A randomised controlled trial of the effectiveness of the Action Tutoring Programme
Statistical Analysis Plan

Education Endowment Foundation

Evaluator (institution): NFER

Principal investigator(s): Dr Stephen Welbourne

PROJECT TITLE	A randomised controlled trial of the effectiveness of the Action Tutoring Programme
DEVELOPER (INSTITUTION)	Action Tutoring
EVALUATOR (INSTITUTION)	National Foundation for Educational Research (NFER)
PRINCIPAL INVESTIGATOR(S)	Dr Stephen Welbourne
SAP AUTHOR(S)	Ruth Staunton, Stephen Welbourne
TRIAL DESIGN	Two-armed RCTs with randomisation at pupil-level within year group by school blocks. Primary population is Y11, secondary populations are Y10 and Y7. Nimble trial within main trial
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	Years 7, 10 and 11 (age 11-16) Key stages 3, 4
NUMBER OF SCHOOLS	62 at design 58 at randomisation
NUMBER OF PUPILS	5,208 (average 28 per year group, per school) at design 1,848 in Y11, 1,795 in Y10 & 2,028 in Y7 at randomisation
PRIMARY OUTCOME MEASURE AND SOURCE	Maths Attainment: GCSE maths fractional grades collected from schools (Year 11) Access Maths assessment (Years 7 and 10)
SECONDARY OUTCOME MEASURE AND SOURCE	Mathematical self-perceptions and enjoyment of maths scales from Maths and Me Survey Reduction in school absence: School absence data (authorised and unauthorised absence)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]		*N/A*

Table of contents

Introduction

Pupils from disadvantaged backgrounds have less access to the tools that support them to progress in school and reach their full academic potential. This results in an attainment gap at GCSE level. Just 52% of PP pupils achieved a GCSE grade 4 in maths in 2024 compared with 79% of other pupils[1]. The Action Tutoring small group tuition programme aims to help close the attainment gap in GCSE maths between pupils from disadvantaged backgrounds and other pupils, contributing to improved longer-term education, training and other life outcomes associated with improved grades. The Action Tutoring programme evaluated in this trial is a targeted secondary maths small-group tutoring programme that deploys volunteer tutors to work with pupils from socio-economically disadvantaged backgrounds who are at risk of achieving below grade 4 in GCSE maths.

This evaluation aims to establish whether receipt of the Action Tutoring programme has an impact on pupils' maths attainment. For the purpose of this evaluation, the programme will be delivered to pupils in Years 10 and 11 in the 2025-26 academic year. A variation of the programme, in an earlier stage of development, will be delivered in Year 7. Each school will select between 26 and 40 pupils in each year group. Year 11 pupils must start the year working at GCSE grades 3–5 in maths. Year 10 pupils must start the year working at GCSE grades 2–5 in maths. Year 7 pupils must have narrowly achieved or narrowly missed the expected standard in their maths SATs. At least 65% of pupils selected by the school to take part in the trial must be eligible for the Pupil Premium grant. All pupils must meet the attainment criteria, but schools can offer up to 35% of places to pupils who are not eligible for the Pupil Premium. Within each year group in each school, half of the nominated pupils will be randomly assigned to receive tutoring and half to a control group receiving usual practice.

The primary analysis focuses on Year 11 pupils and tests for a difference in fractional GCSE maths grade[2] between the intervention and control pupils. Secondary analyses explore impacts on maths attainment in the other year groups, in the free school meals (FSM) subgroup, by dosage, by mode of delivery and by prior attainment, and impacts on other outcomes comprising absence rates, mathematical self-perception and enjoyment of mathematics.

[1] https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-4-performance/2023-24

[2] Fractional grades will be calculated from the raw GCSE maths scores using the published grade boundaries to convert to a fractional grade, assuming a linear relationship between score and fractional grade within each adjacent pair of grade boundaries.

Design overview

Table 1 - Design overview table

Trial design, including number of arms		Two-arm pupil randomised RCTs. Primary population is Y11, secondary populations are Y10 and Y7.
Unit of randomisation		Pupil with year group by school blocks
Stratification variables (if applicable)		None
Primary outcome	variable	Maths attainment in Y11
	measure (instrument, scale, source)	Maths GCSE fractional grades, 0-9.99, raw scores received by schools then shared with NFER along with exam board and paper to allow conversion into fractional grades
Secondary outcome(s)	variable(s)	i. Maths attainment in Y10 ii. Maths attainment in Y7 iii. Absence iv. Mathematical self-perception v. Enjoyment of mathematics
	measure(s) (instrument, scale, source)	i. Access Mathematics Test, 60-150, Hachette Learning ii. Access Mathematics Test, 60-150, Hachette Learning iii. Weekly absence rates (both authorised and unauthorised absences included), 0-100%, recorded by schools then shared with NFER iv. Mathematical self-perception, Likert, endpoint Maths and Me survey v. Enjoyment of mathematics, Likert, endpoint Maths and Me Survey
Baseline for primary outcome	variable	Maths attainment for incoming Y11
	measure (instrument, scale, source)	Action Tutoring baseline assessment, 0-100%, Action Tutoring
Baseline for secondary outcome	variable	i. Maths attainment for incoming Y10 ii. Maths attainment for incoming Y7 iii. Absence iv. Mathematical self-perception v. Enjoyment of mathematics
	measure (instrument, scale, source)	i. Action Tutoring baseline assessment, 0-100%, Action Tutoring[3] ii. KS2 maths scaled score, 80-120, NPD

[3] As the baseline assessment is a core feature of Action Tutoring's delivery model and appears within the programme Theory of Change, the decision was taken to use this measure rather than requiring pupils to sit an additional assessment that is externally validated. This minimises adaptations to Action Tutoring's model and reduces the burden of assessment on pupils before tuition begins, to protect pupil experience and engagement at a critical point in the delivery. The psychometric properties of the measure have not been evaluated but the correlation with GCSE maths score is high at 0.69.

		iii. Average weekly absence rates prior to tutoring starting (both authorised and unauthorised absences included), 0-100%, recorded by schools then shared with NFER iv. Mathematical self-perception, Likert, baseline Maths and Me survey v. Enjoyment of mathematics, Likert, baseline Maths and Me survey

This study comprises three two-arm RCTs with randomisation at pupil level. The cohorts of the three trials are Year 11 pupils, Year 10 pupils and Year 7 pupils. Pupils randomised to the intervention arm are undertaking the Action Tutoring Programme. Pupils randomised to the control arm are experiencing business as usual. The primary outcome in this trial is GCSE maths fractional grade as the primary research question identifies Year 11 pupils. The analyses relating to Year 11 and Year 10 pupils form an effectiveness trial to evaluate the impact of the Action Tutoring Programme. The Year 7 elements of the trial are considered exploratory. All schools providing evaluation data will receive a financial payment of £500.

In addition, a nimble trial has been incorporated within the Year 10 and Year 11 intervention arms. NFER has formed a partnership with the Behavioural Insights Team (BIT) who are leading on this element of the project. The nimble trial is randomised at the tutor level and explores the impact of a social belonging intervention on pupils' attendance at tutoring sessions, motivation towards maths, and maths attainment. A statistical analysis plan for the nimble trial can be found in Appendix B.

Sample size calculations overview

We have presented sample size calculations for all three year groups. The primary analysis is for Y11 (Table 2). Since we only have one primary model, no adjustment for multiple comparisons has been made and the significance threshold has been maintained at 5%.

