Independent evaluation of the Oxford Teaching Effective Early Mathematics and Understanding in Primary schools (TEEMUP) professional development programme: A two-armed cluster randomised controlled trial

Statistical Analysis Plan

Evaluator: University of York

Principal investigators: Dr Lyn Robinson-Smith, Hannah Ainsworth and Caroline Fairhurst

PROJECT TITLE	Independent evaluation of the Oxford Teaching Effective Early Mathematics and Understanding in Primary schools (TEEMUP) Professional Development Programme: A two-armed cluster randomised controlled trial
DEVELOPER (INSTITUTION)	Department of Education, University of Oxford
EVALUATOR (INSTITUTION)	York Trials Unit, University of York
PRINCIPAL INVESTIGATORS	Dr Lyn Robinson-Smith, Hannah Ainsworth (up to August 2022) and Caroline Fairhurst (from August 2022)
PROTOCOL AUTHORS	Hannah Ainsworth, Caroline Fairhurst, Jess Hugill-Jones, Heather Leggett, Katie Whiteside, Kalpita Baird, Louise Elliott, Imogen Fountain, Kerry Bell, Carole Torgerson, David Torgerson, Lyn Robinson-Smith
SAP AUTHORS	Kalpita Baird, Caroline Fairhurst
TRIAL DESIGN	Two-armed cluster randomised controlled trial with random allocation at the school level
TRIAL TYPE	Efficacy
CHILD AGE RANGE AND KEY STAGE	KS1 – 4-6 years: Cohort 1: Starting reception in September 2021, followed to the end of academic year 2023; Cohort 2: Starting reception in September 2022, followed to the end of academic year 2023
NUMBER OF SCHOOLS	Planned: 106 primary schools; Actual: 93 primary schools
NUMBER OF CHILDREN	Planned: 3,180 (average 15 per school per cohort); Actual: (estimated for Cohort 1) 1,583 children (average 17 per school). Cohort 2 to be recruited September 2022.
PRIMARY OUTCOME MEASURE AND SOURCE	**Child maths attainment at the end of Year 1 for Cohort 1:** British Ability Scales 3 Early Number Concepts (BAS3 ENC) by GL assessment.
SECONDARY OUTCOME MEASURE AND SOURCE	**Self-Regulation** Children's Self-Regulation and Social Behaviour Questionnaire (CSBQ), 3 self-regulation subscales: Cognitive, Behavioural and Emotional. Combined mean

	score. At the end of Reception and Year 1 for Cohort 1, and the end of Reception for Cohort 2. Self-Regulation Early Years Foundation Stage Profile (EYFSP) Early Leaning Goal (ELG) at the end of Reception for both cohorts. **Personal, Social and Emotional Development** CSBQ – 7 subscales: Sociability, Prosocial behaviours, Externalising problems, Internalising problems, Cognitive self-regulation, Emotional self-regulation, Behavioural self-regulation. Each subscale scored separately. At the end of Reception and Year 1 for Cohort 1, and the end of Reception for Cohort 2. Self-Regulation EYFSP ELG, Managing Self EYFSP ELG and Building Relationships EYFSP ELG. Combined. At the end of Reception for both cohorts. **Child maths attainment** BAS3 ENC at the end of Reception for Cohort 2. Number EYFSP ELG and Numerical Patterns EYFSP ELG. Combined. At the end of Reception for both cohorts. **Child general attainment at the end of Reception for both cohorts** All 17 EYFSP ELGs average total point score. Good Level of Development achieved. **Teacher Confidence: Maths** Adapted version of Chen et al.'s (2014) 'Early Math Beliefs and Confidence Survey'.

This analysis plan was written post-randomisation and prior to receipt of any outcome data and deals only with the statistical analysis of effectiveness for the main trial. This document has been written based on information in the study protocol version 1.1 dated 04.07.2023, published on the EEF website in which full details of the background and design of the trial are presented.

SAP version history

Any changes made to the protocol which impact on the SAP, and any changes made to the SAP after its initial publication, will be specified here. There are no such changes to note to date.

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	TBC	N/A. Creation of original document

Table of Contents

Introduction

TEEMUP is an evidence-based professional development (PD) programme developed by researchers at the University of Oxford. Nominated Reception (YR) and Year 1 (Y1) teachers will receive specialist training from the Oxford TEEMUP PD team in improving maths content/domain knowledge and how to support children's mathematics and self-regulation.

The TEEMUP PD allows teachers to:

- explore best practice in mathematics teaching,
- work together to support transitions into and across classrooms,
- effectively engage the children's home in their maths education,
- build their mathematical confidence, knowledge and understanding,
- explore novel techniques to strengthen children's self-regulation, and
- effectively self-evaluate, plan for improvement and monitor their own children's progress.

The primary goals of the TEEMUP PD are to:

- improve pupils' maths attainment at the end of YR and Y1
- improve pupils' personal, social and emotional development (PSED) and self-regulation at the end of YR and Y1.

Nominated teachers will be offered two full consecutive days of training followed by seven half day workshops, once a fortnight, allowing time between sessions to implement new ideas, as well as a final half day follow-up workshop in 2023. In addition to workshops, a minimum of three in school mentoring/coaching sessions will be provided and a dedicated website with PD resources and materials to support in-class teaching.

This two-armed cluster randomised controlled trial (RCT) with randomisation at the school level will evaluate the effectiveness of the TEEMUP PD programme on the maths development of children in YR and Y1 of primary school in England.

Two cohorts of children will be recruited to take part in the evaluation:

- Cohort 1 consists of YR children aged 4-5 years old in the academic year 2021-22 without significant SEND or EAL, and will be followed until the end of Y1 (June/July 2023) when they will be 5-6 years old.
- Cohort 2 comprises YR children aged 4-5 years old in the academic year 2022-23 without significant SEND or EAL, and will be followed until the end of YR (June/July 2023).

The primary analysis will compare outcomes for Cohort 1 at the end of Y1 to investigate the impact of up to two years of the intervention. Comparisons involving Cohort 2, who will have reduced exposure to the intervention, will be assessed as a secondary outcome.

At the outset of the evaluation, before randomisation, schools will nominate a minimum of two teachers, one from YR and one from Y1 (3 teachers welcome), who will participate in the TEEMUP PD, if their school is allocated to the intervention group. Teachers in intervention schools will receive TEEMUP PD and support over a 16-month period (Jan 2022 - May 2023). Changes to practice would be expected to build over this period and therefore the evaluation

seeks to investigate the impact of the TEEMUP PD on children in Cohort 1 (who, at the end of Y1, will have been taught by YR and Y1 teachers receiving TEEMUP PD) and Cohort 2 (who, at the end of YR, will have been taught by the nominated YR teachers at the end of receiving the full TEEMUP PD). Participating schools/teachers will be asked to (1) retain nominated YR and Y1 teachers in their respective year groups for the duration of the trial, and (2) keep participating children in classes taught by nominated YR and Y1 teachers.

