL does not include pupil names. As well as the RL test scores, the school name, postcode and URN will be provided along with the DOB for every child. RL has also requested UPN, year group, age and gender for each child, but these are not mandatory variables. Therefore there is a concern that matching this data into the NPD will be challenging. We intend to over-recruit schools to mitigate this risk.

Timeline

Date	Activity	Responsible/ leading
Oct 2020	Project set up, logic model development, materials development, study plan development	Evaluator
Oct 2020 – July 2021	*Mentoring period (whole programme)*	AMs
5th January – 8th March 2021	National lockdown period – many pupils learning from home, schools only open to children of keyworkers and vulnerable children. AM provision moved to online only.	
April 2021	Submit NPD request	UoW
Mid May to early July 2021	GCSEs (year 11) *Also window for end-point testing (+ to end of summer term)*	Schools
Summer 2021	Study plan finalisation and publish	Evaluator
Mid August to early Sept 2021	Data cleaning (MI/pupil data)	NFER
September 2021	Assessment data from RL sent to Teach First and shared with DfE	Teach First
November 2021	Study plan publication, to match with revised TP study plan publication	Evaluator
Mid October 2021	Draw comparison sample and placebo check	UoW
Mid October 2021	Run placebo check again on the selected comparison sample (weighting will be applied if required)	UoW
December 2021	NPD (unamended) data available and matched into dataset	NPD team/ UoW
February-April 2022	Impact analysis	UoW
May 2022	Write first draft of report	All
June-September 2022 (tbc)	Report revisions and report publication	All

References

Allen, R., Jerrim, J., Parameshwaran, M. and Thompson, D., 2018. *Properties of commercial tests in the EEF database*. [online] Available at: <https://educationendowmentfoundation.org.uk/public/files/Support/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf> [Accessed 5 Aug. 2021].

Carpenter, J.R., Kenward, M.G. and White, I.R., 2007. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, 16(3), pp.259–275. https://doi.org/10.1177/0962280206075303.

Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T. and Hausmann, R.G., 2001. Learning from human tutoring. *Cognitive Science*, 25(4), pp.471–533. https://doi.org/10.1207/s15516709cog2504_1.

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

Coe, R., 2020. *Impact of school closures on the attainment gap: Rapid Evidence Assessment*. [online] London, UK: Education Endowment Foundation. Available at: <https://dera.ioe.ac.uk/35707/1/EEF_%282020%29_-_Impact_of_School_Closures_on_the_Attainment_Gap.pdf>.

Collins, A. and Stevens, A., n.d. Goals and methods for inquiry teachers. In: R. Glaser, ed. *Advances in instructional psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cullinane, C. and Montacute, R., 2020. COVID-19 and Social Mobility Impact Brief #1: School Closures. [online] Available at: <https://www.suttontrust.com/our-research/Covid-19-and-social-mobility-impact-brief/>.

Dietrichson, J., Bøg, M., Filges, T. and Jørgensen, A.-M.K., 2017. Academic Interventions for Elementary and Middle School Students With Low Socioeconomic Status: A Systematic Review and Meta-Analysis: *https://doi.org/10.3102/0034654316687036*, 87(2), pp.243–282. https://doi.org/10.3102/0034654316687036.

Education Endowment Foundation, 2018a. *One to one tuition | Toolkit Strand | Education Endowment Foundation | EEF*. [online] Available at: <https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/one-to-one-tuition/> [Accessed 5 Aug. 2021].

Education Endowment Foundation, 2018b. *Small group tuition | Toolkit Strand | Education Endowment Foundation | EEF*. [online] Available at: <https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/small-group-tuition/> [Accessed 5 Aug. 2021].

Elbaum, B., Vaughn, S., Hughes, M.T. and Moody, S.W., 2000. How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, 92(4), pp.605–619. https://doi.org/10.1037/0022-0663.92.4.605.

Every Child a Chance Trust, 2009a. *Every Child a Reader. The first year of the national rollout*. London, UK: Every Child a Chance Trust.