Table 2 – Sample size calculations for Y11 – primary outcome of the trial

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.148	0.166	0.163	0.186	0.149	0.168	0.164	0.186
Pre-test/ post-test correlations	level 1 (pupil)	0.69		0.69		0.69		0.69	
	level 2 (school)	0		0		0		0	
Intracluster correlations (ICCs)	level 2 (school)	0.10		0.10		0.10		0.10	
Treatment effect heterogeneity (omega)		1		1		1		1	

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
Alpha		0.05		0.05		0.05		0.05	
Power		0.8		0.8		0.8		0.8	
One-sided or two-sided?		Two-sided		Two-sided		Two-sided		Two-sided	
Average cluster size		28	21	18.2	13.65	31.86	23.90	20.71	15.53
Number of schools	intervention	62	56	62	56	58	52	58	52
	control	62	56	62	56	58	52	58	52
	total	62	56	62	56	58	52	58	52
Number of pupils	intervention	868	588	564	382	924	622	601	404
	control	868	588	564	382	924	621	600	404
	total	1736	1176	1128	764	1848	1243	1201	808

Table 3 – Sample size calculations for Y10 – a secondary outcome of the trial

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.152	0.170	0.167	0.189	0.154	0.173	0.169	0.191
Pre-test/ post-test correlations	level 1 (pupil)	0.69		0.69		0.69		0.69	
	level 2 (school)	0		0		0		0	
Intracluster correlations (ICCs)	level 2 (school)	0.11		0.11		0.11		0.11	
Treatment effect heterogeneity (omega)		1		1		1		1	
Alpha		0.05		0.05		0.05		0.05	
Power		0.8		0.8		0.8		0.8	
One-sided or two-sided?		Two-sided		Two-sided		Two-sided		Two-sided	
Average cluster size		28	21	18.2	13.65	30.95	23.21	20.12	15.09
Number of schools	intervention	62	56	62	56	58	52	58	52
	control	62	56	62	56	58	52	58	52
	total	62	56	62	56	58	52	58	52
Number of pupils	intervention	868	588	564	382	897	603	583	392
	control	868	588	564	382	898	604	584	393
	total	1736	1176	1128	764	1795	1207	1167	785

Table 4 – Sample size calculations for Y7 – a secondary outcome of the trial

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.158	0.178	0.175	0.201	0.157	0.176	0.172	0.196
Pre-test/ post-test correlations	level 1 (pupil)	0.60		0.60		0.60		0.60	
	level 2 (school)	0		0		0		0	
Intracluster correlations (ICCs)	level 2 (school)	0.11		0.11		0.11		0.11	
Treatment effect heterogeneity (omega)		1		1		1		1	
Alpha		0.05		0.05		0.05		0.05	
Power		0.8		0.8		0.8		0.8	
One-sided or two-sided?		Two-sided		Two-sided		Two-sided		Two-sided	
Average cluster size		28	21	18.2	13.65	35.0	26.2	22.7	17.0
Number of schools	intervention	62	56	62	56	58	52	58	52
	control	62	56	62	56	58	52	58	52
	total	62	56	62	56	58	52	58	52
Number of pupils	intervention	868	588	564	382	1014	681	659	443
	control	868	588	564	382	1014	681	659	443
	total	1736	1176	1128	764	2028	1362	1318	886

Table 5 - Expected attrition

	School level	Pupil level
Expected attrition (%)	10%	25%

We have allowed for a larger than typical pupil level attrition as this intervention is aimed at a population that is difficult to engage, and 25% is in line with Action Tutoring's experience of tutoring dropout rates for secondary school pupils.

Sample size calculations were undertaken using the *PowerUpR* package in R statistical software. ICCs are taken from the EEF power parameters guide[4]. The Year 11 value is taken from Table 1 – Key Stage 4, EEF studies, Maths. The Year 10 and Year 7 value is taken from Table 36 – median of the Maths rows. The pre-post correlation used for Year 11 and Year 10 is from Action Tutoring and represents the historic correlation between maths GCSE score and their baseline test score. The pre-post correlation used for Year 7 is the median of the Maths rows in Table 36 in the EEF power parameters guide. The treatment effect heterogeneity has been set to 1 in the absence of data as this is the most conservative value. The proportion of pupils eligible for FSM is set at 65% for these calculations. As described in the Introduction section of this document, at least 65% of the study pupils will be eligible for PP and given the large crossover in pupils between these two groups[5] and the fact that this is a minimum threshold for PP pupils in the trial, we believe 65% FSM to be an appropriate estimate for these calculations. We anticipate 25% pupil attrition (from unpublished Action Tutoring data) and 10% school attrition (from previous EEF trials run by NFER) so to achieve the MDES above we aimed to recruit 28 pupils per year group from 62 schools. Post recruitment at randomisation, we found that the target number of schools had not been met (58 out of 62) but the target average number of pupils per school has been exceeded (32 vs. 28). We have updated the MDES based on the number of schools and pupils randomised but have continued to use an estimate for the FSM percentage as this information will only become available through the NPD. The effect on the MDES was balanced by the two opposing influences (fewer schools but more pupils per school) and the MDES at randomisation of 0.168 (primary outcome, overall i.e. all pupils, adjusted for attrition) was only marginally different to the MDES at protocol stage of 0.166. All pupils remaining in the evaluation after attrition will be analysed i.e. no sampling will take place.

All combinations of year group and overall/FSM subgroup have 80% power to detect effects smaller than 0.201. The EEF Teaching and Learning Toolkit[6] describes effect sizes from 0.19 to 0.26 as equivalent to 3 months' progress or a 'moderate impact'.

Randomisation

The primary randomisation is a 50/50 pupil level randomisation within each year group in each school. This was completed for Y11 and Y10 in October 2025 and for Y7 in February 2026. No stratification is implemented so simple randomisation within each block was applied. Randomisation was done using R statistical software[7] and a seed was set for reproducibility. Code will be included in the appendix of the final report.

---

[4] https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/methodological-research-and-innovations/improving-power-calculations-in-educational-trials

5 27.4% of secondary students were known to be eligible for FSM in the 2024/25 academic year, PP was received for 27.5% of secondary students in the financial year up to March 2025 and all students eligible for FSM are entitled to PP

6 https://educationendowmentfoundation.org.uk/education-evidence/using-the-toolkits

7 R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

In addition to this randomisation, schools were randomised into one of two groups for the dosage analysis. Each group was offered different programme lengths in Years 10 and 7. Group 1 are completing a 20-week programme in Year 10 and a 10-week programme in Year 7. Group 2 are completing a 15-week programme in both Years 10 and 7.

Analysis

The main analyses will be intention-to-treat and will follow the October 2022 EEF Statistical Analysis Guidance[8]. It will not be possible to blind analysts to group allocation due to difference in data structure between groups i.e. tuition dosage only recorded for intervention pupils. All analyses will be done using R statistical software[7].

Research questions

Primary research question

RQ1: How effective is Action Tutoring at promoting maths attainment among Y11 pupils?

Secondary research questions

RQ2: How does Action Tutoring effect the proportion of Y11 pupils achieving grade 4 or above in their GCSE maths?

RQ3: How effective is the Action Tutoring programme at promoting maths attainment among Y10 pupils?

RQ4: How effective is the Action Tutoring programme at promoting maths attainment among Y7 pupils?

RQ5: How effective is the Action Tutoring programme at promoting maths attainment among disadvantaged pupils in each of the three year groups (a) Y11, b) Y10, and c) Y7)?

RQ6: How does the effectiveness of the programme at promoting maths attainment vary with dosage offered among a) Y10 and b) Y7 pupils?

RQ7: How does the effectiveness of the programme at promoting maths attainment vary with dosage received among Y11 pupils?

RQ8: How effective is Action Tutoring at reducing absence in Y7, Y10 and Y11?

RQ9a: How effective is Action Tutoring at improving pupils' mathematical self-perception in Y7, 10 and 11?

RQ9b: How effective is Action Tutoring at improving pupils' enjoyment of mathematics in Y7, 10 and 11?

8 https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

RQ10: Is there any evidence of difference in the effectiveness of Action Tutoring on Y11 pupils depending on the mode of delivery (Action Tutoring's models of delivery – face-to-face; standard online; hybrid; virtual online)?

RQ11: How does the effectiveness of Action Tutoring on GCSE outcomes vary for Y11s with different prior attainment (estimated GCSE grades of 3, 4 or 5)?

Primary outcome analysis

Measures

The primary outcome measure in this evaluation will be GCSE maths fractional grade. This will be calculated from the raw GCSE maths scores using the published grade boundaries to convert to a fractional grade (assuming a linear relationship between score and fractional grade within each adjacent pair of grade boundaries). The conversion of raw scores to fractional grades is necessary because we anticipate that schools may use different exam boards where the raw scores would not be comparable with each other. The raw GCSE maths scores will be collected from schools for all Year 11 pupils in the evaluation. Due to the evaluation timeline, the first score provided to the school will be used for analysis. Re-marks will not be included. GCSE maths exams occur soon after the end of tutoring delivery. The last week of Y11 delivery is the week commencing 27th April and the 2026 GCSE exam window is from 4th May to 26th June. The baseline measure will be the Action Tutoring baseline assessment. This was administered by Action Tutoring's Programme Coordinators to pupils in Year 11 in both the intervention and control arms before randomisation and before tutoring commenced. The test is marked automatically in Action Tutoring's CRM. As the baseline assessment is a core feature of Action Tutoring's delivery model and appears within the programme Theory of Change, the decision was taken to use this measure rather than requiring pupils to sit an additional assessment that is externally validated. This minimises adaptations to Action Tutoring's model and reduces the burden of assessment on pupils before tuition begins, to protect pupil experience and engagement at a critical point in the delivery. The psychometric properties of the measure have not been evaluated but (from unpublished Action Tutoring analysis) the correlation with GCSE maths score is high at 0.69 (938 observations across 2022/23 and 2023/24 academic years).