The research questions are:

What is the impact of the TEEMUP PD, in comparison to usual teaching practice, on:

RQ 1. children's maths attainment at the end of Y1 as measured by the BAS3 ENC? [Cohort 1 only; primary outcome]

RQ 2. children's self-regulation and PSED as measured using the CSBQ at the end of YR and Y1? [Cohort 1; secondary outcome]

RQ 3. children's maths attainment at the end of YR as measured by the BAS3 ENC? [Cohort 2; secondary outcome]

RQ 4. children's self-regulation and PSED as measured using the CSBQ at the end of YR? [Cohort 2; secondary outcome]

RQ 5. children's EYFSP scores at the end of YR, including the ELGs of Mathematics, Self-Regulation, PSED and general development? [Both cohorts; secondary outcome]

RQ 6. the maths attainment of children who are eligible for free school meals (FSM) at the end of YR (Cohort 2) and Y1 (Cohort 1) as measured by the BAS3 ENC? [Both cohorts; secondary outcome]
For the purposes of this document, pupils eligible for FSM will be denoted as EVER6FSM. We will use the National Pupil Database (NPD) variable, EVERFSM_6_P, to identify these pupils.

RQ 7. nominated teacher's confidence in supporting children's maths development? [YR and Y1 Teachers; secondary outcome].

These research questions will be answered by analyses due to be conducted in Autumn/Winter 2023, and written up in a report due to be submitted to the EEF in early 2024.

Design overview

Table 1: Study design overview

Trial design, including number of arms		Two-armed cluster randomised controlled efficacy trial, 2 cohorts. Cohort 1 followed for 2 years: YR 2021-22 to Y1 2022-23 Cohort 2 followed for 1 year: YR 2022-23
Unit of randomisation		Primary schools
Minimisation factors		Percentage of pupils eligible for free school meals (EVER6FSM) in the school (latest available data) (2 levels: dichotomised at the median ≤16%; >16%) Percentage of pupils identified as having English as an Additional Language (EAL) in the school (latest available data) (2 levels; dichotomised at the median ≤8%; >8%) Geographical location (6 levels: Peterborough, Norwich, Newmarket/Bury St Edmunds, Milton Keynes, Oxford and Barnet)
Primary outcome	variable	Maths attainment at the end of Y1 (Cohort 1 only)
	measure (instrument, scale, source)	British Ability Scales 3 Early Number Concepts (BAS3 ENC) 0-35, GL Assessment. Collected by blinded evaluation team research assistants.
Secondary outcome(s)	variable(s)	**<u>Cohort 1 (YR 2021-22, Y1 2022-23)</u>** Self-regulation at the end of YR and end of Y1. PSED at end of YR and end of Y1. Routinely collected maths, self-regulation, PSED, and general attainment at the end of YR. **<u>Cohort 2 (YR 2022-23)</u>** Maths attainment at end of YR. Self-regulation at end of YR. Child PSED at end of YR. Routinely collected maths, self-regulation, PSED, and general attainment at the end of YR. **<u>Teachers</u>** Teacher confidence (in teaching children maths), during intervention and at the end of intervention (YR and Y1 teachers).

	measure(s) (instrument, scale, source)	**Maths Attainment** British Ability Scales 3 Early Numbers Concepts (BAS3 ENC) 0-35, GL Assessment. Collected by blinded evaluation team research assistants. **Self-Regulation** Children's Self-Regulation and Social Behaviour Questionnaire (CSBQ), 17-items yielding 3-self regulation subscales: Cognitive, Behavioural and Emotional. Collected by nominated YR and Y1 teachers. **PSED** CSBQ, 34-items yielding 7 subscales: Sociability, Prosocial behaviour, Externalising problems, Internalising problems, Cognitive self-regulation, Emotional self-regulation, Behavioural self-regulation. Collected by nominated YR and Y1 teachers. **Routinely collected maths, self-regulation, PSED, and general attainment at the end of YR.** Early Years Foundation Stage Profile (EYFSP) data collected by teachers at the end of Reception accessed from the National Pupil Database (NPD). • Maths (Number ELG and Numerical Patterns ELG, combined) • Self-Regulation ELG PSED (Self-Regulation ELG, Managing Self ELG and Building Relationships ELG combined) • General Attainment (All 17 EYFSP ELGs average total point score and whether Good Level of Development has been 'met' **Teacher confidence: Maths** 'Early Math Beliefs and Confidence Survey' Adapted by Chen et al. (2014). Only subscale: Confidence in Helping Children Aged 4-6 Learn Maths;
Baseline for primary outcome	**variable**	Maths attainment at start if YR (Cohort 1 only)
	measure (instrument, scale, source) **variable**	British Ability Scales 3 Early Number Concepts (BAS3 ENC) 0-35, GL Assessment. Collected by blinded evaluation team research assistants. **<u>Cohort 1 (YR 2021-22, Y1 2022-23)</u>**

		Self-regulation at start of YR. PSED at start of YR. **<u>Cohort 2 (YR 2022-23)</u>** Self-regulation at start of YR. PSED at start of YR. **<u>Teachers</u>** Teacher confidence (in teaching children maths) at baseline (YR and Y1 teachers).
Baseline for secondary outcome	measure (instrument, scale, source)	**Self-Regulation** Children's Self-Regulation and Social Behaviour Questionnaire (CSBQ), 17-items yielding 3 self-regulation subscales: Cognitive, Behavioural and Emotional. Collected by nominated YR and Y1 teachers. **PSED** CSBQ, 34-items yielding 7 subscales: Sociability, Prosocial behaviour, Externalising problems, Internalising problems, Cognitive self-regulation, Emotional self-regulation, Behavioural self-regulation. Collected by nominated YR and Y1 teachers. **Teacher confidence and beliefs: Maths** Adapted 'Early Math Beliefs and Confidence Survey' by Chen at al. (2014). Only subscale: Confidence in Helping Children Aged 4-6 Learn Maths.

For Cohort 1, at the start of the academic year 2021/22, recruited schools were asked to provide a list of pupils in YR at the school who were aged 4-5 years and being taught by the nominated YR teacher. Information sheets and withdrawal forms were sent to the parents/carers of these eligible pupils. Children for whom a withdrawal form was not received were considered eligible to pre-testing. At the start of the Autumn term 2022, we will ask recruited schools to undertake the same tasks for the new YR for Cohort 2.

Outcome measures (see also Tables 2 and 3)

British Ability Scales 3 Early Number Concepts (BAS3 ENC)

The primary outcome is maths attainment measured by the 30-item BAS3 ENC (Elliot and Smith, 2011) for Cohort 1 only. It is scored 0-35 (Raw Score) and a higher score indicates greater attainment. Further information on the administration of the BAS3 ENC and interpretation of its scores can be found in the trial protocol and are only provided in brief here.

