# YEAR 1 OF THE NATIONAL TUTORING PROGRAMME (2020/21): EVALUATION CONTEXT, CHALLENGES AND CONSIDERATIONS

Paper authors: Pippa Lord, Helen Poet and Palak Roy

October 2022

This evaluation was undertaken by a consortium led by NFER

The evaluation was commissioned by the Education Endowment Foundation (EEF)

# Contents

How to cite this report: Lord, P., Poet, H. and Roy, P. (2022). *Year 1 of the National Tutoring Programme (2020/21): Evaluation Context, Challenges and Considerations*. NFER. From an evaluation conducted by NFER, the University of Westminster and Kantar Public. London: Education Endowment Foundation.

# National Tutoring Programme (2020/21): Evaluation context, challenges and considerations

## 1. About the year 1 evaluation

### 1.1 The programme

This paper sets out the challenges of evaluating year 1 of the National Tutoring Programme (NTP), a large-scale, complex programme delivered during the Covid-19 pandemic. The NTP (2020/21) was made up of two pillars: the Tuition Partners (TP) programme (which provided tutoring support to pupils) and Academic Mentoring (AMs) (in which mentors were placed in schools to work with small groups of pupils).[1] This paper outlines how these programmes responded to the ongoing disruptions relating to the Covid-19 pandemic, and how the evaluations of both TP and AM (2020/21) were revised to reflect the changing context, the challenges encountered, and the considerations needed when interpreting the results.

This paper was written prior to analyses being finalised, as a preface to the context in which the evaluation was undertaken. Hence, it is written in the present tense at the time of writing, and anticipating analytical challenges as well as which analyses would and would not be able to go ahead.

The 2020/21 NTP programme was funded by the Department for Education (DfE) and was originally developed by the Education Endowment Foundation (EEF), Nesta, Impetus, The Sutton Trust, Teach First, and with the support of the KPMG Foundation.

The 2020/21 evaluation of the NTP was commissioned by the EEF, and was carried out by an independent Consortium led by the National Foundation for Educational Research (NFER), with Kantar Public and the University of Westminster (UoW). The study plans for each of TP[2] and AM[3] are available online.

The year 1 programme was set up during the Covid-19 pandemic. It required rapid expansion of the tutoring market/provision; accreditation of tutoring organisations; and recruitment of tutors and mentors. Schools were contending with continued periods of remote teaching, staff and pupil absences, wellbeing and mental health, as well as learning recovery. It was a challenging context for schools, the programme and the evaluation.

According to the monitoring data provided by TPs, 46% of the tutored pupils for whom data was provided were eligible for Pupil Premium (PP-eligible) (this compares to *c*. 24% PP-eligible pupils nationally at the time). In AM, 89% of the schools met Teach First's priority criteria, which is based on the proportion of children living in income deprived families (IDACI) and whether the school is in an area of chronic and persistent underperformance (AEA). The remaining 11% of schools had an above average proportion of pupils eligible for Pupil Premium (Teach First, 2021). For AM, 49% of the pupils who took part were eligible for PP or Free School Meals (FSM) (Teach First, 2021) (this compares to 21% FSM pupils nationally in 2020/21).

### 1.2 The evaluation

The evaluation of TP aimed to assess the impact of the programme on pupils' learning, and explore the implementation and delivery of the programme in 2020/21. The evaluation of AM aimed to assess the impact of the programme on pupils' learning; a separate process evaluation of AM was carried out by Teach First.[4]

The evaluations used data provided by Tuition Partners, Teach First and schools, as well as data from the National Pupil Database (NPD) and from assessment providers. The evaluations aimed to compare the attainment outcomes of pupils who took part with the attainment outcomes of pupils who did not take part. To do this, the evaluations used a quasi-

[1] Note, school-led tutoring was not a feature of year 1 of the NTP.
[2] Available online: https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/TP_Overarching-Eval_Study-Plan_V2.pdf?v=1637742905
[3] Available online: https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/Study-Plan-Academic-Mentors-Evaluation-version-2.pdf?v=1641812799
[4] A separate process evaluation was carried out by Teach First, and is published here: https://www.teachfirst.org.uk/reports/amp-phase-1-report

experimental design (QED) that aimed to create similar groups of schools and pupils that did and did not take part. The challenges in identifying and constructing the comparison groups are outlined in this paper.