Every Child a Chance Trust, 2009b. *Every Child Counts. The results of the first year 2008/09*. London, UK: Every Child a Chance Trust.

Fantuzzo, J.W., King, J.A. and Heller, L.R., 1992. Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology*, 84(3), pp.331–339. https://doi.org/10.1037/0022-0663.84.3.331.

Hodgen, J., Taylor, B., Jacques, L., Tereshchenko, A., Kwok, R. and Cockerill, M., 2020. *Remote mathematics teaching during COVID-19: intentions, practices and equity*. [Report] *UCL Institute of Education: London, UK.* London, UK: UCL Institute of Education. Available at: <https://www.ucl.ac.uk/ioe/> [Accessed 5 Aug. 2021].

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. *Educational Research*, 49(8), pp.549–565. https://doi.org/10.3102/0013189X20965918.

Leinhardt, G. and Greeno, J.G., 1986. The cognitive skill of teaching. *Journal of Educational Psychology*, 78(2), pp.75–95. https://doi.org/10.1037/0022-0663.78.2.75.

Lepper, M.R., Drake, M.F. and O'Donnell-Johnson, T., 1997. Scaffolding techniques of expert human tutors. In: *Scaffolding student learning: Instructional approaches and issues.*, Advances in learning & teaching. Cambridge, MA, US: Brookline Books.pp.108–144.

Lepper, M.R. and Woolverton, M., 2002. Chapter 7 - The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. In: J. Aronson, ed. *Improving Academic Achievement: Impact of Psychological Factors on Education*, Educational Psychology. [online] San Diego: Academic Press.pp.135–158. https://doi.org/10.1016/B978-012064455-1/50010-5.

McArthur, D., Stasz, C. and Zmuidzinas, M., 1990. Tutoring Techniques in Algebra. *Cognition and Instruction*, 7(3), pp.197–244.

Merrill, D.C., Reiser, B.J., Ranney, M. and Trafton, J.G., 1992. Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. *Journal of the Learning Sciences*, 2(3), pp.277–305. https://doi.org/10.1207/s15327809jls0203_2.

Mischo, C. and Haag, L., 2002. Expansion and Effectiveness of Private Tutoring. *European Journal of Psychology of Education*, 17(3), pp.263–273.

Moss, G., 2020. Reasons to be cautious about lockdown learning loss estimates. *Schools Week*. [online] Available at: <https://schoolsweek.co.uk/5-reasons-to-be-cautious-about-estimates-of-lockdown-learning-loss/>.

Nickow, A., Oreopoulos, P. and Quan, V., 2020. The Impressive Effects of Tutoring on Prek-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *National Bureau of Economic Research working paper no. w27476*, 20(20), pp.267–350. https://doi.org/10.3386/w27476.

Pintrich, P.R. and Schunk, D., 2002. *Motivation in education: Theory, research, and applications*. 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Ritter, G.W., Barnett, J.H., Denny, G.S. and Albin, G.R., 2009. The Effectiveness of Volunteer Tutoring Programs for Elementary and Middle School Students: A Meta-Analysis. *Review of Educational Research*, 79(1), pp.3–38. https://doi.org/10.3102/0034654308325690.

Rogoff, B., 1990. *Apprenticeship in Thinking. Cognitive Development in Social Context*. New York: Oxford University Press.

Schafer, J.L., 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), pp.3–15. https://doi.org/10.1177/096228029900800102.

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

Smyth, E., 2008. The More, the Better? Intensity of Involvement in Private Tuition and Examination Performance. *Educational Research and Evaluation*, 14(5), pp.465–476.

Torgerson, C.J., Bell, K., Coleman, E., Elliott, L., Fairhurst, C., Gascoine, L., Hewitt, C.E. and Torgerson, D.J., 2018. Tutor Trust: Affordable Primary Tuition Evaluation report and executive summary. [online] Available at: <https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Tutor_Trust.pdf>.

Weidmann, B. and Miratrix, L., 2021. Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *Journal of Policy Analysis and Management*, 40(3), pp.964–986. https://doi.org/10.1002/PAM.22236.

Appendix A: 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)		✓	✓