Analysis

RQ1: How effective is Action Tutoring at promoting maths attainment among Y11 pupils?

The primary outcome measure of GCSE maths fractional grade will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 11 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the GCSE maths fractional grade for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.

- β_2 is the coefficient for the baseline score.
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

Secondary outcome analysis

Measures

For RQ2, the outcome measure will be a binary variable indicating whether the pupil achieved grade 4 or above in GCSE maths. The GCSE maths grade will be collected from schools for all Year 11 pupils in the evaluation.

For RQs 3, 4, 5b, 5c, 6a and 6b, the outcome measure will be Access Mathematics Test scores. Access Mathematics Tests[9] are a validated measure of maths attainment. The measure is present in the EEF attainment measures database[10], designed for our target age group, UK standardised and curriculum aligned. This outcome has been chosen to assess the theory of change short term outcome "Overall improvement in attainment in maths, including greater mathematical fluency and mastery of high-value exam topics" in year groups where no national test is expected.

NFER test administrators will administer these tests to all Year 10 (AMT Form 3) and Year 7 (AMT Form 1) pupils. NFER will not inform test administrators of pupil's randomisation allocation. For Year 10 students this will be after their tutoring block has ended, while for Year 7 there is a slight overlap with the last week of tutoring to allow assessments to take place prior to the end of the summer term. NFER will liaise directly with schools, at the end of the spring term, to co-ordinate dates for NFER Test Administrator assessment visits. NFER will provide some guidance, and schools will be asked to consider what works best for them and their pupils in terms of administration (this could be during the school day during their usual tutoring slot or a different time) and confirm to NFER their preferred assessment date and time.

For RQ8, the outcome measure will be the average % weekly absence (all absences i.e. authorised and unauthorised absences will both be included) during the period of tutoring. This will be collected from schools for all pupils in the evaluation.

For RQ9, the outcome measures will be the Maths and Me survey, consisting of two scales: mathematical self-perceptions (RQ9a) and enjoyment of mathematics (RQ9b). The Maths and Me survey is a validated measure present in the EEF SPECTRUM database[11].

The endpoint Maths and Me survey will be administered online by NFER Test Administrator for Year 7 and 10 pupils (after their end assessment) or by school staff for Year 11 pupils (who do not do an end assessment). For Year 10 and 11 students this will be after their tutoring block has ended, while for Year 7 there is a slight overlap with the last week of tutoring to allow surveys to take place prior to the end of the summer term.

For RQs 5a, 7, 10 and 11, the primary outcome measure of GCSE maths fractional grade will be used.

[9] https://www.hachettelearning.com/assessment/access-mathematics-tests

[10] https://educationendowmentfoundation.org.uk/measures-database/access-mathematics-tests-1-and-2-second-edition1

[11] https://educationendowmentfoundation.org.uk/measures-database/math-and-me-survey

For RQs 2, 3, 5a, 5b, 6a, 7, 10 and 11, the baseline measure will be the Action Tutoring baseline assessment (see primary outcome analysis).

For RQs 4, 5c and 6b, the baseline measure will be Key Stage 2 maths scaled score[12]. This will be accessed through the National Pupil Database (NPD) for all Year 7 pupils in the evaluation.

For RQ8, the baseline measure will be the average % weekly absence for a three week period of the academic year. This period will be before tutoring starts and will be the same period for all schools. It will occur in September/October 2025 for Year 11 and Year 10 and in January/February 2026 for Year 7. All absences will be included i.e. both authorised and unauthorised, in line with DfE's methodology for calculating metrics such as persistent absence. This will be collected from schools for all pupils in the evaluation.

For RQ9, the baseline measures will be the Maths and Me survey[13] consisting of two scales: mathematical self-perceptions (RQ9a) and enjoyment of mathematics (RQ9b). For Years 10 and 11, the baseline Maths and Me survey was administered online before tutoring commenced by Action Tutoring Programme Coordinators (alongside the baseline assessment) and then scored/analysed by an NFER analyst. For Year 7, the survey will be administered by schools as they did not complete the baseline assessment (Key Stage 2 scores are used as a baseline measure).

Analysis

For all research questions where Year 10 or Year 7 pupils are noted as included in the analysis population, this will be regardless of the length of programme that they are exposed to.

For all models including free school meals (FSM) as a variable, this will be defined as a binary variable indicating if the pupil has been recorded as eligible for FSM on any census day in the last six years, sourced from the NPD (EVERFSM_6_P).

RQ2

How does Action Tutoring effect the proportion of Y11 pupils achieving grade 4 or above in their GCSE maths?

The secondary outcome of 'achieved grade 4 or higher in GCSE maths' will be used as the dependent variable in a logistic multilevel model with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 11 pupils included in the evaluation. The model can be represented as:

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij}$$

[12] Scaled score is preferred over raw score to account for any pupils who took their KS2 maths in any year other than immediately prior to the intervention year.

[13] http://journals.sagepub.com/doi/pdf/10.1177/0748175611418522

Where:

- p_{ij} is the probability that pupil *i* in school *j* achieves grade 4 or higher in GCSE maths.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in log odds between the two groups.
- β_2 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.

RQ3

How effective is the Action Tutoring programme at promoting maths attainment among Y10 pupils?

The outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 10 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Access Mathematics Test score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ4

How effective is the Action Tutoring programme at promoting maths attainment among Y7 pupils?

The outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-

by-treatment interactions in the random structure. The analysis population will be Year 7 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Access Mathematics Test score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the KS2 maths scaled score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ6a

How does the effectiveness of the programme at promoting maths attainment vary with dosage offered among Y10 pupils?

The outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with offered dosage as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random. The analysis population will be Year 10 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 DosageGroup_j + \beta_2 BaselineScore_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Access Mathematics Test score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the dosage group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score
- $DosageGroup_j$ is the dosage group (20 weeks or 15 weeks of tuition) that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ6b

How does the effectiveness of the programme at promoting maths attainment vary with dosage offered among Y7 pupils?

The outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with offered dosage as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random. The analysis population will be Year 7 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 DosageGroup_j + \beta_2 BaselineScore_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Access Mathematics Test score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the dosage group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score
- $DosageGroup_j$ is the dosage group (10 weeks or 15 weeks of tuition) that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the KS2 maths scaled score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ7

How does the effectiveness of the programme at promoting maths attainment vary with dosage received among Y11 pupils?

The outcome measure of GCSE maths fractional grade will be used as the dependent variable in a linear multilevel model with received dosage as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random. The analysis population will be intervention group Year 11 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 ReceivedDosage_{ij} + \beta_2 BaselineScore_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the GCSE maths fractional grade for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the received dosage representing the difference in Y for an increase of one unit of dosage.
- β_2 is the coefficient for the baseline score
- $ReceivedDosage_{ij}$ is the received dosage i.e. the number of sessions that pupil *i* in school *j* attended.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ8

How effective is Action Tutoring at reducing absence in Y7, Y10 and Y11?

The outcome measure of average % weekly absence will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and year group, and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 11, Year 10 and Year 7 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 YearGroup_{ij} + \beta_3 IntGrp_{ij}{:}\,YrGrp_{ij} + \beta_4 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the average % weekly absence for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient for the intervention group.
- β_2 is the coefficient for year group.
- β_3 is the coefficient for the intervention group by year group interaction.
- β_4 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $YearGroup_{ij}$ is the year group (11, 10 or 7) for pupil *i* in school *j*.
- $IntGrp_{ij}{:}\,YrGrp_{ij}$ is the intervention group and year group for pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the average % weekly absence for a three week period of the academic year before tutoring starts for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

A contrast between intervention groups will be estimated and reported for each year group from this model.

RQ9a

How effective is Action Tutoring at improving pupils' mathematical self-perception in Y7, 10 and 11?

The outcome measure of Maths and Me mathematical self-perception subscale will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and year group, and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 11, Year 10 and Year 7 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 YearGroup_{ij} + \beta_3 IntGrp_{ij}{:}\,YrGrp_{ij} + \beta_4 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Maths and Me mathematical self-perception subscale score for pupil *i* in school *j*.