## 1.3 Considerations for interpreting the findings

This paper discusses these evaluations and the considerations needed when interpreting their results. These include:

- Year 1 of the NTP was set up during a pandemic, with limited time for planning and scale-up of the tutoring market/provision prior to launching (unlike other interventions or programmes which might have considerably more lead-in time); and it necessarily responded to the ongoing circumstances created by the pandemic, which created challenges for both delivery and the evaluation (see sections 3 and 4).

- Not all of the planned analyses could proceed and so the picture we will be able to present will not be as rounded as originally intended. Originally we aimed for the study designs to include all of the year groups experiencing tutoring. However, due to cancellation and changes in national assessments in 2020/21 and the challenges of recruiting during a pandemic, the impact analysis for TP only includes primary schools and Year 11 pupils; Years 7–10 were dropped from the evaluation. And for AM, the impact analysis only includes Year 11 (see sections 4 and 5 for further details).

- The TP and AM impact evaluations aimed to present several different estimators of impact on groups of pupils with similar characteristics in intervention and comparison schools, namely: PP-eligible pupils, all pupils in the year groups involved, and by aiming to match pupils on participant characteristics (i.e., to predict participation). However, it was difficult to identify a group of pupils based on characteristics that would predict participation, and so this particular analysis will not go ahead (see section 5 for more details). In addition, the analysis of PP-eligible pupils suffers from dilution, as not all PP-eligible pupils would be selected for tutoring. Indeed, of PP-eligible pupils, the proportions taking part were relatively low. This means that any impact of tutoring will be harder to detect when analysing all PP-eligible pupils in the year group(s) (as per the evaluation design) – as only a small proportion of them received tutoring. Section 5 has further details.

- The Year 11 analysis is exploratory in nature due to the only available outcome measure (Teacher Assessed Grading – TAG) for which we have no prior data to compare to (see section 5 for further details about the checks we carried out).

- The impact reports will also include analysis comparing outcomes associated with different tutoring models (e.g., face-to-face or online delivery), however this element is descriptive and will therefore provide information about association, but not causation.

- The TP Implementation and Process Evaluation (IPE) includes well-sampled, large-scale qualitative and quantitative perceptual data, providing a rich account of the delivery and experience of NTP TP Year 1. It involved: meetings and workshops with EEF programme managers, over 280 in-depth interviews (with TPs, school leads, classroom teachers and tutors), 34 focus groups (with pupils and tutors), and five online surveys with tutors (over 10,000 responses across two waves), school leads (over 1800 responses across two waves), and school staff (over 800 responses). It explores implementation, including key aspects of high-quality tutoring (such as communication/planning with the school, addressing pupils' needs, tutor–pupil relationships), and the challenges of delivering during a pandemic. The evidence on learning outcomes in the IPE is based on large-scale responses on how pupils are perceived to have benefited, and should be seen as a valuable source of data.

- The TP IPE also explores what else TP schools were doing to support their pupils (e.g., how they were spending their catch-up funding), but it does not explore any learning recovery support in place in the comparison group. This may be a limitation of the evaluation; however, we know from other studies (Nelson, Lynch and Sharp, 2021; Rose *et al*., 2021; Harland *et al*., 2022) that schools are putting in place a range of recovery strategies and support, and so it is likely that all schools in the evaluation may be recovering to some extent. This makes it harder to isolate the effect of the tutoring support – as this would be part of the mix of support schools were putting in place (schools could also use the 'one-off universal' catch-up premium[5] for learning recovery).

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

- Teach First undertook a separate IPE evaluation and published a report (Teach First, 2021). This will be referred to, but not integrated within, the evaluation of AM conducted here.

# 2. Aims and scope of year 1 of the NTP evaluation

## 2.1 The TP (2020/21) evaluation was designed to:

i) collate data provided by 33 TPs about the schools (planned to be 6000+), pupils (215,000–265,000) and tutors (20,000+) involved

ii) explore programme implementation and delivery through a large-scale IPE using interviews with school leaders, teachers, TPs, and tutors; online surveys with school leaders, tutors and school staff; and focus groups with pupils and with tutors

iii) analyse the impact on pupils' attainment in English and maths using a QED (i.e., it aimed to create similar groups of schools and pupils that had and had not taken part in TP), to assess the impact of the programme. This was originally designed to use several estimators of impact for the following samples:

- primary English, primary maths, secondary English, secondary maths – each of these would entail a sample of schools (both TP and non-TP schools) that signed up to share their pupil assessment data (i.e., routinely conducted standardised assessments) for the evaluation – known as 'Research Champion' samples

- Year 6 English, Year 6 maths, Year 11 English, Year 11 maths – each of these would entail a sample of schools (both TP and non-TP) and use attainment data from the NPD for the whole year groups of pupils involved – known as 'population analyses'.