Cohort 1 will complete the BAS3 ENC twice with an independent research assistant blind to group allocation, once at baseline in Oct/Nov 2021 and again for outcome assessment in Jun/Jul 2023. The administrator will record the raw scores for the test. This will be converted to a total score by a member of the trial team in accordance with the BAS3 scoring manual.

At baseline, we aimed to assess at least 15 eligible children (and more where possible) with the BAS3 ENC. In cases where a school had 15 or fewer pupils, all pupils were assessed at baseline where possible. If there were more than 15 eligible children in a school, purposive and random sampling of children to pre-test was performed as follows. A key priority for the funder of this trial, the EEF, is raising the attainment of disadvantaged children (i.e., those eligible for EVER6FSM); therefore, a sub-group analysis will be conducted to explore the impact of TEEMUP PD on children eligible for EVER6FSM. In order to ensure a sufficient sample size to conduct this analysis, in schools where there were 3 or fewer children eligible for EVER6FSM, all these children were selected for assessment and then the remaining eligible children were randomly ordered for assessment. In cohorts with more than 3 children eligible for EVER6FSM, 3 children eligible for EVER6FSM were randomly sampled from the EVER6FSM group for assessment, and then all remaining children (EVER6FSM and non-EVER6FSM) were randomly ordered for assessment. Research assistants (RAs) completing the BAS3 ENC were advised to work their way through the provided list, starting with the child who was first on the list (up to 3 children eligible for EVER6FSM appeared first), and continue until the assessment had been completed for at least 15 children. If a child was absent on the day of baseline BAS3 ENC testing, the RA assessed the next available child from the randomly ordered list, or the next available EVER6FSM child if one of the first three were absent. Once the first 3 assessments were complete, research assistants reverted to assessing children in order of the list. This process also served to prevent unconscious assessor bias during data collection.

The group of Cohort 1 children for whom baseline BAS3 ENC assessments are completed will form the randomised evaluation group. When RAs revisit the school to complete the post-test with Cohort 1, they will only complete assessments with children who were assessed at baseline.

For Cohort 2, the BAS3 ENC will be completed once with an RA, for outcome assessment in Jun/Jul 2023. The assessment sampling process as described above will be followed for outcome assessments for Cohort 2.

At post-test, RAs will spend two days in schools, assessing Cohort 1 and Cohort 2. Children that are absent on the first scheduled post-test assessment visit will be captured on the second, if they are in attendance. Pupil-level attrition has been accounted for in the sample size calculations. However, a further visit to the school will be considered on case-by-case basis and only if a significant number of children are missing on both previous assessment days.

Child Self-Regulation and Social Behaviour Questionnaire (CSBQ)

The 34-item CSBQ developed by Howard and Melhuish (2017) will be implemented to collect data on children's self-regulation and PSED overall. It yields seven subscales that all contain at least five items (some items are used in more than one scale):

1. Cognitive self-regulation (items 5,6,8,12,18)
2. Emotional self-regulation (items 2,10,11,14,23,26)
3. Behavioural self-regulation (items 7,13,15,29,30,31)

4. Sociability (items 1,4,9,16,22,27,32)
5. Prosocial behaviour (items 15,19,24,27,30)
6. Externalising problems (items 3,20,23,26,28)
7. Internalising problems (items 17,21,25,33,34)

For each item, the respondent is asked to evaluate the child's frequency of target behaviours on a five-point scale (1= 'not true' to 3 = 'partly true' to 5 = 'very true'). The items are either positively worded or negatively worded. The items are scored whereby the higher the child scores on these scales, the more they show these behaviours. For subscales 1-5, scores on negatively worded items are reversed prior to analysis. Higher scores for these subscales indicate a more favourable outcome. For subscales 6-7, negatively worded items are not reversed before analysis, therefore the higher the children score on these, the more they show Externalising and Internalising problems. The seven subscales scores are obtained by taking the average of the component item scores (first reversing any relevant score). The developers offer no guidance on how to handle missing item-level data for this instrument and so the subscales will only be scored if a valid response is provided to all items.

To assess self-regulation, we shall use a single overall index of children's self-regulatory capacities that represent the mean of the CSBQ's three subscales of cognitive, emotional and behavioural self-regulation, where all three have a valid summary score.

To assess PSED, we will consider each of the seven CSBQ subscales separately.

For Cohort 1, teachers will be asked to complete the CSBQ at the start and end of the 2021-22 academic year (Oct/Nov 2021 and Jun/July 2022) and again with the same children at the end of Y1 (Jun/Jul 2023). For Cohort 2, the participating YR teacher will be asked to complete the CSBQ at the beginning and end of the 2022-2023 academic year, for the relevant children in their cohort.

At baseline, for Cohort 1, nominated teachers were asked to complete the CSBQ for a minimum of 15 participating children. Where possible, teachers were asked to complete the CSBQ for all eligible children, and not just those that were assessed for BAS3 ENC. The same randomly ordered list as generated for the BAS3 ENC assessments was provided to nominated teachers to enable them to conduct the CSBQ. This same process will be followed for baseline CSBQ assessments for Cohort 2. Pupils with a valid CSBQ score will be post tested where possible.

Early Years Foundation Stage Profile

The EYFSP is an observational measure completed by teachers when children are in the Summer term of YR. EYFSP data will be obtained from the National Pupil Database (NPD), via the Office for National Statistics Secure Research Service (ONS SRS). The EYFSP measures 17 ELGs whereby the teacher assigns the child as being 'emerging' or 'expected' for each ELG (DfE 2021).

Mathematics Early Learning Goals

ELGs Number and Numerical Patterns will be combined and analysed as a categorical outcome (Expected level met for both ELGs).

Self-Regulation Early Learning Goal

This outcome will be analysed as a categorical variable (Expected level met).

Personal, Social and Emotional Development Early Learning Goals

ELGs Self-Regulation, Managing Self and Building Relationships will be combined and analysed as a categorical outcome (Expected level met for all three ELGs).

General Attainment/Development

The EYFSP provides a general measure of good level of development (GLD). This outcome will be analysed as a binary variable.

The average total point score for the 17 ELGs, each assigned a score of 1 = Emerging and 2 = Expected, will also be considered.

Teacher Confidence: Maths

Teacher confidence (in teaching children maths) will be assessed, for nominated YR and Y1 teachers in each school, using an adapted short survey 'Early Math Beliefs and Confidence Survey' by Chen et al (2014).

We will request for the survey to be completed by all nominated YR and Y1 teachers in each school. The survey will be completed at baseline in Sept/Oct 2021, in Jun/Jul 2022 and in Jun/Jul 2023. The original survey consists of three subscales: Beliefs about Children Aged 4-6 and Maths (5 items); Confidence in Helping Children Aged 4-6 Learn Maths (11 items); and Confidence in Own Maths Abilities (9 items). However only the second: Confidence in Helping Children Aged 4-6 Learn Maths will be used.