- β_0 is the intercept.
- β_1 is the coefficient for the intervention group.
- β_2 is the coefficient for year group.
- β_3 is the coefficient for the intervention group by year group interaction.
- β_4 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $YearGroup_{ij}$ is the year group (11, 10 or 7) for pupil *i* in school *j*.
- $IntGrp_{ij}: YrGrp_{ij}$ is the intervention group and year group for pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the baseline Maths and Me mathematical self-perception subscale score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

A contrast between intervention groups will be estimated and reported for each year group from this model.

RQ9b

How effective is Action Tutoring at improving pupils' enjoyment of mathematics in Y7, 10 and 11?

The outcome measure of Maths and Me enjoyment of mathematics subscale will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and year group, and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 11, Year 10 and Year 7 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 YearGroup_{ij} + \beta_3 IntGrp_{ij}: YrGrp_{ij} + \beta_4 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Maths and Me enjoyment of mathematics subscale score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient for the intervention group.
- β_2 is the coefficient for year group.
- β_3 is the coefficient for the intervention group by year group interaction.
- β_4 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $YearGroup_{ij}$ is the year group (11, 10 or 7) for pupil *i* in school *j*.
- $IntGrp_{ij}: YrGrp_{ij}$ is the intervention group and year group for pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the baseline Maths and Me enjoyment of mathematics subscale score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.

- ϵ_{ij} is the residual error for pupil *i* in school *j*.

A contrast between intervention groups will be estimated and reported for each year group from this model.

RQ10

Is there any evidence of difference in the effectiveness of Action Tutoring on Y11 pupils depending on the mode of delivery (Action Tutoring's models of delivery – face-to-face; standard online; hybrid; virtual online)?

The outcome measure of GCSE maths fractional grade will be used as the dependent variable in a linear multilevel model with mode of delivery as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts. The analysis population will be intervention group Year 11 pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 ModeOfDelivery_{ij} + \beta_2 BaselineScore_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the GCSE maths fractional grade for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the mode of delivery.
- β_2 is the coefficient for the baseline score
- $ModeOfDelivery_{ij}$ is the mode of delivery (online, face to face, hybrid) for the tuition that pupil *i* in school *j* received.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

Subgroup analyses

Research questions RQ5a, RQ5b, RQ5c and RQ11 all represent subgroup analyses. The subgroup of interest for the RQ5 questions is pupils eligible for FSM, with the three sub questions representing Y11, Y10 and Y7 pupils. The subgroups of interest for RQ11 are pupils estimated (by their school) to be on track for grade 3, grade 4, or grade 5 in GCSE maths.

RQ5a

How effective is the Action Tutoring programme at promoting maths attainment among disadvantaged pupils in Y11?

The outcome measure of GCSE maths fractional grade will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 11 pupils included in the evaluation who are eligible for FSM. The model equations will be identical to the RQ1 model specification.

In addition, the outcome measure of GCSE maths fractional grade will be used as the dependent variable in a linear multilevel model with the intervention group by FSM eligibility interaction as predictors, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis populations for the model will be Year 11 pupils included in the evaluation. This model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 FSM_{ij} + \beta_3 IntGrp_{ij}: FSM_{ij} + \beta_4 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the GCSE maths fractional grade for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient for the intervention group.
- β_2 is the coefficient for FSM eligibility.
- β_3 is the coefficient for the intervention group by FSM interaction.
- β_4 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- FSM_{ij} is the FSM eligibility (EVERFSM_6) for pupil *i* in school *j* (Y or N/unclassified).
- $IntGrp_{ij}: FSM_{ij}$ is the intervention group and FSM eligibility for pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ5b

How effective is the Action Tutoring programme at promoting maths attainment among disadvantaged pupils in Y10?

The outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 10 pupils included in the evaluation who are eligible for FSM. The model equations will be identical to the RQ3 model specification.

In addition, the outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with the intervention group by FSM eligibility interaction as predictors, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis populations for the model will be Year 10 pupils included in the evaluation. This model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 FSM_{ij} + \beta_3 IntGrp_{ij}: FSM_{ij} + \beta_4 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Access Mathematics Test score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient for the intervention group.
- β_2 is the coefficient for FSM eligibility.
- β_3 is the coefficient for the intervention group by FSM interaction.
- β_4 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- FSM_{ij} is the FSM eligibility (EVERFSM_6) for pupil *i* in school *j* (Y or N/unclassified).
- $IntGrp_{ij}$: FSM_{ij} is the intervention group and FSM eligibility for pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ5c

How effective is the Action Tutoring programme at promoting maths attainment among disadvantaged pupils in Y7?

The outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis population will be Year 7 pupils included in the evaluation who are eligible for FSM. The model equations will be identical to the RQ4 model specification.

In addition, the outcome measure of Access Mathematics Test score will be used as the dependent variable in a linear multilevel model with the intervention group by FSM eligibility interaction as predictors, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis populations for the model will be Year 7 pupils included in the evaluation. This model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 FSM_{ij} + \beta_3 IntGrp_{ij}: FSM_{ij} + \beta_4 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the Access Mathematics Test score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient for the intervention group.
- β_2 is the coefficient for FSM eligibility.
- β_3 is the coefficient for the intervention group by FSM interaction.
- β_4 is the coefficient for the baseline score

- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- FSM_{ij} is the FSM eligibility (EVERFSM_6) for pupil *i* in school *j* (Y or N/unclassified).
- $IntGrp_{ij}: FSM_{ij}$ is the intervention group and FSM eligibility for pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the KS2 maths scaled score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.
- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ11

How does the effectiveness of Action Tutoring on GCSE outcomes vary for Y11s with different prior attainment (estimated GCSE grades of 3, 4 or 5)?

The outcome measure of GCSE maths fractional grade will be used as the dependent variable in three linear multilevel models with intervention group as a predictor, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis populations for the three models will be Year 11 pupils included in the evaluation who are estimated by their school for i) grade 3, ii) grade 4 and iii) grade 5 in GCSE maths. Each of the model equations will be identical to the RQ1 model specification.

In addition, the outcome measure of GCSE maths fractional grade will be used as the dependent variable in a linear multilevel model with the intervention group by prior attainment group interaction as predictors, controlling for baseline scores and accounting for clustering of pupils at school level as random intercepts as well as school-by-treatment interactions in the random structure. The analysis populations for the model will be Year 11 pupils included in the evaluation. This model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 PriorAttainmentGroup_{ij} + \beta_3 IntGrp_{ij}: PAGrp_{ij} + \beta_4 BaselineScore_{ij} + b_{0j} + b_{1j} InterventionGroup_{ij} + \epsilon_{ij}$$

Where:

- Y_{ij} is the GCSE maths fractional grade for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient for the intervention group.
- β_2 is the coefficient for prior attainment group.
- β_3 is the coefficient for the intervention group by prior attainment group interaction.
- β_4 is the coefficient for the baseline score
- $InterventionGroup_{ij}$ is the intervention group that pupil *i* in school *j* was randomly assigned to.
- $PriorAttainmentGroup_{ij}$ is the GCSE maths grade estimated by school *j* for pupil *i* (3, 4 or 5).
- $IntGrp_{ij}: PAGrp_{ij}$ is the intervention group and prior attainment group for pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the Action Tutoring baseline assessment score for pupil *i* in school *j*.
- b_{0j} is the random intercept for school *j*.

- b_{1j} is the random slope for intervention group in school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

Robustness check

In January 2026 it was identified that one school in the trial was also participating in another EEF evaluation that focused on improving maths ability in year 7 pupils. We consider that there could be a small interaction effect for pupils participating in both trials, so as a robustness check we will rerun the RQ4 analysis model but excluding the dual participating school. The size, direction and significance of the intervention group coefficient will be qualitatively compared to the same coefficient in the RQ4 model to assess any impact of participating in both trials.

Imbalance at baseline

Tables showing the balance between intervention groups for all randomised pupils and for the analysis populations will be reported (separately for Y11, Y10 and Y7). These tables will present the number and percentage of pupils or schools for the following characteristics: eligible for FSM, establishment type, urban/rural and Ofsted rating. They will present means and standard deviations for the following characteristics: pupil to qualified teacher ratio, school progress 8 score in the 2024/25 academic year and Action Tutoring baseline assessment score (or KS2 maths scaled score for Y7). The difference in the baseline assessment score between the intervention groups will be reported as an effect size. The baseline assessment scores will also be presented as histograms. School levels variables where category numbers are less than 3 and pupil level variables where category numbers are less than 10 will be censored according to statistical disclosure control practice.