The impact analysis was also designed to explore how impact varies by pupil- and school-level characteristics, and by mode of delivery (online/face-to-face, 1:1/small group).

## 2.2 The AM (2020/21) evaluation was designed to:

i) analyse the impact on pupils' attainment in English and maths using a QED (i.e., it aimed to create similar groups of schools and pupils that had and had not taken part in AM), to assess the impact of the programme. The evaluation was designed to use data from year groups that used Renaissance Learning (RL) assessments (primary school sample Years 1–6, and secondary school sample Years 7–10), and NPD attainment data for Year 6 and Year 11. As per TP, the design would use several estimators of impact, including those based on participation of PP-eligible pupils in AM schools, and similar pupils in comparison schools.

ii) The impact analysis was also designed to explore how impact varies by pupil- and school-level characteristics, and by mode of delivery.

# 3. A summary of changes to the NTP programme, delivery and policy context

## 3.1 How did the TP programme respond to the challenges of the ongoing pandemic? Were any changes made to the programme or delivery?

As noted in the introduction, the year 1 programme was set up during the Covid-19 pandemic, requiring continued responsiveness to the challenges faced as schools re-opened. The ongoing pandemic affected implementation of, and participation in, the programme. This was especially the case during the period of school closures for most pupils from January to March 2021. During this period, EEF made a change to approve providers to deliver at-home online tutoring in specific circumstances. This resulted in some schools taking up online at-home tutoring instead of the designed in-school tutoring model. However, many schools chose to wait to commence tutoring after schools reopened, and therefore started tutoring later than planned. This resulted in TPs needing to rehire and train more tutors in the summer term. Once schools were reopened fully, more schools opted for online provision rather than face-to-face tutoring for Covid-related reasons.

Delivery was also disrupted in the summer term of 2021 due to Covid-related absences of pupils and tutors. This affected delivery of and attendance at tutoring sessions; including group delivery, and whole year group absences in cases where all pupils were recommended to self-isolate. A shift was also seen in the pupils selected for tutoring, from Year 6/Year 11 to Year 5/Year 10, potentially related to the cancellation of the national assessments.

To support increased tuition delivery in the shorter time available once schools reopened fully, EEF introduced more flexibility to the offer, including expanding online at-home tuition into weekend and half-term provision, extending the TP programme into the summer holidays, and allowing shorter blocks of 10 hours of tuition for schools that had not yet started tuition, later in the summer term. Just under a third of tutoring sessions took place later in the year (i.e., after summer term assessments) – sections 4 and 5 provide further details about how we took this into account in the evaluation and the considerations needed.

## 3.2 How was the AM programme affected by the ongoing challenges of the pandemic? Were any changes made to the programme or delivery?

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

## 3.3 Changes to national assessments

**Other key policy decisions** with implications for the evaluation of both TP and AM included: the cancellation of KS2 (Year 6) assessments, and the usual summer exams process for Year 11 pupils could not go ahead as planned in summer 2021, and GCSEs were determined by a Teacher Assessed Grading (TAG) process instead. Sections 4 and 5 provide further details.

# 4. A summary of changes to the evaluation

## 4.1 Were any changes made to the TP evaluation?

Changes made to the TP study design in 2021 in response to the changes outlined above included:

- removing the Year 6 analysis on the full population of TP schools from the evaluation due to the cancellation of the KS2 national assessments in 2021, as there was no alternative national-level data available

- amending the outcome measure for the Year 11 analysis from GCSE grades to using the TAGs that was implemented for 2021. As this was a new and unique assessment approach for one year, we conducted additional analysis checks prior to the main analysis to understand whether the TAGs would be a suitable outcome measure for our analysis (e.g., in terms of sensitivity and reliability) (section 5 provides further details)

- dropping the secondary school English and maths Research Champion evaluation samples, as recruitment was affected by the period of school closures to most pupils in January–March 2021 and other priorities in schools, and we had insufficient schools to create an evaluation and comparison sample (see further detail in section 5)

- extending the evaluation period to include data monitoring about summer holiday delivery of tutoring

- collecting more detailed dosage data on dates of sessions. This was in response to much of the tutoring being delivered later in the school year as a result of the school closures to most pupils in early 2021, as a way of establishing how much tutoring had taken place at the time of the end-point assessment in the summer term. EEF anticipated that around one-third of delivery would take place after mid-June 2021. According to the delivery data provided by TPs, of the sessions where session delivery dates were recorded, 29% of tutoring sessions happened after 11th June 2021 (note, 41% of booking rows did not provide detailed dates per session). Any schools that had not started delivery before the assessments were excluded from the TP samples.