Teachers will be asked to rate their agreement with each item on a Likert scale, from 1=strongly disagree to 5=strongly agree. Each item is scored from one to five. Scores for items in the subscale will be summed to produce a summary score ranging from 11-55, and a higher score indicates greater confidence. The developers offer no guidance on how to handle missing item-level data for this instrument and so the scale will only be scored if a valid response is provided to all 11 items.

Table 2: Pupil-level outcome measures and associated baseline measures of prior attainment for Cohort 1

Outcome measure	End of academic year	Measure/instrument	Scoring	Outcome measure	Start of academic year	Measure/instrument	Scoring
Outcome				**Measure of prior attainment**			
Maths attainment	Y1	BAS3 ENC	0-35	**Maths attainment**	YR	BAS3 ENC	0-35
Maths attainment	YR	EYFSP Mathematics ELGs: Number, and Numerical Patterns	Dichotomous: 1=Expected on both ELGs; 0 otherwise	**Maths attainment**	YR	BAS3 ENC	0-35
Self-regulation	YR and Y1	CSBQ 3 subscales (total 17 items): Cognitive, Behavioural and Emotional	Mean of three subscale scores (0-5)	**Self-regulation**	YR	CSBQ 3 subscales (total 17 items): Cognitive, Behavioural and Emotional	Mean of three subscale scores (0-5)
Self-regulation	YR	EYFSP Self-regulation ELG	Dichotomous: 1=Expected; 0=Emerging	**Maths attainment**	YR	BAS3 ENC	0-35
PSED	YR and Y1	CSBQ 7 subscales: Sociability, Prosocial behaviour, Externalising problems, Internalising problems, Cognitive self-regulation, Emotional self-regulation, Behavioural self-regulation.	Consider each subscale score (0-5) separately	**PSED**	YR	Respective CSBQ subscale: Sociability, Prosocial behaviour, Externalising problems, Internalising problems, Cognitive self-regulation, Emotional self-regulation, Behavioural self-regulation.	0-5
PSED	YR	EYFSP PSED ELGs: Self-regulation, Managing Self; and	Dichotomous: 1=Expected on	**Maths attainment**	YR	BAS3 ENC	0-35

		Building Relationships	all three ELGs; 0 otherwise				
General attainment	YR	EYFSP all 17 ELGs	Average total point score, where 1=Emerging, and 2=Expected	**Maths attainment**	YR	BAS3 ENC	0-35
General attainment	YR	EYFSP GLD	Dichotomous 1= achieved at least the expected level for the ELGs in the prime areas of communication and language, physical development and PSEC, and the specific areas of mathematics and literacy; 0 otherwise	**Maths attainment**	YR	BAS3 ENC	0-35

Table 3: Pupil-level outcome measures and associated baseline measures of prior attainment for Cohort 2

Outcome measure	End of academic year	Measure/instrument	Scoring	Outcome measure	Start of academic year	Measure/instrument	Scoring
Outcome				**Measure of prior attainment**			
Maths attainment	YR	BAS3 ENC	0-35	**Maths attainment**	YR for Cohort 1	BAS3 ENC – school-level mean from Cohort 1	0-35
Maths attainment	YR	EYFSP Mathematics ELGs: Number, and Numerical Patterns	Dichotomous: 1=Expected on both ELGs; 0 otherwise	**Maths attainment**	YR for Cohort 1	BAS3 ENC – school-level mean from Cohort 1	0-35
Self-regulation	YR	CSBQ 3 subscales (total 17 items): Cognitive, Behavioural and Emotional	Mean of three subscale scores (0-5)	**Self-regulation**	YR	CSBQ 3 subscales (total 17 items): Cognitive, Behavioural and Emotional	Mean of three subscale scores (0-5)
Self-regulation	YR	EYFSP Self-regulation ELG	Dichotomous: 1=Expected; 0=Emerging	**Maths attainment**	YR for Cohort 1	BAS3 ENC – school-level mean from Cohort 1	0-35
PSED	YR	CSBQ 7 subscales: Sociability, Prosocial behaviour, Externalising problems, Internalising problems, Cognitive self-regulation, Emotional self-regulation, Behavioural self-regulation.	Consider each subscale score (0-5) separately	**PSED**	YR	Respective CSBQ subscale: Sociability, Prosocial behaviour, Externalising problems, Internalising problems, Cognitive self-regulation, Emotional self-regulation, Behavioural self-regulation.	0-5

PSED	YR	EYFSP PSED ELGs: Self-regulation, Managing Self; and Building Relationships	Dichotomous: 1=Expected on all three ELGs; 0 otherwise	**Maths attainment**	YR for Cohort 1	BAS3 ENC – school-level mean from Cohort 1	0-35
General attainment	YR	EYFSP all 17 ELGs	Average total point score, where 1=Emerging, and 2=Expected	**Maths attainment**	YR for Cohort 1	BAS3 ENC – school-level mean from Cohort 1	0-35
General attainment	YR	EYFSP GLD	Dichotomous 1= achieved at least the expected level for the ELGs in the prime areas of communication and language, physical development and PSEC, and the specific areas of mathematics and literacy; 0 otherwise	**Maths attainment**	YR for Cohort 1	BAS3 ENC – school-level mean from Cohort 1	0-35

Randomisation

Schools were randomised after child recruitment and baseline data collection had been completed in that school. A statistician at York Trials Unit randomised schools 1:1 to either the intervention arm (offered the TEEMUP PD programme) or the control arm (continue with usual provision for the duration of the evaluation).

A dedicated computer program, MinimPy (Saghaei and Saghaei, 2011), was used for randomisation via minimisation using the factors[1]:

- School geographic location – 6 levels: Peterborough, Norwich, Newmarket/Bury St Edmunds, Milton Keynes, Oxford and Barnet, for logistical reasons, to ensure a balanced spread of intervention and control schools in each area.
- School deprivation level – the percentage of pupils eligible for free school meals (EVER6FSM) in the school (latest available data; dichotomised at the median for the 95 schools who expressed interest in the trial, ≤16%, >16%) to ensure balance between the randomised groups, since this school characteristic and individual child deprivation may moderate outcomes.
- School English as an Additional Language (EAL) level – the percentage of pupils identified as having EAL in the school (latest available data; dichotomised at the median for the 95 schools who expressed interest in the trial, ≤8%, >8%) to ensure balance between the randomised groups, since this school characteristic and individual child EAL status may moderate outcomes.

Randomisation was carried out in batches (groups of schools that were ready to be randomised at that time) to avoid delays in programme induction and to maximise programme delivery for as many schools as possible. Naïve minimisation with base probability 1.0 was conducted (i.e., 1:1 deterministic minimisation). Naïve minimisation was deemed to be sufficient as the allocations were conducted in batches, rather than one-by-one prospectively, meaning predictability was not a concern and hence a random element was not required.

In total, 93 settings were randomised (47 Intervention, 46 control) in 5 ‘batches’.