Missing data

The number and proportion of pupils with missing GCSE maths fractional grades (primary outcome variable) will be reported. If the percentage of pupils missing this outcome variable is less than 5%, no further missing data analysis will take place. If the percentage of pupils missing is greater than 5%, a logistic multilevel model will be run with a binary outcome variable indicating missing GCSE maths fractional grade. This model will include the intervention variable as a predictor, along with the following pupil and school characteristics: gender, pupil premium status, establishment type, school size (number of pupils), geographical region, urban/rural. Any of the additional variables which demonstrate a significant association with missingness will be included as a covariate in a rerun of the RQ1 analysis as a sensitivity check.

Compliance

Compliance will be calculated at the pupil level. Pupils who have attended at least 10 of 20 tutoring sessions will be categorised as compliant. This benchmark has been used by Action Tutoring for its own impact reporting for several years and is based on attendance patterns as the programme has developed. Pupil participation at tutoring sessions will be collected by Action Tutoring and shared with NFER. A complier average causal effect (CACE) analysis will be undertaken using a two stage least squares instrumental variable approach. This will only be conducted for the primary analysis i.e. the Year 11 cohort.

For the first stage the compliance indicator will be regressed on the intervention group, together with the covariate from the primary analysis model (Action Tutoring baseline assessment score). This first stage linear regression model will be:

$$compliance_{ij} = \beta_0 + \beta_1 InterventionGroup_{ij} + \beta_2 BaselineScore_{ij} + \epsilon_{ij}$$

The compliance indicator is expected to take the value zero for all pupils in the control group (one-sided non-compliance).

For the second stage GCSE fractional maths grade will be regressed on each pupil's predicted compliance value, $\widehat{compliance}_{ij}$ (estimated from the first stage model), in the following linear regression model:

$$Y_{ij} = \beta_0 + \beta_1 \widehat{compliance}_{ij} + \beta_2 BaselineScore_{ij} + \epsilon_{ij}$$

The coefficient for predicted compliance β_1 in the second stage model is the CACE (complier

average causal effect) estimate for the effect of compliance on GCSE fractional maths grade. Results from both regression stages will be reported.

Intra-cluster correlations (ICCs)

The ICCs (school, treatment by school and residual) will be calculated as the proportion of the total model variance attributed to each level. These will be reported for the primary analysis model and for a null model with only the intercept as a fixed effect. Total variance is defined in the Effect size section below.

Effect size calculation

Effect sizes will be calculated by dividing the adjusted difference in means (e.g. β_1 from the primary analysis model) by the square root of the total variance from the null version of the primary analysis model i.e. a model with only intercept as the fixed effect.

$$ES = \frac{\beta_1}{\sqrt{\sigma_T^2}}$$

For models with a treatment random slope (including the primary analysis model), the total variance from the null version of the model will be calculated using the formula suggested in Singh et. al., 2022[14]:

$$\sigma_T^2 = \frac{N^T(\sigma_B^2 + \sigma_W^2 + \sigma_E^2 + 2\tau) + N^C(\sigma_B^2 + \sigma_W^2)}{N}$$

Where N is the average number of pupils per school, N^T and N^C are the average number of pupils per school in the treatment group (T) and control group (C) respectively, σ_B^2 is the between school intercept variance, σ_W^2 is the within school (residual) variance, σ_E^2 is the between school slope variance, and τ is the covariance between within school variance and between school intercept variance.

For models without random slopes, σ_E^2 and τ are assumed to be 0 and the equation simplifies to:

[14] Singh, A., Uwimpuhwe, G., Li, M., Einbeck, J., Higgins, S., & Kasim, A. (2022). Multisite educational trials: estimating the effect size and its confidence intervals. International Journal of Research & Method in Education, 45(1), 18-38. https://doi.org/10.1080/1743727x.2021.1882416

$$\sigma_T^2 = \sigma_B^2 + \sigma_W^2$$

For the binomial generalised linear mixed effects model, effects will be reported as odds ratios.

Transparency

All analyses detailed in this SAP along with randomisation and sample size calculations have been or will be subject to the NFER Centre for Statistics QA process. Briefly, all code will be reviewed by a statistician who was not involved in the analysis to check i) coding accuracy and ii) adherence to the statistical analysis plan.

Sample size calculation code is included as an Appendix A to this document. Randomisation code will be included as an appendix to the final report.

Appendix A: Power Calculation R Code

```
rm(list=ls())

library(PowerUpR)

library(ggplot2)

npupils_p=28

propFSM=0.65

npupils_r_11=1848/58

npupils_r_10=1795/58

schoolAttrition=0.1

pupilAttrition=0.25


par<-data.frame(Label=rep(c("Protocol, overall","Protocol, FSM","Randomisation, overall","Randomisation, FSM"),3)[1:10],

        Population=c(rep("Y11",4),rep("Y10",4),rep("Y7",2)),

        NSch_noAttrition=rep(c(62,62,58,58),3)[1:10],

        NSch_withAttrition=NA,


NumberOfStudentsPerSchool_noAttrition=c(npupils_p,npupils_p*propFSM,npupils_r_11,npupils_r_11*propFSM,npupils_p,npupils_p*propFSM,npupils_r_10,npupils_r_10*propFSM,npupils_p,npupils_p*propFSM),

        NumberOfStudentsPerSchool_withAttrition=NA,

        Outcome=c(rep("KS4 Maths",4),rep("Commercial Maths Test",6)),

        Baseline=c(rep("AT Maths Test",8),rep("KS2 Maths",2)),

        PrePostCorrelation=c(rep(0.69,8),rep(0.6,2)),

        ICC=c(rep(0.1,4),rep(0.11,6)))

par$NSch_withAttrition<-par$NSch_noAttrition*(1-schoolAttrition)

par$NumberOfStudentsPerSchool_withAttrition<-par$NumberOfStudentsPerSchool_noAttrition*(1-pupilAttrition)


par$MDES_noAttrition<-round(sapply(1:dim(par)[1],function(i){
```

```
mdes.bira2(rho2=par$ICC[i],omega2=1,r21=par$PrePostCorrelation[i]^2,n=round(par$Numbe
rOfStudentsPerSchool_noAttrition[i],2),J=round(par$NSch_noAttrition[i]))$mdes[1,1]

}),3)

par$MDES_withAttrition<-round(sapply(1:dim(par)[1],function(i){


mdes.bira2(rho2=par$ICC[i],omega2=1,r21=par$PrePostCorrelation[i]^2,n=round(par$Numbe
rOfStudentsPerSchool_withAttrition[i],2),J=round(par$NSch_withAttrition[i]))$mdes[1,1]

}),3)


par
```

Appendix B: Nimble Trial Statistical Analysis Plan

Programme Evaluator (institution):
Behavioural Insights Team

Principal investigator(s): Dr. Patrick Taylor

PROJECT TITLE	Action Tutoring Sense of Belonging Nimble Trial
DEVELOPER (INSTITUTION)	Behavioural Insights Team
EVALUATOR (INSTITUTION)	Behavioural Insights Team
PRINCIPAL INVESTIGATOR(S)	Dr Patrick Taylor
PROTOCOL AUTHOR(S)	Callum O'Mahony, Niall Daly, Lal Chadeesingh, Dr. Patrick Taylor, Lily Margaroli, Dr. Akanksha Vardani Quality assurance by Dr. Giulia Tagliaferri
TRIAL DESIGN	Two-arm, cluster randomised control trial
TRIAL TYPE	Nimble trial
PUPIL AGE RANGE AND KEY STAGE	Years 10 and 11 (age 14-16) Key stage 4
NUMBER OF SCHOOLS	62
NUMBER OF PUPILS	1568
PRIMARY OUTCOME MEASURE AND SOURCE	Sense of belonging: Two survey items added to the Maths and Me survey administered at endpoint
SECONDARY OUTCOME MEASURE AND SOURCE	Maths attainment: GCSE maths z-score (Y11); Access Maths test z-score (Y10)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]		*This is the first version of the SAP associated with the Action Tutoring Sense of Belonging nimble trial (hereafter mentioned simply as nimble trial)*

Table of contents

Introduction

This SAP sets out the analytic approach for evaluating the Action Tutoring Sense of Belonging nimble trial, which involves a short, 15-minute social-belonging activity delivered in Week 3 or 4 of Action Tutoring's maths tutoring programme for Year 10 and Year 11 pupils. The activity is divided into three parts where first, the pupils read messages from former Action Tutoring pupils, then, reflect on their own experiences, and lastly write a message for future pupils.