## 4.2 Were any changes made to the AM evaluation?

Changes made to the AM study during 2021 included:

- removing the Year 6 analysis on the full population of AM schools due to the cancellation of the KS2 national assessments in 2021, as there was no alternative national-level data available

- removing the RL evaluation samples (primary school Year 1–Year 6 and secondary school Year 7–10) from the study design – after a change in data sharing arrangements by RL, despite considerable efforts to re-contact schools, the number of schools providing agreement was insufficient to warrant impact analyses on the evaluation samples using RL data

- amending the outcome measure for the Year 11 analysis from GCSE grades to using the TAGs that was implemented for 2021, with similar additional checks carried out as for TP.

# 5. Methodological challenges and limitations

## 5.1 What are the main challenges of a QED evaluation design?

The evaluation was not able to randomise tutoring to pupils: given the urgency of the requirement for catch-up support in schools it was not considered ethical to do so. QEDs are the next best impact evaluation tool, but they have challenges and limitations, chiefly relating to creating a suitable comparison group.

In contrast to most EEF evaluations, which usually use a randomised controlled trial (RCT), this evaluation uses a QED. It was not possible to randomise schools or pupils to receive the tutoring in TP or in AM. This was because it was deemed necessary to roll out the programme as soon as possible, given the urgency of addressing the missed learning. In the TP evaluation, we gave TPs the option of randomising pupils within schools to early or later delivery, but due to a number of practical considerations, including the speed at which the programme was set up and the subsequent disruptions to schools in January–March 2021, no schools came forward with this pattern of delivery and so this did not proceed. For more information see the TP study plan.

A QED requires a suitable comparison group to be constructed. In an RCT, we can assume that unobservable factors (such as attitude, or motivation to participate in the programme) are randomly distributed and therefore evenly assigned to the intervention/programme group and comparison group. In a QED we recruit schools that have signed up for the programme and then use matching methods to identify a similar group of schools (or pupils) based on a range of observable characteristics, but that are not participating in the programme. This addresses school-level selection bias if intervention and comparison schools are well matched on observable characteristics and if there are no other important unobservable factors that influence take-up of the programme. We compare outcomes between the two groups of schools. The success of this comparison is in part determined by the quality of the comparison group identified. Recent research by Weidmann and Miratrix (2020), which compared comparison groups created in this way with randomised control groups, found little trace of unobserved factors that might invalidate conclusions from such a QED.

However, addressing school-level selection bias in this way only gets us so far. Tutoring is a pupil-level intervention, and as TP and AM schools were able to decide themselves which pupils would take part, we need to consider pupil-level selection bias. Indeed, TP schools selected pupils on factors such as perceived likelihood to engage with and/or benefit from tuition. In AM, when selecting pupils, schools typically identified pupils who were below the expected level for their year, or whose 'learning loss' had been greatest since the start of the pandemic (Teach First, 2021). To counter this, the QEDs were designed to focus on identifying similar groups of pupils in the intervention and comparison schools that would potentially participate, based on their observable characteristics. By analysing pre-identified groups of pupils in intervention and comparison schools (such as pupils eligible for pupil premium), we would remove the pupil-level selection problem but as we see below, this approach has its own issues.

## 5.2 What were the challenges when recruiting to the evaluation samples?

Recruitment to evaluation samples was hampered by the disruptions in schools relating to the ongoing Covid-19 pandemic. We were only able to recruit an evaluation sample to TP and only in the primary phase.

TP secondary, AM primary and AM secondary evaluation samples were not large enough to proceed with the analyses.

For year groups where national assessment data was available, our analysis could use attainment data from all schools in England from the NPD (cancelled for Year 6, and proceeded with for Year 11 through exploratory analyses using TAGs). However, for year groups where this data is not available, we had to recruit a sample of schools to the evaluation and obtain their permission to use assessment providers' tests routinely conducted in these schools.

### *5.2.1 TP evaluation samples*

Recruiting schools to the TP evaluation samples (Research Champions) was challenging in the context of a pandemic, particularly after the further period of school closures to most pupils in January–March 2021 when schools were focused on supporting their pupils.