[1] NB: baseline for primary outcome (BAS3 ENC) was not used in the minimisation as although this assessment was completed before randomisation, scores were not confirmed before randomisation (due to time taken to mark and verify scores). Baseline score will be included as a covariate in the analysis so it was not necessary to additionally specify this as a minimisation factor.

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	EVER6FSM	OVERALL	EVER6FSM
Minimum Detectable Effect Size (MDES)		0.21	0.31	0.22	0.33
Pre-test/ post-test correlations	level 1 (child)	0.6	0.6	0.6	0.6
	level 2 (class)	N/A	N/A	N/A	N/A
Intracluster correlations (ICCs)	level 2 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two
Average cluster size		15	3	17	3
Number of schools	Intervention	50	50	47	47
	Control	50	50	46	46
	total*	100	100	93	93
Number of pupils	Intervention	750	150	794	135
	Control	750	150	789	143
	total**	1500	300	1583	278

* the trial aimed to recruit 106 schools to allow for some school level attrition without comprising the MDES; ** figures in Randomisation columns are current estimates, which may change when scoring of BAS3 ENC is finalised.

From protocol

For the primary analysis, we will compare BAS3 ENC scores at the end of Y1 for Cohort 1 between the intervention and control groups, adjusted for baseline BAS3 ENC score measured when the children are at the start of YR. Therefore, for the sample size for the primary analysis, we make the following assumptions: a school-level intracluster correlation coefficient (ICC) of 0.15, 15 children per school (at randomisation); a baseline and outcome testing correlation of 0.6 and 1:1 allocation at school level. The ICC and pre to post-test correlation can be justified based on the following previous EEF-funded trials, though these do differ slightly in the age of the population to this trial. The 1stClass@Number evaluation in Year 2 children found an ICC of 0.22 for its primary outcome of the Quantitative Reasoning Test (Nunes et al., 2018) with a pre-post correlation of 0.29 (sample restricted to those struggling with maths) but 0.63 for the whole sample which better reflects our population; for the secondary outcome of Key Stage 1 Maths, the ICC was 0.15 with a pre-post correlation of 0.26 for the restricted sample but 0.63 for the whole sample. The Mathematical Reasoning evaluation in Year 2 children found an ICC of 0.11 for its primary outcome of the GL Assessment Progress Test in Maths with a pre-post correlation of 0.58 (Stokes et al., 2018). The Maths Champions trial in early years, in the year before children started primary school,

found an ICC of 0.17 for its primary outcome of the CEM ASPECTS assessment in Maths with a pre-post correlation of 0.59 (Robinson-Smith et al., 2018).

Based on 100 schools (i.e., 1500 children), we would have 80% power to show an effect size of 0.21 of a standard deviation between the control and the intervention groups in the primary analysis, allowing for 15% attrition at child level at post-test. The trial aimed to recruit 106 schools to allow for some school level attrition without comprising the MDES.

Based on the sampling strategy, we might conservatively assume that we will achieve an average of 3 EVER6FSM children per school (300 from 100 schools). Assuming a baseline and outcome testing correlation of 0.6, an ICC of 0.15 and 15% attrition at the child level, with 100 schools we would have 80% power to show an effect size of 0.31 in the EVER6FSM subgroup for Cohort 1.

At randomisation

The primary analysis will include Cohort 1. In total, 93 settings were randomised, from which, there are 1,583 participating pupils (Intervention, n=794; Control, n=789). NB. This figure is subject to change as data are finalised. This is an average of 17 pupils per school. Assuming a pre- and post-test correlation of 0.6, an ICC of 0.15, and 15% pupil-level attrition, the MDES with this sample size would be approximately 0.22.

Approximately 278 of the randomised pupils are eligible for EVER6FSM (average of 3 per school). With this sample size, under the same assumptions, the MDES would be approximately 0.33.

Analysis

Analysis will follow the EEF's (2018) most recent guidance[2]. The trial statistician will not be blind to group allocation.

Analysis will be conducted in STATA v17[3].

Analyses and summaries will be presented separately for the two cohorts. All analyses will be conducted on an intention to treat basis (ITT), where data are available, including all schools and pupils in the group to which they were randomised irrespective of whether or not they actually received the intervention, using two-sided tests at the 5% significance level. The EEF analysis guidance states that the ITT population should exclude any pupils and school that dropped-out after randomisation, but before allocation is revealed; we do not have any such cases in this trial.

A CONSORT diagram will be produced to show the flow of schools and children through the trial. The number of children identified as eligible for the evaluation and the numbers actually assessed at baseline and post-test will be reported with reasons for non-participation given where available.

The number of schools who return a completed Expression of Interest form (EOI), are identified as eligible, who complete a Memorandum of Understanding (MoU) and a Data

[2] Please see the Statistical Analysis Guidance.

[3] A later version of STATA may be used. The version used will be confirmed in the final report.

Sharing Agreement (DSA) and the number that are actually randomised and complete post-testing will be reported.

All outcome data will be summarised descriptively by trial arm for each assessment point. The correlation of outcome measures and measures of prior attainment will be presented with a 95% confidence interval (CI). Effect sizes based on the adjusted difference between the groups at the outcome assessment point will be presented as mean differences for continuous outcomes, and odds ratios and difference in proportions for dichotomous outcomes, with their associated 95% CI and p-value. Treatment effects will also be presented as (estimated) Hedges' g effect sizes.

Imbalance at baseline

School and pupil characteristics and outcome measures measured at baseline will be summarised descriptively by randomised group both as randomised and as analysed in the primary analysis. At school level the following data will be summarised: geographical location (Peterborough, Norwich, Newmarket/Bury St Edmunds, Milton Keynes, Oxford and Barnet), percentage of pupils ever eligible for EVER6FSM, percentage of pupils with English as an Additional Language, and whether or not the participating school is involved in the NCETM Maths Hubs Programme (data permitting). At child level, the following data will be summarised: gender, EVER6FSM status and EAL status, plus measures of prior attainment.

Continuous measures will be reported as a mean, standard deviation (SD) (and/or median, minimum and maximum) while categorical data will be reported as a count and percentage. No formal comparison of the baseline data will be undertaken, except for a comparison of the difference in prior attainment (BAS3 ENC and CSBQ scores, as appropriate) between the groups, reported as the Hedge's g effect size, with a 95% confidence interval (CI).

Primary outcome analysis

BAS3 ENC score will be analysed using a mixed effects linear regression model at the child-level. Group allocation, baseline BAS3 ENC score, and the minimisation factors (geographical location of school, EVER6FSM and EAL) will be included as fixed effects in the model, and school as a random effect. Robust standard errors will be specified to account for any potential heteroscedasticity.

Whilst EVER6FSM and EAL will be used as aggregate measures at the school level in the minimisation, pupil level indicators of EVER6FSM and EAL will be included in the analysis model, since these are more granular measures and so are likely to correlate better with the outcome than the school-level data.