The primary aim of this intervention is to promote a sense of belonging within tutoring sessions, with further expected impacts on maths attainment, attendance, and mathematical self-perception. The activity is framed to empower pupils and reduce belonging uncertainty by reinforcing that early challenges in tutoring are common and temporary.

The SAP aligns with the design and logic model described in detail in Appendix C of the effectiveness Trial Protocol.

Design overview

The trial employs a two-arm, cluster randomised controlled design in which the tutor is the unit of randomisation. Pupils are nested within tutors, and tutors operate within schools, resulting in a three-level data structure. This choice reflects the operational structure of Action Tutoring, where tutors work with small groups of pupils, making pupil-level randomisation impractical and increasing the risk of contamination. Tutor-level randomisation ensures that all pupils within a tutor group receive the same trial condition. The vast majority of tutors work exclusively within a given school (at randomisation, four tutors worked across different schools). At the RCT planning stage, we estimated 523 to 784 tutors delivering the intervention across participating schools, accounting for school attrition of 10%.[15] Stratification is applied at the school level to ensure a balanced representation of intervention and control tutors within each school.

The control condition consists of business-as-usual delivery, where tutors continue with standard warm-up activities that do not include belonging elements. Tutors in this group will not receive the belonging activity materials and communication will be managed using separate mailing lists to prevent contamination. The intervention condition consists of the belonging activity delivered during the warm-up in Week 3 (or Week 4 if required).

The primary outcome is a composite measure of sense of belonging in tutoring sessions, based on two adapted Likert-scale items administered at the endpoint through the Maths and Me survey. Maths attainment is measured as the secondary outcome with attendance, and mathematical self-perception being the exploratory outcomes. Baseline data are collected for all outcomes where meaningful baseline measures exist. Baseline belonging is proxied by adapting the endpoint belonging items to the reference school rather than tutoring sessions, as pupils will not yet have started tutoring when baseline surveys are administered.

[15] At the time of writing, randomisation has taken place, with 645 tutors randomised.

Table 1: Trial Design

Trial design, including number of arms		Two-arm, cluster randomised control trial
Unit of randomisation		Tutor
Stratification variables (if applicable)		School
Primary outcome	variable	Sense of belonging
	measure (instrument, scale, source)	Instrument: Two survey items added to the Maths and Me survey administered at endpoint. The items, created by BIT researchers, are adapted from validated belonging scales (including Walton and Cohen's (2007) Belonging uncertainty scale, PISA's (2022) Computer-based student questionnaire). Scale: Two individual 6-point Likert scale questions, whose responses will be combined (by taking the mean score of both responses) to form a composite sense of belonging score.
Secondary outcome(s)	variable(s)	Maths attainment
	measure(s) (instrument, scale, source)	GCSE maths z-score (Y11); Access Maths test z-score (Y10)
Exploratory outcome(s)	variable(s)	1. Tutoring attendance 2. Mathematical self-perception
	measure(s) (instrument, scale, source)	1. Attendance data from Action Tutoring (measured as a fraction of sessions attended) 2. Mathematical self-perception scale in the Maths and Me survey
Baseline for primary outcome	variable	Sense of belonging
	measure (instrument, scale, source)	Instrument: Two survey items will be added to the Maths and Me survey administered at baseline. The items, created by BIT researchers, are adapted from validated belonging scales (including Walton and Cohen's (2007) Belonging uncertainty scale, PISA's (2022) Computer-based student questionnaire). Because pupils have not started tutoring at baseline, the items relate to sense of belonging in school, not tutoring. Scale: Two individual 6-point Likert scale questions, whose responses will be combined (by taking the mean score of both responses) to form a composite sense of belonging score.
	variable	Maths attainment

Baseline for secondary outcome	measure (instrument, scale, source)	Action Tutoring baseline assessments (Y10 and Y11)
Baseline for exploratory outcome	variable	1. Attendance at Maths tutoring sessions 2. Mathematical self-perception
	measure (instrument, scale, source)	1. Attendance before Week 3 (i.e. in weeks 1 and 2) (measured as a fraction of sessions attended) 2. Self-perception items in "Maths and Me" survey

Sample size calculations overview

Table 2: Sample size calculations

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.123	0.15	0.153	0.186				
Pre-test/ post-test correlations	level 1 (pupil)	0.69		0.69					
	level 2 (tutors)	NA		NA					
Intracluster correlations (ICCs)	level 2 (tutor)	0.4		0.4					
	level 3 (school)	NA[16]		NA					
Alpha		0.05		0.05					
Power		0.8		0.8					
One-sided or two-sided?		two-sided		two-sided					

[16] A separate school-level ICC was not incorporated into the power calculation because we expect the tutor-level clustering to capture the majority of the correlation among pupils within the same school, given that pupils are grouped within tutors and treatment is assigned at the tutor level. Any remaining school-level correlation is therefore expected to be small and unlikely to materially inflate the design effect.

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
Average cluster size		2.5		2.5					
Number of tutors	intervention	347	314	347	314				
	control	347	314	347	314				
	total	694[17]	628	694	628				
Number of pupils	intervention	871	588	566	382				
	control	871	588	566	382				
	total	1742	1176	1132	764				

	School level	Pupil level
Expected attrition (%)	10%	25%

Power calculations are based on the primary outcome only. Calculations were conducted in Stata 16.0 to estimate the Minimum Detectable Effect Size (MDES) under the expected trial design and sample constraints. The trial is a two-arm clustered RCT, with clustering at the tutor-group level. The available sample for the study is restricted by the sample size in the main evaluation.

The planned analysis assumes a total of 1,568 pupils after 10% school attrition, with an average tutor-group size of 2.5 pupils and 65% of pupils eligible for pupil premium. Separate MDES estimates are presented for the full sample and the FSM-eligible subgroup. We assume 25% pupil-level attrition (based on unpublished Action Tutoring data) and 10% school-level attrition (based on rates observed in previous EEF trials delivered by NFER).

The ICC is modelled across a range from 0.30 to 0.50 to illustrate sensitivity to alternative clustering structures. An ICC of 0.40 is used as the main design scenario for planning purposes and is what is presented in the table above. Sample size calculations for ICC values 0.3 and 0.5 are presented in Appendix B.1. The range for ICC is informed by three NTP nimble trials with a similar clustered design, in which the ICC for attendance outcomes was 0.37. A pre-post correlation of 0.69 is assumed, based on evidence of the NFER figure for maths attainments as per Action Tutoring, we expect the pre-post correlation for the Sense of Belonging variable to be along the same lines. Under these assumptions, the MDES in the main design scenario (ICC = 0.40) is:

[17] According to the Appendix C, 628 tutors are associated with 56 schools (after applying a 10 percent attrition assumption to an initial sample of 62 schools). Assuming proportionality between the number of schools and tutors, this implies approximately 628*62/56 ≈694 tutors in the absence of school-level attrition.

Full sample: 0.15 with the expected attrition.

FSM subgroup: 0.186 with the expected attrition.

All pupils in intervention and control tutor groups will be included in testing where possible. An updated table will be provided at randomisation, reporting the actual sample size and revised MDES using the same assumptions as at protocol stage.

At the time of writing, randomisation at the tutor level has been completed, with the achieved sample of 645 tutors. However, pupil numbers by trial arm are not yet available. Consequently, post-randomisation MDES values cannot yet be computed and will be updated once final pupil-level allocation data become available.

Analysis

The analyses described in this SAP follow the approach presented in Appendix C , with all models estimated on an intention-to-treat basis. All analyses account for the clustered nature of the trial using multilevel modelling. Outcomes are analysed only once, as the nimble trial collects no follow-up data beyond the endpoint surveys and assessments. Analyses will be conducted using complete cases only in either R or STATA software.

Primary outcome analysis

The primary outcome is pupils' sense of belonging in their tutoring sessions, measured at endpoint using two adapted Likert-scale items (refer to Table 1: Trial Design). These items capture pupils' comfort in tutoring sessions and their feelings of belonging uncertainty[18] and are combined into a composite score by taking their mean.