For primary schools, whilst we recruited sufficient numbers of intervention schools (participating in TP), the matched comparison schools we approached (matched on a range of school characteristics) showed lower interest. We supplemented the sample with comparison schools drawn from schools that expressed interest in the programme but had not (yet) taken it up, and succeeded in recruiting enough schools. A challenge was whether the group, now made up of schools that were similar to TP schools in terms of characteristics and schools that were similar in terms of their interest in tutoring (but not necessarily in terms of characteristics), would make a robust comparison for the TP group. Checks on the samples from the TP group and the non-TP comparison group indicate that overall the samples are similar in terms of observable school characteristics.

With secondary schools, we found that the number of schools that met our eligibility criteria was much lower than anticipated. Coupled with a low response to our recruitment generally at secondary level – particularly from comparison schools – we had to drop the secondary phase Research Champion sample.

### *5.2.2 AM evaluation samples*

The evaluation samples for AM were planned to be derived from an RL data-feed provided to Teach First and DfE, by identifying AM schools and suitably matched non-AM schools in this data-feed. The samples would cover primary English (Years 1–6), primary maths (Years 1–6), secondary English (Years 7–10) and secondary maths (Years 7–10). However, there were both insufficient schools agreeing to a new opt-in consent arrangement with RL and insufficient AM schools using RL data to proceed with these evaluation samples. These evaluation samples were dropped from the AM study.

## 5.3 What were the challenges in identifying groups of comparison pupils that were similar to the pupils participating in TP and in AM?

The study designs for both TP and AM included a number of estimators of impact, in order to approximate comparison groups of pupils that would have been identified for tutoring had the programme been available in their school.

### *5.3.1 Estimating impact using pupil premium*

In TP, 46% of the pupils selected for tuition were eligible for pupil premium. For AM, 89% of the schools met Teach First's priority criteria, which is based on the proportion of children living in income deprived families (IDACI) and whether the school is in an area of chronic and persistent underperformance (AEA). The remaining 11% of schools had an above average proportion of pupils eligible for Pupil Premium (Teach First, 2021). In AM, 49% of the pupils who took part were eligible for PP or FSM.

This estimator focused on all PP-eligible pupils in the relevant year groups in both the intervention and comparison schools. Of PP-eligible pupils, the proportion that was selected for TP and for AM was relatively low (see below), meaning that any effect of tutoring would be diluted as the estimates of impact include all PP-eligible pupils (in the relevant year groups) and not only those participating in the programmes.

One of the key intentions of the TP programme was to focus was on supporting disadvantaged pupils, including those eligible for PP, FSM or those identified by schools as having an equivalent need for support. Participating schools had discretion to identify which of their pupils they felt would benefit from additional support, and decide whether face-to-face or online tuition would be more suitable for them in the current environment. Whilst there was no target for PP participation in TP (2020/21), PP-eligible pupils were expected to be a key participant group: in addition to the stated aim of the programme to support disadvantaged pupils, in a pilot of online tutoring in the summer term of 2020, over 60% of targeted learners were eligible for pupil premium (Marshall *et al.*, 2021)[6]. However, the proportion of PP-eligible pupils taking part in TP was not as high as in the pilot. According to the monitoring data provided by TPs, 46% of the

[6] In that pilot (https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/online-tuition-pilot), 79% of the primary school learners working with Action Tutoring and Tutor Trust, and 60% of the secondary school learners who worked with Action Tutoring, The Access Project and Tutor Trust were receiving PP.

tutored pupils for whom data was provided were PP-eligible.[7] When matched to NPD, we found that 43% of pupils taking part in TP were PP-eligible[8] (i.e., a similar proportion to the data provided by TPs).

In AM, 49% of the pupils who took part were eligible for PP or FSM (Teach First, 2021), and 46% in our analysed Y11 sample were eligible for PP[9].

Both the TP and AM evaluations were designed to estimate impact using PP-eligibility (i.e., to analyse the progress of PP-eligible pupils in schools deploying tutoring compared to the progress of PP-eligible pupils in schools not involved in NTP). This was an important analysis, given both TP and AM had a specific objective to help disadvantaged pupils whose learning had particularly suffered during the course of the pandemic. Whilst schools had discretion over which pupils would receive tutoring, we anticipated that, due to the focus on supporting disadvantaged pupils and the guidance provided to schools, a high proportion of PP-eligible pupils would be selected. For TP, this was based on briefing materials for schools and TPs, and the percentage of PP-eligible pupils that participated in a pilot project at the start of the pandemic (Marshall *et al*., 2021); and similar assumptions were also made for the evaluation of AM.