Model equation:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 I_{Ai} + \beta_3 I_{Bi} + \beta_4 I_{Ci} + \beta_5 I_{Di} + \beta_6 I_{Ei} + \beta_7 FSM_{ij} + \beta_8 EAL_{ij} + \beta_9 I_{Gi} + u_i + y_{ij}$$

Y_{ij} = response (post-test BAS3 ENC score) of the j-th of n_i members of the i-th cluster (school), *i=1,…,m, j=1,…,n_i*

m = number of clusters (school)

n_i = size of cluster (school) i

x_{ij} = baseline BAS3 ENC score for j-th member of i-th cluster (school)

I_{Ai} = indicator variable for location of i-th school (1= Peterborough)

I_{Bi} = indicator variable for location of i-th school (1= Norwich)

I_{Ci} = indicator variable for location of i-th school (1= Newmarket/Bury St Edmunds)

I_{Di} = indicator variable for location of i-th school (1= Milton Keynes)

I_{Ei} = indicator variable for location of i-th school (1= Oxford)

FSM_{ij} = indicator variable for EVER6FSM status for j-th member of i-th cluster (school)

EAL_{ij} = indicator variable for EAL status for j-th member of i-th cluster (school)

I_{Gi} = indicator variable for group allocation of i-th cluster (school) (0=Control, 1=Intervention)

β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7, β_8, β_9 = fixed effect parameters

$u_i \sim N(0, \varphi_b^2)$ = setting-specific random effect and $\gamma_{ii} \sim N(0, \varphi_w^2)$ = individual-specific random effect

Model assumptions will be checked as follows: the normality of the standardised residuals will be checked using a qq plot. If the model assumptions are in doubt, a sensitivity analysis will be conducted in which transformations of the outcome and/or covariate data will be tried to improve the model fit.

Secondary outcome analysis

Secondary outcomes will be analysed in an exactly analogous way to the primary outcome, adjusting for the appropriate associated measure of prior attainment.

Cohort 1

CSBQ

CSBQ scores will be analysed via a mixed effects linear regression model incorporating both outcome time points for each child, adjusting for respective baseline CSBQ score, location, EVER6FSM and EAL indicators, group allocation, time and group by time interaction as fixed effects, and school and child as random effects to account for the repeated measures over time (using an unstructured covariance structure).

EYFSP

The EYFSP dichotomous measures (Mathematics ELGs, Self-Regulation ELG, PSED ELGs and GLD) will be compared using mixed-effects logistic regression at the child-level, adjusted for group allocation, baseline BAS3 ENC score, location, EVER6FSM and EAL as fixed effects, and school as a random effect. The treatment effect expressed as an adjusted odds ratio will be reported with a 95% CI and p-value. We will also present the unadjusted and adjusted (i.e. predicted, using the postestimation command *margins, dydx(allocation)*) percentage point difference between the two groups with a 95% CI (Ge et al. 2011), and convert the adjusted OR (and 95% CI limits) to an estimated Hedges' g effect size using the Cox index as follows (What Works Clearinghouse):

$$d_{Cox} = \omega[ln(OR)]/1.65$$

Where $\omega = \left[1 - 3/(4N - 9)\right]$ and *N* is the total sample size.

The continuous measure (average total point score for the 17 ELGs) will be analysed as described for the BAS3 ENC.

Cohort 2

BAS3 ENC

Individual baseline BAS3 ENC scores will not be available for Cohort 2. Therefore, we will consider the use of a lagged school-level measure of prior attainment for these children as follows. We will calculate the mean baseline BAS3 ENC score per school from Cohort 1 and calculate the correlation between this and the outcome for Cohort 2. We will conduct analyses with and without including this measure as a school-level covariate in the analysis for Cohort 2.

Maths attainment for children in the intervention group and those in the control group will be compared using a mixed effects linear regression model at the child-level. Group allocation, baseline BAS3 ENC score (school-level mean from previous year group), location, EVER6FSM and EAL will be included as fixed effects in the model, and school as a random effect. This analysis will be repeated omitting the BAS3 covariate.

CSBQ

CSBQ scores will be analysed via a mixed effects linear regression model adjusting for respective baseline CSBQ score, location, EVER6FSM and EAL as fixed effects, and school as a random effect.

EYFSP

The EYFSP dichotomous measures (Mathematics ELGs, Self-Regulation ELG, PSED ELGs and GLD) will be compared using mixed-effects logistic regression at the child-level, adjusted for group allocation, baseline BAS3 ENC score (school-level mean from previous year group), location, EVER6FSM and EAL as fixed effects, and school as a random effect. The treatment effect expressed as an adjusted odds ratio will be reported with a 95% CI and p-value. We will also present the unadjusted and adjusted (i.e. predicted, using the postestimation command *margins, dydx(allocation)*) percentage point difference between the two groups with a 95% CI (Ge et al. 2011), and convert the adjusted OR (and 95% CI limits) to an estimated Hedges' g effect size using the Cox index as follows (What Works Clearinghouse):

$$d_{Cox} = \omega[ln(OR)]/1.65$$

Where $\omega = \left[1 - {}^{3}/_{(4N-9)}\right]$ and *N* is the total sample size.

This analysis will be repeated omitting the BAS3 covariate.

The continuous measure (average total point score for the 17 ELGs) will be analysed as described for the BAS3 ENC.

Teachers

Responses to items in the confidence survey will be summarised descriptively by trial arm. The summary score will be compared between the two arms using mixed effects linear regression, adjusting for baseline score, the school level minimisation factors and pertinent

teacher level factors as fixed effects, and school as a random effect (plus teacher as a random effects to account for repeated measures where appropriate).

Subgroup analyses

For both cohorts, subgroup analyses looking at gender and EVER6FSM eligibility will be undertaken for the BAS3 ENC outcome. In addition, a subgroup analysis looking at whether schools are taking part in the National Centre for Excellence in the Teaching of Mathematics (NCETM) Maths Hubs Programme will be considered, dependent on the level of missing data for this factor.

The subgroup analyses will be conducted by including the factor and an interaction term between the factor and allocation in the primary analysis model. We shall also repeat the primary analysis within the subset of participants eligible for EVER6FSM.

Additional analyses

In sensitivity analyses, the analysis models for EYFSP Self-regulation and PSED outcomes will be repeated adjusting for baseline CSBQ scores for these domains, rather than baseline BAS3 ENC score, since they propose to measure the same domains and so we may anticipate a reasonable correlation between baseline score and outcome. The correlations will be calculated and reported.

A further sensitivity analysis will be included, to adjust for whether or not the participating school is involved in the NCETM Maths Hubs Programme for the BAS3 ENC outcome, in both cohorts, dependent on the level of missing data for this factor.

Baseline CSBQ scores for Cohort 2 will be collected between November 2022 and January 2023. Children's abilities may change over this time. Therefore, we will consider the proportion of baseline CSBQ data that were collected before and after Christmas, and present this by trial arm. In a sensitivity analysis for the Cohort 2 CSBQ analysis, we shall include an indicator for whether the data were collected before or after Christmas.

Missing data

The amount of missing baseline and outcome data will be summarised, and reasons for missing data will be explored and provided in the report where available. Where less than 5% of ITT pupils are missing from the primary analysis model, no further action will be taken. If the percentage of missing cases exceeds 5%, then multi-level logistic regression models will be used to model presence or absence of the primary outcome including all available pupil and school-level baseline data as fixed effects, and school as a random effect. Significant predictors and possible mechanisms for the missing data will be discussed in the report.

The impact of missing data on the primary analysis (if >5%) will additionally be assessed using multilevel imputation via the REACOM-Impute macro, which is compatible with Stata (http://www.bristol.ac.uk/cmm/software/realcom/imputation.html), including all available pupil and school-level baseline variables (school: location, percentage of pupils every eligible for EVER6FSM, percentage of pupils with English as an Additional Language; pupil: gender, EVER6FSM status, EAL status). This imputation procedure can account for the two-level (pupil and school) nature of the data.

A 'burn-in' of 10 will be used, which means that the first 10 iterations of the imputation are not used to allow the iterations to converge to a stationary distribution, and 30 imputed datasets

will be created. (The values of 10 and 30 are subject to the convergence of the model and other values may be used during analysis). The primary analyses will then be rerun within the imputed datasets and Rubin's rules (Rubin, 1987) will be used to combine the multiply imputed estimates.

Compliance

There will be a Complier Average Causal Effect (CACE) analysis, conducted on the primary outcome of maths attainment as measured by the BAS3 ENC on Cohort 1 and Cohort 2. These CACE analyses will aim to obtain a treatment effect estimate among 'compliers', which may differ from the primary ITT analysis. This particular trial does not differentiate between compliance and fidelity for the CACE analysis and seeks to capture information on both compliance and fidelity within one measure. The Implementation and Process Evaluation will seek to explore compliance and fidelity as separate constructs where possible.

As suggested by the EEF analysis guidance, two thresholds for compliance are defined to conduct two CACE analyses for **good** compliance and **at least minimal** compliance.

Cohort 1

Compliance will be measured at the school-level, since the impact of the intervention will depend on engagement of both the YR and Y1 teachers that the children in Cohort 1 will have been taught by. Each teacher in the intervention arm will be assessed for their compliance with the intervention. A school will be classed as having good compliance if they fulfil all of the following core criteria in Table 4.

Table 4: Cohort 1 CACE analysis GOOD compliance 'core' criteria

GOOD compliance 'core' criteria	Data collection by/from
TEEMUP trained YR and Y1 teachers complete 7 of the first 9 core training sessions, at least by watching recorded sessions. (NB. Attendance/watching final half day session is not required for compliance), and the TEEMUP trained YR teacher/s remains at the school and teaching Reception during the majority (>50%) of the 2021/2022 academic year and TEEMUP trained Y1 teacher/s remains at the school and teaching Y1 for the majority (>50%) of the 2022/2023 academic year.	Attendance at training collected by DT (Delivery Team) via attendance registers/training completion records for each school and shared with ET. Teachers and teacher changes collected by ET (Evaluation Team) directly from schools at the end of each academic year, and by DT through PD. Lists shared and cross referenced between the two teams
The school hosts 3 face to face visits from a mentor/coach, at least 2 of which teachers should be well prepared for (DT will define 'preparedness', which will consist of two elements: Ensuring there is time during the meeting to 1. review existing change plans and write/agree new ones; and 2. gather evidence of changes made relating to previous agreed actions and/or TEEMUP PD.	Collected by DT mentor records for each school and shared with ET

A minimum of 8 school logins to the online knowledge base over the course of the whole intervention period.	Automated data held by DT team shared with ET or self-reported data from teachers collected by DT and shared with ET.
>75% or at least 15 children, whichever is lower (e.g. 12 or more of 15, and 15 or more of 22) of children in the evaluation (Cohort 1 Reception children for whom a baseline BAS3 ENC assessment was conducted) move to a Year 1 class being taught maths by a TEEMUP trained Y1 teacher in the 2022/2023 academic year.	ET
School's TEEMUP mentor considers the school to have been 'good' compliers, i.e. the school can provide sufficient evidence of change in practice resulting from TEEMUP training/resources. The mentors will assess evidence of change in practice on a scale of 0-3 with 0=no change, 1=minimal change, 2=good change, and 3=excellent change. A school must score at least 2 to be classed as a good complier.	Collected by DT mentor records for each school and shared with ET

A school will be classed as having at least minimal compliance if they fulfil all the following criteria detailed in Table 5.

Table 5: Cohort 1 CACE analysis MINIMAL compliance criteria

MINIMAL compliance criteria	Data collection by/from
TEEMUP trained YR and Y1 teachers complete 5 of the first 9 core training sessions, at least by watching recorded sessions. (NB. attendance/watching final half day session is not required for compliance) and a TEEMUP trained Y1 teacher/s remains at the school and teaching Y1 for the majority (>50%) of the 2022/2023 academic year.	Attendance at training collected by DT (Delivery Team) via attendance registers/training completion records for each school and shared with ET. Teachers and teacher changes collected by ET (Evaluation Team) directly from schools at the end of each academic year, and by DT through PD. Lists shared and cross referenced between the two teams
The school hosts 2 face to face visits from a mentor/coach, at least 1 of which teachers should be well prepare for (DT will define 'preparedness' which may include , class cover arranged, an appropriate meeting place organised, read through questions provided by DT prior to meeting and prepared to answer them).	Collected by DT mentor records for each school and shared with ET
A minimum of 4 school log-ins to the online knowledge base over the course of the whole intervention period.	Automated data held by DT team shared with ET or self-reported data from teachers collected by DT and shared with ET.

>50% or at least 11 children, whichever is lower (e.g. 8 of 15 children, or 11 of 22) of children in the evaluation (Cohort 1 YR children for whom a baseline BAS3 ENC assessment was conducted) move to a Y1 class being taught maths by a TEEMUP trained Y1 teacher in the 2022/2023 academic year.	ET

Cohort 2

Compliance will be measured primarily at the teacher-level, since the impact of the intervention will depend on engagement of the Reception teacher only for Cohort 2. Each Reception teacher in the intervention arm will be assessed for their compliance with the intervention.

A YR teacher will be classed as having good compliance if they fulfil all of the core criteria in Table 6 and the school's TEEMUP mentor considers the school to have been at least 'good' compliers, i.e. the school can provide sufficient evidence of change in practice resulting from TEEMUP training/resources. .