The primary outcome will be analysed using a linear multilevel model (LMM). The model will adjust for baseline belonging, which is measured using adapted versions of the endpoint items, but framed to refer to school rather than tutoring. Random intercepts will be included at both tutor-group and school levels to account for the hierarchical structure of the data. The intervention indicator will be entered as the primary predictor. If some tutors deliver sessions across more than one school[19], tutor and school effects will be specified as cross-classified rather than nested. As a robustness check, we will estimate a model in which tutor groups are treated as nested within schools (by defining school-specific tutor identifiers) and report whether the conclusions are materially unchanged.

A multilevel model is used because the trial data are hierarchically structured, with pupils clustered within tutor groups and tutors operating within schools. This approach appropriately accounts for the lack of independence between pupils taught by the same tutor.

The mathematical specification for the LMM regression model for pupil *i with* tutor *j* in school *s* is defined as:

$$Y_{ijs} = \alpha + \beta_1 * Treatment_j + \beta_2 * Baseline\_measure_{ijs} + u_s + v_j + \varepsilon_{ijs}, \quad (1)$$

[18] We define "belonging uncertainty" as the feeling that one may not fully belong or that belonging is conditional.

[19] We have a total of four tutors in the sample that teach at two different schools.

where: Y_{ijs} is the measure of the outcome at the endpoint point; $Treatment_j$ is the indicator variable for the treatment arm; $Baseline\_measure_{ijs}$ represents the outcome as measured at baseline; u_s is the random intercept associated with each school mean, v_j is the random intercept associated with each tutor mean and ε_{ijs} is the residual error term. The random effects are assumed to be independently normally distributed with mean zero.

Secondary outcome analysis

The secondary outcome is maths attainment. For Year 10 pupils, maths attainment is measured using the Access Maths assessment administered under exam-style conditions by trained NFER test administrators who are blind to the treatment condition. For Year 11 pupils, the measure of attainment will be GCSE maths fractional grade. The fractional grade will be calculated from the raw GCSE maths scores using the published grade boundaries to convert to a fractional grade.[20] The conversion of raw scores to fractional grades circumvents the issue in comparing raw scores when pupils may sit examinations across different exams boards or at different tiers (Foundation vs Higher).

Mirroring the specification used for the primary outcome, the impact on maths attainment will be analysed using a linear multilevel model. The model will include the intervention indicator as the primary predictor and adjust for baseline maths attainment using Action Tutoring baseline assessments. Random intercepts will be included at tutor-group and school levels to account for clustering. The mathematical specification for the LMM regression model for pupil *i with* tutor *j* in school *s* is defined as:

$$Y_{ijs} = \alpha + \beta_1 * Treatment_j + \beta_2 * Baseline\_measure_{ijs} + u_s + v_j + \varepsilon_{ijs}, \quad (2)$$

where: Y_{ijs} is the measure of maths attainment at the endpoint point; $Treatment_j$ is the indicator variable for the treatment arm; $Baseline\_measure_{ijs}$ represents the maths attainment as measured at baseline; u_s is the random intercept associated with each school mean, v_j is the random intercept associated with each tutor mean and ε_{ijs} is the residual error term. The random effects are assumed to be independently normally distributed with mean zero.

Exploratory outcome analysis

We will examine two exploratory outcomes: pupil attendance at tutoring sessions and mathematical self-perception.

Attendance will be measured as a fraction of sessions attended (for example, the proportion of sessions attended after week 4) to account for variation in the length of programme delivery across Year 10 cohorts. Mathematical self-perception will be constructed from relevant items in the Maths and Me survey.

Both outcomes will be analysed using the same model specification as the primary analysis: a linear multilevel model adjusting for baseline values, with random intercepts at the tutor and school levels and standard errors clustered at the tutor level. The mathematical specification for the LMM regression model for pupil *i with* tutor *j* in school *s* is defined as:

20

$$Y_{ijs} = \alpha + \beta_1 * Treatment_j + \beta_2 * Baseline\_measure_{ijs} + u_s + v_j + \varepsilon_{ijs}, \quad (3)$$

where: Y_{ijs} is either attendance or mathematical self-perception as measured at the endpoint point; $Treatment_j$ is the indicator variable for the treatment arm; $Baseline\_measure_{ijs}$ is the corresponding baseline value of the outcome; u_s is the random intercept associated with each school mean, v_j is the random intercept associated with each tutor mean and ε_{ijs} is the residual error term. The random effects are assumed to be independently normally distributed with mean zero.

Subgroup analyses

We will conduct a subgroup analysis focusing on FSM-eligible pupils. This analysis will replicate the primary model specification (Eq. 1) but will be estimated on the subsample of FSM-eligible pupils only. FSM eligibility will be defined using the EVERFSM_6_P indicator from the National Pupil Database, linked by NFER using standard procedures. The mathematical specification for the LMM regression model for FSM-eligible pupil *i with* tutor *j* in school *s* is defined as:

$$Y_{ijs} = \alpha + \beta_1 * Treatment_j + \beta_2 * Baseline\_measure_{ijs} + u_s + v_j + \varepsilon_{ijs}, \quad (4)$$

where: Y_{ijs} is pupil's sense of belonging as measured at the endpoint point; $Treatment_j$ is the indicator variable for the treatment arm; $Baseline\_measure_{ijs}$ is the corresponding baseline value of the outcome; u_s is the random intercept associated with each school mean, v_j is the random intercept associated with each tutor mean and ε_{ijs} is the residual error term. The random effects are assumed to be independently normally distributed with mean zero.

Additional analyses

No further robustness checks or additional analyses are planned for this trial. The analyses described – primary, secondary, exploratory, and subgroup – fully comprise the analytic plan for the nimble trial.

Longitudinal follow-up analyses

The nimble trial does not include any longitudinal follow-up beyond the endpoint measurements described in Appendix C . All outcomes—primary, secondary, and exploratory—are measured once only. For Year 11 pupils, the endpoint Maths and Me survey is administered by school staff either during tutoring weeks 19 and 20 or in the week immediately after. For Year 10 pupils, the same survey is administered four or nine weeks after programme completion, depending on whether their tutoring programme lasts 20 or 15 weeks. All analyses therefore relate exclusively to these single measurement points, and no additional follow-up is planned.

Imbalance at baseline

We will assess balance between the treatment and control groups at baseline to evaluate the success of the randomisation and to examine whether differential attrition may have introduced imbalance in the analysed sample.

We will study the baseline characteristics that are correlated with the endline outcomes measures and that are available at baseline. The baseline characteristics that we will include are:

- Pre-test pupil-level measures of primary and secondary outcomes, as described in the above sections
- Proportion of FSM-eligible pupils

Consistent with EEF guidance, we will not rely on statistical significance testing to assess baseline balance. Instead, we will report means and standard deviations for the treatment and control groups and calculate standardised effect sizes, defined as the difference in means divided by the pooled standard deviation.

Any notable differences in baseline characteristics between trial arms, particularly in the analysis sample, will be described and discussed, with consideration given to whether such differences may plausibly be attributable to differential attrition.

Missing data

The extent of missing data for the primary outcome (Sense of Belonging) and the key secondary outcome (Maths Attainment) will be reported by presenting the number and proportion of pupils with missing data for each outcome. Where available, known reasons for missingness (e.g. pupil absence, withdrawal from school, incomplete assessment) and their prevalence will also be described.

If the proportion of missing data for an outcome is 5% or less, no further missing data analysis will be conducted and the complete-case analysis will be reported. If the proportion of missing data for an outcome exceeds 5%, a missing data strategy will be implemented as described below.

We will assess whether missing data are plausibly Missing At Random (MAR) by:

- Creating an indicator for missingness of the outcome, and
- Estimating logistic regression models to test whether missingness is predicted by observed baseline characteristics used in the main analysis (including treatment assignment, baseline outcome measures/proxies, FSM status, and school-level characteristics).

If the missingness is predicted by any of the covariates (MAR), complete case analysis may yield biased estimates. We will run the complete case analysis with and without covariates and compare the results. If the results are similar across the two scenarios, we will proceed with a complete case analysis as it is likely to be unbiased but potentially underpowered.

If the results are dissimilar, the data is likely to be Missing Not At Random (MNAR). In this case, we will use Manski-type bounds: a pre-specified, resource-appropriate sensitivity analysis that provides a necessary interval for the true effect without resorting to high-resource modeling that is not proportionate to the constraints of the nimble trial. Specifically, we will assign missing outcomes the minimum feasible value and, separately, the maximum feasible value (based on outcome measurement scale), to obtain an interval for the treatment effect under extreme assumptions about unobserved outcomes. This provides a conservative range for the true effect in the presence of differential attrition without relying on imputation models.