Hence, one of the research questions (RQs) in each of the TP and AM evaluations focuses on all PP-eligible pupils in the year groups involved as a way of identifying would-be participants and avoiding selection bias. Any effect of tutoring would be diluted amongst all the PP-eligible pupils (as not all would take part), but this was outweighed against being able to identify a key group of potential participants in both intervention and comparison groups.

However, of PP-eligible pupils, the proportion actually selected to do TP was low (less than 25% across each of our different samples)[10], so any effect of tutoring would be highly diluted amongst the PP-eligible pupils, as the level of dilution means that the analysis is on a group where the majority did not participate in TP. In response to this we include a sensitivity analysis for the TP Year 11 evaluation sample, whereby we re-run the analysis, restricting the sample to schools that target a majority (50% or 70%) of PP-eligible pupils for tuition, thereby reducing the level of dilution.

Similarly, of PP-eligible pupils in Year 11, a small proportion were taking part in AM in Year 11, and smaller proportions still in English and in maths, which means the estimated effect in the analysis will be diluted as the analysis includes all Year 11 PP-eligible pupils from these schools.

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

In order to manage or counter the possibility of variation in pupil selection for TP and AM, we also included estimates on all pupils, and a predicted participation analysis in the study plans.

### *5.3.2 Using observable characteristics to predict participation*

The impact evaluations for TP and AM were unable to identify exactly who would participate in the programmes using observable characteristics – and hence unable to create a comparison group based on predicted participation.

In order to counter the dilution issue in the analysis that uses all PP-eligible pupils (outlined above), the evaluation also sought to identity exactly who would participate in the programme and to focus on them rather than on all pupils of a certain type. Hence, one of the RQs in both the TP and AM evaluations aimed to model predicted participation using

[7] Note, this is 46% of 184,597 pupils for whom TPs provided data on PP eligibility (PP=yes/no). There was missing/blank or withdrawn data for 20% of the 232,892 pupils for this field on pupil premium. (We could not assume the missing/blank meant no, as sometimes a whole school's data was missing for this field.)

[8] That is, of the 188,250 pupils that were matched to NPD, 80,986 were eligible for pupil premium. Note that when pupil data provided by TPs was matched to the National Pupil Database (NPD), via the Office for National Statistics (ONS) Secure Research Service (SRS), 43% of the 188,250 pupils that could be matched were identified as in receipt of Free School Meals (FSM) (the NPD does not record Pupil-Premium eligibility in one field; FSM was the most relevant field for this purpose). Note, this and the data cited in footnotes 9 and 10, pertains to data accessed through the SRS. No other SRS held data is presented in this paper.

[9] Note, this and the data cited in footnotes 8 and 10, pertains to data accessed through the SRS. No other SRS held data is presented in this paper.

[10] Note, this and the data cited in footnotes 8 and 9, pertains to data accessed through the SRS. No other SRS held data is presented in this paper.

observable characteristics from administrative data sources, in order to create a comparison group that has similar observable characteristics.

However, in both TP and AM the models aiming to predict participation had very poor predictive power and this analysis was not able to proceed. This aligns with the finding from the TP IPE that schools were using a variety of unobservable factors to select pupils for tutoring, including their perceptions about how likely the pupil was to benefit from and engage with tutoring (which are not 'visible' in the dataset). Similarly in AM, schools typically identified pupils who they felt were 'most behind' based on classroom assessments and also on lower engagement during remote learning (Teach First, 2021, p.22) – again, these are unobservable characteristics in the datasets available for matching. This means that trying to identify similar comparison groups in the TP and AM data is very difficult.

The issues 5.3.1 and 5.3.2 above attempt to find pre-identified groups of pupils in each of the intervention and comparison groups in order to mitigate pupil-level selection bias in school (or year-group)-level analyses. Given the limitations of both of these and in addition to the pre-specified analyses in the TP study plan, we also explored an approach to comparing the progress of tutored children with their peers in TP schools. Although vulnerable to selection bias, this approach could be a way of producing a pupil-level estimate of the effect of tutoring at scale. The results would need to be treated with caution.

## 5.4 Are the samples large enough to detect an effect?

The samples on which we are reporting each have sufficient numbers of schools and pupils to detect an effect – however, the calculations do not take into account the level of dilution seen in practice.

Previous research suggests a weighted mean effect size of 0.37 for one-to-one tuition and 0.31 for small-group tuition.[11] However, we expect 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 TPs. The samples recruited to the TP Research Champion samples were smaller than anticipated. This led to the secondary phase Research Champion sample being dropped from the analysis. Although the samples for the primary school Research Champions were smaller than planned, some of the assumptions behind the calculations have been updated since the study plan and show that we have sufficient power to detect an effect of 0.125. The Year 11 TP analysis is able to detect an effect of 0.04. However, the calculations do not take into account the level of dilution seen in practice.