Table 6: Cohort 2 CACE analysis GOOD compliance 'core' criteria

GOOD compliance criteria 'core' criteria	Data collection by/from
TEEMUP trained YR teacher/s completes 7 of the first 9 core training sessions, at least by watching recorded sessions. (NB. attendance/watching final half day session is not required for compliance), and the TEEMUP trained YR teacher/s remains at the school and teaching Reception for the majority (>50%) of the 2022/2023 academic year.	Attendance at training collected by DT via attendance registers/training completion records for each school and shared with ET Teachers and teacher changes collected by ET directly from schools at the end of each academic year, and by DT through PD. Lists shared and cross referenced between the two teams
The school hosts 3 face to face visits from a mentor/coach, at least 2 of which the YR teacher should be well prepared for (DT will define 'preparedness' which may include class cover arranged, an appropriate meeting place organised, read through questions provided by DT prior to meeting and prepared to answer them). It is acceptable for these visits to be conducted with different teachers (if there is a change in YR teacher in the school between 2021/2022 and 2022/2023), provided at least 2 are conducted with a TEEMUP trained teacher.	Collected by DT mentor records for each school and shared with ET
A minimum of 8 school log-ins to the online knowledge base over the course of the whole intervention period.	Automated data held by DT team shared with ET or self-reported data from teachers collected by DT and shared with ET.

School's TEEMUP mentor considers the school to have been 'good' compliers, i.e. the school can provide sufficient evidence of change in practice resulting from TEEMUP training/resources. The mentors will assess evidence of change in practice on a scale of 0-3 with 0=no change, 1=minimal change, 2=good change, and 3=excellent change. A school must score at least 2 to be classed as a good complier.	Collected by DT mentor records for each school and shared with ET

A YR teacher will be classed as having at least minimal compliance if they fulfil all of the criteria as detailed in Table 7.

Table 7: Cohort 2 CACE analysis MINIMAL compliance criteria

MINIMAL compliance criteria	Data collection by/from
TEEMUP trained YR teacher/s completes 5 of the first 9 core training sessions, at least by watching recorded sessions. (NB. attendance/watching final half day session is not required for compliance) and the TEEMUP trained YR teacher/s remain at the school and teaching Reception for the majority (>50%) of the 2022/2023 academic year.	Attendance at training collected by DT via attendance registers/training completion records for each school and shared with ET Teachers and teacher changes collected by ET directly from schools at the end of each academic year, and by DT through PD. Lists shared and cross referenced between the two teams
	Collected by DT via attendance registers/training completion records for each school and shared with ET
The school hosts 2 face to face visits from a mentor/coach, at least 1 of which YR teacher should be well prepare for (DT will define 'preparedness' which may include class cover arranged, an appropriate meeting place organised, read through questions provided by DT prior to meeting and prepared to answer them). It is acceptable for these visits to be conducted with different teachers (if there is a change in YR teacher in the school between 2021/2022 and 2022/2023) as long as at least one of these mentor meetings is with the trained YR teacher.	Collected by DT mentor records for each school and shared with ET
A minimum of 4 school log-ins to the online knowledge base over the course of the whole intervention period.	Automated data held by DT team shared with ET[RL4] or self-reported data from teachers collected by DT and shared with ET.

Two CACE analyses (Dunn, Maracy and Tomenson, 2005) for each cohort will be conducted for the BAS3 ENC outcome defining compliance of the schools as a dichotomous variable in the two ways described above. These analyses will use a Two Stage Least Square (2SLS) approach with group allocation as the instrumental variable for the compliance indicator, with cluster standard errors to account for clustering at the school level. CACE analyses will be conducted at the pupil-level. Results for the first stage (which predicts the compliance indicator using the treatment allocation as instrumental variable alongside all other covariates

included in the second stage) will be reported alongside i) the correlation between the instrument and the endogenous variable; and ii) a F test.

Intra-cluster correlations (ICCs)

The intra-cluster correlation coefficient (ICC) associated with school for the outcomes (both pre and post-test where available) will be presented alongside a 95% CI. The ICC at post-test will be computed for the analysis model, and also for an empty model (i.e. one without covariates). The ICC at pre-test will be calculated for a linear model with pre-test as the outcome and setting as a random effect.

Effect size calculation

Effect sizes will be calculated by dividing the adjusted mean difference between the intervention and control group (accounting for baseline measures and the minimisation factors) by the pooled unconditional standard deviation obtained from the model run without these covariates. A 95% CI for the effect size will be calculated by dividing the 95% confidence limits for the adjusted mean difference by this same denominator. All parameters used in these calculations will be provided in the final report.

$$\mathrm{ES} = \frac{(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}}{sd_{pooled}}$$

where, $(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}$ denotes the difference in means between trial groups adjusting for pre-test score and the minimisation factors, from the multilevel analysis model; and sd_{pooled} denotes the pooled, unconditional standard deviation of the two groups (square root of the sum of the within- and between-cluster variances).

References

Chen, J.-Q., McCray, J., Adams, M. and Leow, C., 2014. A survey study of early childhood teachers' beliefs and confidence about teaching early math. Early Childhood Education Journal, 42(6), pp.367–377.

Department for Education (2021). Early years foundation stage profile 2022 handbook. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1024319/Early_years_foundation_stage_profile_handbook_2022.pdf (Accessed: 24 March 2022)

Dunn, G., Maracy, M. and Tomenson, B. (2005) 'Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods', *Statistical Methods in Medical Research*, 14(4), pp. 369–395. doi: 10.1191/0962280205sm403oa.

Education Endowment Foundation (2018) *Statistical analysis guidance for EEF evaluations*. Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf (Accessed: 20 November 2019).

Elliot, C. D. and Smith, P. (2011) *British Ability Scales: Third Edition (BAS3)*. London: GL Assessment.

Ge M, Durham LK, Meyer RD, Xie W, Thomas N. Covariate-Adjusted Difference in Proportions from Clinical Trials Using Logistic Regression and Weighted Risk Differences. Drug Information Journal. 2011;45(4):481-493. doi:10.1177/009286151104500409

Howard, S. J. and Melhuish, E. (2017) 'An Early Years Toolbox for assessing early executive function, language, self-regulation, and social development: validity, reliability, and preliminary norms', *Journal of Psychoeducational Assessment*, 35(3), pp. 255–275. doi: 10.1177/0734282916633009.

Nunes, T. et al. (2018) 1stClass@Number: Evaluation report and executive summary. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/1stClass@Number_evaluation_report.pdf (Accessed: 24 March 2022).

Robinson-Smith, L. et al. (2018) Maths Champions: Evaluation report and executive summary. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Maths_champions_evaluation_report.pdf (Accessed: 24 March 2022).

Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987.

Saghaei, M. and Saghaei, S. (2011) 'Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials', *Journal of Biomedical Science and Engineering*, 4(11), pp. 734–739. doi: 10.4236/jbise.2011.411090.

Stokes, L. et al. (2018) Mathematical Reasoning: Evaluation report and executive summary. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Mathematical_Reasoning.pdf (Accessed: 24 March 2022).

What Works Clearinghouse (n.d). Procedures Handbook, Version 4.0, p.13-14: https://ies.ed.gov/ncee/wwc/docs/referenceresources/wwc_procedures_handbook_v4.pdf