Compliance

Compliance with the intervention will be assessed to describe implementation fidelity and to support analysis in the presence of non-compliance. Compliance will be measured at both the tutor and pupil levels. At the tutor level, two indicators will be used. The first is engagement with

the intervention delivery training materials, which will be shared with tutors in the treatment group via email ahead of Week 3. This will be measured using Google Analytics data on the total number of unique users who accessed the training materials, providing an overall indication of engagement with the materials.[21] The second tutor-level indicator is intervention delivery, defined as whether the intervention is delivered by the tutor in Week 3 (or Week 4). This will also be measured as a binary variable, with expected compliance of around 90 per cent in the intervention arm and 100 per cent in the control arm. Data for this measure will be collected through programme coordinator confirmation of whether intervention group tutors complete the activity in Week 3 or 4.

At the pupil level, compliance will be defined as completion of the belonging activity. This will be measured using evidence of activity completion, either via submitted SmartSurvey forms or via completed in-person worksheets, and/or through recorded attendance at the session in which the intervention was delivered. This measure will be constructed as a binary indicator (1 = compliant, 0 = non-compliant). The expected level of compliance at the pupil level is approximately 70 per cent in the intervention arm, closely aligned with attendance levels, and 95 per cent in the control arm.[22]

In addition to the ITT analysis for all outcomes, Complier Average Causal Effect (CACE) analysis will be undertaken for the primary outcome, Sense of Belonging in Tutoring Sessions. This analysis will use a two stage least squares (2SLS) method, where the randomised assignment to the intervention arm will act as an instrumental variable for pupil-level intervention completion.[23] The approach assumes that randomised assignment influences the outcome only through its effect on intervention completion and that there are no unobserved common causes of compliance and the outcome. The feasibility of this CACE analysis is contingent on the availability of pupil-level compliance data from completed worksheets and/or SmartSurvey submissions.

Pupil-level compliance will be used in the CACE analysis, while tutor-level compliance indicators will be used descriptively to improve reporting on the different stages of compliance and to understand where breakdowns in implementation occur in the event of non-compliance.

[21] The delivery training materials will be hosted on Action Tutoring's website as a PDF and the link will be shared with tutors. Engagement will be tracked using Google Analytics. Due to data limitations, it was not feasible to define this measure at the individual-tutor level.

[22] Because tutor assignment and class lists are centrally controlled, the risk of control pupils being exposed to the intervention is expected to be very low. Nevertheless, we assume 95 percent compliance in the control arm to allow for the small possibility of administrative or human error.

[23] We will use a standard 2SLS approach with linear first-stage and second-stage equations. To account for the hierarchical structure of the data, standard errors will be clustered at the tutor level and school fixed effects will be included at both stages.

Intra-cluster correlations (ICCs)

We will estimate the intra-cluster correlation coefficients (ICCs) for the pre- and post-test measures at the tutor level to quantify the similarity of outcomes within clusters (tutors) and to support interpretation of the clustered trial results.

To do so, we will employ an empty variance components model as follows:

$$Y_{ijs} = \alpha + u_s + v_j + \varepsilon_{ijs}, \qquad (2)$$

where individual i is taught by tutor j in school s, Y_{ij} is the outcome of interest (Sense of Belonging or Maths Attainment), v_j is a tutor-level random effect, u_s is a school-level random intercept included to account for the hierarchical structure of the data, and ϵ_{ijs} is the individual-level residual error. The random effects are assumed to be independently normally distributed with mean zero.

The tutor-level ICC will be estimated from this model using the equation:

$$\rho = \frac{var(v_j)}{var(v_j) + var(u_s) + var(\epsilon_{ijs})} \qquad (3)$$

Tutor-level ICCs will be re-estimated using the primary analysis model, which adjusts for baseline belonging and includes the treatment indicator, as specified in Equation (1) above.

Effect size calculation

For all continuous outcomes (Sense of Belonging, Maths Attainment, Mathematical Self-Perception, and the continuous measure of Tutoring Attendance), effect sizes will be calculated as Hedges' g. These will be estimated using the unconditional standard deviation, in line with the EEF guidance.

Hedges' g effect size will be calculated as follows:

$$g = J(n_1 + n_2 - 2)\, x \, \frac{\widehat{\beta_1}}{\hat{\sigma}} \qquad (4)$$

where:

- $\widehat{\beta_1}$ is the conditional estimate of the treatment effect recovered from the treatment coefficient in the relevant multilevel model (Eq 1),
- $\hat{\sigma}$ is the residual standard deviation from that multilevel model, and
- $J(n_1 + n_2 - 2)$ is the small-sample correction factor calculated as:

$$J(n_1 + n_2 - 2) = \frac{\Gamma(\frac{n_1 + n_2 - 2}{2})}{\sqrt{\frac{n_1 + n_2 - 2}{2}}\, \Gamma(\frac{n_1 + n_2 - 3}{2})} \qquad (5)$$

where $\Gamma(.)$ is the Gamma function, n_1 is the sample size of pupils in the control group and n_2 is the sample size in the treatment group.

If calculating $J(n_1 + n_2 - 2)$ as per Eq (5) proves computationally intractable, we will instead use the following approximation:

$$J(n_1 + n_2 - 2) \approx 1 - \frac{3}{4(n_1 + n_2) - 9} \qquad (6)$$

Ninety-five per cent confidence intervals (95% CIs) of the effect size will be estimated by inputting the upper and lower confidence limits of $\widehat{\beta_1}$ from the regression model into the effect size formula. All of these parameters will be made available in the report.

Appendix B.1

Table B1: Sample size calculations sensitivity analysis

		Protocol							
		OVERALL		FSM		OVERALL		FSM	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.118	0.143	0.146	0.177	0.129	0.157	0.16	0.194
Pre-test/ post-test correlations	level 1 (pupil)	0.69		0.69		0.69		0.69	
	level 2 (tutors)	NA		NA		NA		NA	
Intracluster correlations (ICCs)	level 2 (tutor)	0.3		0.3		0.5		0.5	
	level 3 (school)	NA		NA		NA		NA	
Alpha		0.05		0.05		0.05		0.05	
Power		0.8		0.8		0.8		0.8	
One-sided or two-sided?		two-sided		two-sided		two-sided		two-sided	
Average cluster size		2.5		2.5		2.5		2.5	
Number of tutors	intervention	347	314	347	314	347	314	347	314
	control	347	314	347	314	347	314	347	314
	total	694	628	694	628	694	628	694	628
Number of pupils	intervention	871	588	566	382	871	588	566	382
	control	871	588	566	382	871	588	566	382
	total	1742	1176	1132	764	1742	1176	1132	764

R Code for Power Calculations

```
 n_pupils <- c(1742, 1176, 1132,764)
cluster_size <- c(2.5)
prepost_corr <- c(0.69)
power <- 0.8
alpha <- 0.05
icc <- c(0.3,0.4,0.5)
std <- 1
# Create full factorial table
school_inputs <- expand.grid(
  n_pupils = n_pupils,
  cluster_size = cluster_size,
  icc = icc,
  prepost_corr = prepost_corr,
  power = power,
  alpha = alpha
)
# Powercals function
cluster_rct_table <- function(df) {
  df %>%
    rowwise() %>%
    mutate(
      # Total sample per arm
      pupils_per_arm = n_pupils/2,
      # Design effect with reduced cluster size
      DE = 1 + (cluster_size - 1) * icc,
      # Effective sample size
      eff_n = n_pupils / DE,
      pupils_per_arm_adj = eff_n / 2,
      # Unadjusted Cohen's h
      d_unadj = pwr::pwr.t2n.test(n1 = pupils_per_arm_adj,
                                  n2 = pupils_per_arm_adj,
```

```
                    d = NULL,
                    sig.level = alpha,
                    power = power,
                    alternative = "two.sided")$d,

    # # Adjust for  pre/post correlation
    # Calculate adjusted sample size based on pre-post correlation
    n_adj_corr = pupils_per_arm_adj / (1 - prepost_corr^2),

    # Calculate the adjusted Cohen's h using the new sample size
    d_adj = pwr::pwr.t2n.test(n1 = n_adj_corr,
                    n2=n_adj_corr,
                    d = NULL,
                    sig.level = alpha,
                    power = power,
                    alternative = "two.sided")$d,
      ) %>%
  ungroup()
}
school_results <- cluster_rct_table(school_inputs)
```