Note, the AM study was designed to detect an impact of 0.07 for the Year 11 PP-eligible pupil analyses; the school sample sizes achieved for the RL primary and secondary phase samples were too small and were dropped from the AM study.

## 5.5 What outcomes are being used and what considerations are needed?

### *5.5.1 Year 11 analysis on the full population of schools involved (TP and AM)*

The usual summer exams process for Year 11 pupils could not go ahead as planned in summer 2021, and GCSE exams were determined by TAGs in 2021 – these have not been used as an outcome measure for an evaluation before, so we needed to run some checks to determine their suitability for our particular analyses. This is not any reflection or statement about the TAGs as an assessment or grade for pupils.

Due to the use of TAGs as an outcome measure, the Year 11 population analysis should be considered exploratory.

[11] There is a large body of evidence that one-to-one tutoring (EEF, 2021a) and small-group tuition (EEF, 2021b) are effective (with effect sizes of five months and four months respectively) – particularly where they are targeted at pupils' specific needs. Meta-analyses show positive impacts of tutoring on learning outcomes to the order of 0.3 standard deviations, and that tutoring can be particularly effective for disadvantaged pupils (Torgerson *et al*., 2018 and Dietrichson *et al*., 2017).

As mentioned above, the TAGs were a temporary assessment method introduced for 2021. The swift implementation of the TAGs and their nature (individual teachers and schools assessing their pupils which were then moderated) led to a number of unknowns while making this change to the evaluation. As outlined in the published study plans for TP and AM, we were concerned that the process for grade determination in the TAGs may mean that it would be difficult to detect any potential impact of the TP programme for the following reasons[12]:

- Consideration 1: That teacher-assessed GCSE grades may be distributed differently compared to previous years (in particular there may be differences around the grade 3/4 boundary).[13]
- Consideration 2: Knowledge/selection of pupils doing TP led to bias (conscious or unconscious, positive or negative) in the teacher assessed grades (TAGs).
- Consideration 3: There are uncertainties around whether the TAGs will reflect pupils' performance after the tutoring because schools may have used evidence from across the school year.
- Consideration 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.

We therefore carried out a number of pre-specified checks on the data, for example to look at the distribution of grades for different years, and different groups of schools (more information about the checks can be found in the TP and AM study plans). We did this to assess whether the TAGs would be a suitable outcome measure for this evaluation. Although the checks helped to determine whether we should proceed with the analysis, they are not 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 findings of these checks will be reported along with the results of the Year 11 analysis. Due to the use of TAGs as an outcome measure, the Year 11 population analysis should be considered exploratory.

### *5.5.2 Primary school Research Champions (TP only)*

In the primary school analysis for TP, we amalgamated a range of standardised assessments already used by schools into a single measure. By including different assessments, we are increasing measurement error, but this is partially mitigated by only selecting tests that have been standardised and aligned with the National Curriculum.

For the outcome measure used by the Research Champion sample we opted to accept a range of standardised assessments amalgamated into a single measure. This decision was so that we could use the tests already being used by schools – especially at primary schools, where standardised tests are more commonly used – and avoid additional testing burden. We note that by including different measures (assessments), we are increasing measurement error and muddying what domain of learning we are measuring. However, this was partially mitigated by only selecting tests that had been designed in alignment with the National Curriculum and that have been standardised.

## 5.6 What challenges were there in gathering data about participation in TP and AM?

The evaluations are based on a high volume of complex data collected from TPs and AMs, as well as directly from schools. The data requests were set up quickly at the start of the delivery, and had to be updated during the programme. This resulted in varying levels of completeness and quality of the data provided.

The TP evaluation uses data from a very large dataset, from multiple sources including: monitoring data about schools (6000+), pupils (230,000+) and tutors (26,000+) involved in the intervention provided by each of 33 TPs at multiple time-points; pupil-level assessment data from four different assessment providers; school-level characteristics; as well as

[12] This is outlined in more detail in the study plan.

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

data matched to the NPD and Get Information About Schools (GIAS). This created a number of considerations and challenges as follows:

- the quality and completeness of the data collected from TPs varied, and while there was some resource for following up with TPs, data was not always supplied or updated
- some of the data collected from TPs was added to the data request part-way through delivery – the main example being dosage data. This meant that TPs did not necessarily have systems set up in a way that could provide the requested data.

This has implications for all of the RQs as information about participation is used throughout the analysis. This is mitigated somewhat for the primary school analysis as the evaluators worked closely with the schools that signed up to ensure the quality of the data submitted was good. However, the same was not possible for the datasets concerning the whole population of TP participants, as this was a very large dataset with data provided by 33 TPs at multiple timepoints.

Similarly, the evaluation of AM relies on data supplied by AMs. However, not all AMs returned data (less than 70% did); the Evaluators are working only with data about pupils for whom information was provided (i.e., there will be pupils who took part in AM in 2020/21 for whom there is no evaluation data at all).

## 5.7 How did the TP evaluation take account of shifted delivery?

As delivery moved later in the academic year, more tutoring than originally expected took place after the end-point assessments. We had to ask TPs to provide additional information about the timing and amount of tutoring delivered.

In TP, tuition delivery shifted later in the academic year (due to the period of school closures to most pupils in January–March 2021) and tuition delivery extended into the summer holidays. Original timings meant that the bulk of tuition would have been completed before summer assessments (particularly before GCSEs and Year 6 national curriculum assessments, as well as other standardised tests administered by schools in other year groups). Later delivery meant that tutoring blocks were only partially complete by the time of the end-point assessment, and therefore the impact estimates that will be reported will be based on partial delivery for some pupils (as well as being subject to dilution, as outlined above).

In response, we collected dosage (number of hours) data from TPs to add into the analysis to compare outcomes of TP pupils that receive more or less tutoring (i.e., analysis within the TP group). The quality of this analysis depends on the quality and completeness of the dosage data and this varied within the dataset.

The AM programme data collection was already underway and specified prior to finalising the AM evaluation study plan, and so the ‘participation information’ being collected remained as planned by the programme (i.e., number of sessions, and first and last date of tuition). As in TP, the first and last dates would help to establish whether tutoring was available before (or after) the assessment cut-off date (11th June in the case of Year 11 TAGs).

# 6. Reporting

The TP volumes are designed to be viewed as a suite of reporting outputs, and will entail volumes on: the IPE, the impact evaluation for primary schools, the impact evaluation for Year 11, and an overarching synthesised summary and interpretation of key findings. Each volume will be reviewed by a technical panel at EEF and by external peer reviewers appointed by EEF. External peer reviewers will assign padlock ratings to estimates of impact to indicate how secure or 'confident' we can be in the findings, highlighting the analytical limitations/caveats relating to aspects such as attrition, MDES, and implementation fidelity. Padlocks will not be assigned to the Year 11 results based on TAGs due to the exploratory nature of the analysis. The caveats and considerations relating to the Year 11 analyses will be highlighted in the report.

For AM, there will be a report on the impact for Year 11 using TAGs, similarly outlining the caveats and considerations relating to the analyses.

The considerations set out in this paper will be important when interpreting the findings of this large-scale evaluation. The reports will include key messages for policy, tutoring organisations, schools and future research.

# References

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

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

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

Harland, J., Fletcher, L., Morton, C., Lord, P. and Styles, B. (2022). *Learning Recovery in Yorkshire and the Humber* [online]. Available at: https://www.nfer.ac.uk/media/4901/learning_recovery_in_yorkshire_and_the_humber.pdf [Accessed 6 May, 2022].

Marshall, M., Bury, J., Wilshart, R., Hammelsbeck, R. and Roberts, E. (2021). *The National Online Tuition Pilot*. [online]. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/online-tuition-pilot [Accessed 16 May 2022].

Nelson, J., Lynch, S., Sharp, C., 2021. *Recovery during a pandemic: the ongoing impacts of Covid-19 on schools serving deprived communities*. [online] Available at: <https://www.nfer.ac.uk/media/4614/recovery_during_a_pandemic_the_ongoing_impacts_of_covid_19_on_schools_serving_deprived_communities.pdf> [Accessed 5 August 2022].

Rose, S., Twist, L., Lord, P., Rutt, S., Badr, K., Hope, C. and Styles, B., 2021. *Impact of school closures and subsequent support strategies on attainment and socio-emotional wellbeing in Key Stage 1: interim paper 2*. [online] Available at: <https://educationendowmentfoundation.org.uk/public/files/Impact_of_School_Closures_KS1_-_Interim_Findings_Paper_2_-_July_2021.pdf> [Accessed 25 August 2022].

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

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

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

Education Endowment Foundation

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

hthttps://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn