Using Peer-to-Peer Coaching to improve Maths attainment, a two-armed cluster randomised trial
Statistical Analysis Plan

Evaluator (institution): RAND Europe
Principal investigator(s): Elena Rosa Speciani

PROJECT TITLE	Using Peer-to-Peer Coaching to improve Maths attainment, a two-armed cluster randomised trial
DEVELOPER (INSTITUTION)	CoachBright
EVALUATOR (INSTITUTION)	RAND Europe & University of Leeds (UoL)
PRINCIPAL INVESTIGATOR(S)	Elena Rosa Speciani
PROTOCOL AUTHOR(S)	Merrilyn Groom, Helen Murphy
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	11 – 12 (Year 7, KS3), 14 – 15 (Year 10, KS4)
NUMBER OF SCHOOLS	93 schools
NUMBER OF PUPILS	30 per school – 15 from Year 7 and 15 from Year 10 (~2,790 pupils)
PRIMARY OUTCOME MEASURE AND SOURCE	Maths attainment (Hachette Learning Access Mathematics Tests [AMT] – 1B for Year 7; 3B for Year 10)
SECONDARY OUTCOME MEASURE AND SOURCE	1. Self-efficacy (Sources of Mathematics Self-Efficacy Scale [SMSES]) 2. Metacognition (Junior Metacognitive Awareness Inventory [JMAI])

SAP version history

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

Table of contents

Introduction

CoachBright's Peer-to-Peer coaching is a programme targeted at disadvantaged pupils that aims to improve attainment outcomes through metacognitive guidance, tailored problems, and mutual concept explanation. This trial focuses specifically on mathematics attainment, with high-attaining Year 10 coaches provided training and support to coach disadvantaged Year 7 coachees who are lower attainers in maths, during weekly 60-minute sessions over the course of 10 weeks.

The core content of the peer coaching sessions centres around the younger pupils and their coaches developing and deploying metacognitive strategies in relation to mathematics. CoachBright offers two types of coaching: pastoral and academic. This evaluation will focus on academic coaching in mathematics. During the coaching sessions, peer coaches support their paired younger pupil to reflect on their academic progress and support them through specific topics they are finding hard. This may be done through tailored mathematics problems, mutual concept explanation, and peer coaching for deeper understanding: coachees are encouraged to explain their reasoning to the peer coach, to promote a deeper understanding of mathematical concepts. Coaches are recommended to reflect on the session so that improvements can be made for the following week. This might include CoachBright or the school providing a wider range of resources to support session planning.

The content of the programme is as follows:

- The selected School Coordinator will first undergo an onboarding session with the CoachBright Programme Manager, who will provide the overview of the programme and discuss pupil data required for the study and provide guidance around selecting and matching pupils.
- The 15 coaches will be selected by the school, either through application or invitation.
- The CoachBright Programme Manager will provide training to the 15 coaches, where the coaches learn about CoachBright and the coaching journey. They also learn about session planning and how to gain and use coachee reflections.
- The launch event will run in each school, during which the coaches and coachees get to know each other and discuss the areas of focus for their coaching sessions, both allowing the pairings to build relationships and giving the School Coordinator and CoachBright Programme Manager the opportunity to rearrange pairs.
- What follows is the 10 weekly coaching sessions, which will take place during the school day during pastoral or PHSE periods, a graduation event (often held at a local university), as well as support throughout for the coaches to earn a Schools, Students and Teachers (SSAT) Network coaching qualification at either bronze or silver level, a UCAS recognised accreditation.

This evaluation will seek to answer the following primary research question:

RQ1 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on mathematics attainment of Year 7 coachees, measured by the Access Mathematics Tests (AMT) compared to similar pupils in control settings receiving business-as-usual?

In addition, it will also respond to the following secondary research questions:

RQ2 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on maths attainment of Year 10 coaches, measured by the Access to Mathematics Tests (AMT), compared to similar pupils in control settings receiving business-as-usual?

RQ3 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on maths self-efficacy of Year 7 coachees, as measured by the Sources of Mathematics Self Efficacy Scale (SMSES), compared to similar pupils in control settings receiving business-as-usual?

RQ4 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on metacognition of Year 7 coachees, measured by the Junior Metacognitive Awareness Inventory (JMAI), compared to similar pupils in control settings receiving business-as-usual?

RQ5 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on maths self-efficacy of Year 10 coaches, as measured by the Sources of Mathematics Self Efficacy Scale (SMSES), compared to similar pupils in control settings receiving business-as-usual?

RQ6 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on metacognition of Year 10 coaches, measured by the Junior Metacognitive Awareness Inventory (JMAI), compared to similar pupils in control settings receiving business-as-usual?

Design overview

Table 1. Design overview

Trial design, including number of arms		Two-arm, cluster randomised controlled trial
Unit of randomisation		School
Stratification variables (if applicable)		Region: The region or LA in which the school is located.
Primary outcome	variable	Mathematics attainment
	measure (instrument, scale, source)	Access Mathematics Tests (AMT), raw score 0-60 / 0-45, Hachette Learning (Hachette Learning Access Mathematics Test (AMT) – Level 1B for Year 7; Level 3B for Year 10)
Secondary outcome(s)	variable(s)	1. Maths self-efficacy 2. Metacognition
	measure(s) (instrument, scale, source)	1. Sources of Mathematica Self-Efficacy Scale (SMSES), 24 – 120, (Usher & Pajares, 2009) 2. Junior Metacognitive Awareness Inventory (JMAI), 1 – 18, (Sperling et al., 2002).
	variable	KS2 SATS maths attainment

Baseline for primary outcome	measure (instrument, scale, source)	KS2 maths scores (KS2_MATSCORE), 0 -999, acquired from the NPD
Baseline for secondary outcome	**variable**	KS2 SATS maths attainment
	measure (instrument, scale, source)	KS2 maths scores (KS2_MATSCORE), 0 -999, acquired from the NPD

This project is designed as a two-armed, cluster-randomised trial. A static stratified randomisation took place at the school level, with region or LA in which the school is located used as the stratification variable during randomisation. This means that an equal number of schools from each region were included in the treatment and control conditions. This took place after baseline data collection in October 2025, with schools being randomly assigned to either receive the Peer-to-Peer coaching programme or to carry on with teaching as normal.

The control condition represents business as usual, with schools randomised into this group receiving £750 in compensation for their support with the trial and associated data collection activities. The treatment condition involves schools receiving access to the CoachBright Peer-to-Peer programme, for which they will pay a heavily discounted (90%) cost of £500 for the service. Before randomisation, schools identified a lead teacher[1] (School Coordinator) who was responsible for identifying 15 coaching pairs (15 Year 7 and 15 Year 10 pupils) that were eligible based on the selection criteria developed by the delivery team.

Schools were recruited during the 2024-25 academic year using the following criteria:

- They were a local authority, multi academy trust, a free school, or a grammar school.
- They have pupils in Year 7 and Year 10 on the same site during the 2025-26 academic year.
- They were from one of eight geographical regions, including London, North East, South East, South West, North West, West Midlands, East Midlands, East of England.
- They were able to identify 15 coaching pairs (15 Year 7 and 15 Year 10 pupils that fit the pupil eligibility criteria.

Year 7 pupils in recruited schools were considered eligible to be a coachee if:

- They have low attainment in maths, measured by their KS2 SATs score being less than 100.
- They are eligible for Free School Meals (at least 80% of Year 7 pupils must meet this criterion) or should meet at least one of the criteria for CoachBright's wider definition of disadvantage.

[1] The only requirement for the lead teacher is that they are a member of staff. Schools had full autonomy over who should be nominated to the lead role of School Coordinator. Within the trial, there have been a mixture of subject specialists, SLT members, heads of departments, and pastoral staff taking on the School Coordinator role.

Year 10 pupils in recruited schools were considered eligible to be a coach if:

- They have high attainment in maths, measured by the potential to achieve GCSE grades of 7-9 in maths (this will be determined by a combination of end-of-year exams, teacher assessments, any standardised tests taken in Year 9).
- At least 50% of the Year 10 pupils recruited are eligible for FSM or meet one of the wider disadvantage criteria.

Year 10 coaches could be selected through application or invitation, with their maths attainment being the more important selection criteria.

CoachBright's wider definition of disadvantaged pupils extends to those:

- Eligible for pupil premium funding;
- Eligible for the Service Premiums;
- Is a young carer;
- Is or has been a Looked After Child;
- Is known to be a refugee or asylum-seeking child;
- Neither parent has attended higher education;
- Otherwise considered disadvantages by the school.

A split-cohort design was adopted due to constraints in programme delivery capacity: that is, it would have been challenging for the delivery team to deliver the programme to the number of schools necessary for a well-powered trial, if the programme was administered with a single cohort. The trial will thus be conducted using two cohorts. The intervention ran between November 2025 and February 2026 for Cohort 1, with endline testing data collected between January and February of 2026. The intervention is running between February 2026 and June 2026 for Cohort 2, with endline testing data collected between May and June of 2026.

Randomisation

In this cluster randomised control trial, 93 schools were allocated in as close to a 50:50 split as possible, to the intervention and control conditions - 47 were assigned to control and 46 assigned to treatment. Originally, it was planned to have a sample of 100 schools, however, a number of schools dropped out of the trial pre-randomisation due to capacity issues and were unable to be replaced within the timeframe. While schools are the unit of randomisation, children are the unit of analysis, reflecting the clustered-RCT design of this evaluation.

Randomisation was stratified by region, ensuring that an equal number of schools from each region were included in the treatment and control conditions. With the school-level approach of the intervention, this ensures that pupils within the programme are experiencing as close to the same condition as possible.

Cohort allocation was decided using two different methods. Once randomisation of treatment had been completed at the school-level, CoachBright was informed of the treatment allocation

of settings. CoachBright assigned the allocation of cohorts for the treatment schools, with a 50:50 split – 23 schools assigned to each Cohort 1 and Cohort 2. Their decision was made based on resource allocation and the feasibility of coordinators to travel between settings to deliver the programme. RAND Europe was responsible for the allocation of control schools, conducting a randomisation of the control schools to determine if they were assigned to Cohort 1 or Cohort 2. This was again stratified by region, with the size of the cohorts being as even as possible achieving a 24:23 split; 24 assigned to Cohort 1 and 23 assigned to Cohort 2.

Table 2: Randomisation results

	Treatment (of which in Cohort 1)	Control (of which in Cohort 1)	Total (of which in Cohort 1)
London and East of England	14 (8)	15 (8)	29 (16)
North West	5 (3)	6 (3)	11 (6)
North East	6 (3)	5 (2)	11 (5)
South East	7 (3)	6 (3)	13 (6)
South West	5 (3)	5 (3)	10 (6)
East and West Midlands	9 (3)	10 (5)	19 (8)
Total	46 (23)	47 (24)	93 (47)

Randomisation was conducted by a member of the evaluation team who was blind to the setting identities. A tailored package in Stata (*randtreat*) was used to implement the settings randomisation with regional stratification. A second researcher at RAND Europe then checked the randomisation code and the outcomes to verify independence. The code used to randomise settings will be included in the final Evaluation Report at the conclusion of the study. A master copy of the final allocation was retained in a locked folder on RAND Europe's servers to prevent editing, with the final allocation communicated to the delivery team and checked against the master copy to ensure no edits occurred in the processing or transfer of data.

Further randomisation was required for the secondary outcome measures as classes were randomly assigned to only one of the secondary outcome tests to reduce burden. For these tests region was once again used as a stratification variable, with the addition of treatment status and cohort as further stratification variables. In order to preserve power, these stratification variables were prioritised in the following order: treatment status, cohort, region. The package *randtreat*

was used again but run separately on the treatment and control groups to ensure an even division. All other randomisation procedures were the same.

Table 3: Secondary randomisation results - Treatment status

	Self-Efficacy	Metacognition	Total
Treatment	23	23	46
Control	23	24	47
Total	46	47	93

Table 4: Secondary randomisation results - Cohort

	Self-Efficacy	Metacognition	Total
Cohort 1	24	23	47
Cohort 2	22	24	46
Total	46	47	93

Table 5: Secondary randomisation results - Region

	Self-Efficacy	Metacognition	Total
London and East of England	15	14	29
North West	6	5	11
North East	4	7	11
South East	6	7	13
South West	5	5	10
East and West Midlands	10	9	19
Total	46	47	93

Outcomes

Primary outcome

The primary outcome of this study will be mathematics attainment, measured through Hachette Learning's Access Mathematics Test (AMT). AMT works with a wide range of curricula, using a skills-based assessment to assess a pupil's comprehension of core mathematical skills. This is a valid and reliable measure, with standardised tests having been subject to a rigorous four stage development process to ensure that the content and outcomes are valid. This included a large-scale standardisation trial run between 2023 and 2024, with 16,867 papers completed by over 7000 students (George et al., 2024).

Following the guidance provided by Hachette Learning, Year 7 pupils will take Test 1B. Hachette Learning suggest that this is suitable for learners aged 10 and upwards, with much of the content covered at an equivalent level of challenge to tests undertaken at the end of Key Stage 2 in the National Curriculum. This includes skills such as conversion between different units of measure, describing positions of shapes on a 2D grid, and identifying numbers that satisfy an equation with two unknowns. Whilst this test is equivalent to those taken at the end of KS2, it has been validated and standardised on pupils up to Spring Term of Year 8, so for Year 7 pupils with maths SATs scores of less than 100, we are unlikely to face significant ceiling or floor effects.

Following the guidance provided by Hachette Learning, Year 10 pupils will take test 3B, which covers upper KS3 content and can be used with KS4 learners. Skills tested in this version of the AMT include solving problems with direct and inverse proportion, calculating and solving problems involving perimeters of 2D shapes (including circles), and the ability to rearrange formulae to change the subject. Despite covering KS3 content, Hatchette Learning suggests that this test is suitable for Year 10 pupils, having been validated and standardised on pupils up to Spring Term of Year 11 (age 16). Despite these reassurances, it is possible that, given the Year 10 students are all predicted to achieve GCSE grades of 7-9 in maths, we may still encounter some ceiling effects in this age group.

The digital form of AMT will be used, taking 45 minutes to complete, with 10 minutes of extra time allowed for students with approved SEND accommodations, following usual criteria for extra time in exams. The results of this will be automatically generated upon completion, minimising score variability. These results include: a raw score, a standardised score, an age standardised score, a mathematics age, a percentile, and diagnostic strand information. Following EEF guidance, we will use the raw score as the primary outcome in all analysis.

Secondary outcomes

The first of the secondary outcomes to be measured is maths self-efficacy due to the research evidence on self-efficacy's role in mediating academic attainment (Wang et al., 2024), its influence on long-term achievement in maths (Parker et al., 2013), and the presence of maths self-efficacy as a short-term outcome on Peer-to-Peer's Theory of Change.

Maths self-efficacy will be measured using the Sources of Maths Self-Efficacy Scale (SMSES) (Usher & Pajares, 2009), which is specifically tailored to self-efficacy behaviour in maths. The SMSES consists of 24 questions ranked on a Likert-type scale, that are designed to capture

students' perceptions of their feelings and experiences related to maths learning. While this test was originally designed for middle school students in America, it has been successfully used to assess maths self-efficacy in a range of countries and cultures including Oman, Turkey and Argentina (Navarro et al., 2025). Within all of these examples, the scale had to be adapted either through translation, rewording, or a complete restructure of order to effectively assess self-efficacy in their respective cultural contexts. As this scale had not yet been used in the English context, it was not possible to determine if the scale would need to similarly be adapted for the English cultural context. Furthermore, few of studies cover the entire age range to be examined in this trial. Navarro et al., (2025) focused on the same age range, with Grades 6 (age 11-12) to 9 (age 14-15) but most other studies have included a much smaller range of ages.

With this novel population set of English school students that includes 11-12 and 14-15 year olds, we have conducted a pilot study in three schools to assess the test's suitability, focusing on the internal validity of SMSES, measured through Cronbach's Alpha and confirmatory factor analysis, to establish if the structure and question selection was representative of self-efficacy in the sample. The SMSES was determined to be suitable and will be used as the secondary measures of maths self-efficacy in this trial. Further details can be found in Appendix A.

Metacognition is the other secondary outcome to be explored. Despite strong evidence overall to suggest that metacognitive knowledge directly impacts maths outcomes (Donker et al., 2014), few trials in the UK have explicitly established this link. EEF's justification for funding this project was funding priority 1a: Mathematical reasoning: approaches designed to develop learners' cognitive, metacognitive and self-regulative knowledge and skills, specifically in maths.

Metacognition will be measured using the Junior Metacognitive Awareness Inventory (JMAI) (Sperling et al., 2002), which has been rated as suitable for use in both KS3 and KS4 by the EEF. It consists of 18 Likert scale questions which measure components of pupils' metacognition, particularly knowledge of cognition and regulation of cognition.

Both secondary outcome measures will be computer-based tests, completed on Qualtrics[2], and will be administered immediately following the primary outcome test. The University of Leeds will add the SMSES and the JMAI to the survey platform and administer them, using independent qualified and trained test administrators. Schools participating in the evaluation will be randomly assigned either the SMSES or the JMAI, both of which take a few minutes to complete, but in order to reduce test burden pupils will not be asked to complete both. Schools were randomly assigned to complete one of the two tests, resulting in an approximately half of pupils in Year 7 completing the SMSES and half of pupils in Year 7 completing the JMAI – and the same for Year 10. This reduces the power, increasing the MDES for secondary outcomes to between 0.24 and 0.29 (see Appendix B for power calculations on secondary outcomes). Randomisation of the secondary measures took place at the school level to reduce burden on test administrators and schools, after schools had been randomly assigned to treatment or control (see Randomisation subsection for further details).

[2] https://www.qualtrics.com/en-gb

Sample size calculations overview

Table 6. Year 7 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)		0.151	0.161	0.165	0.176	0.153	0.168	0.171	0.184
Pre-test/ post-test correlations	level 1 (pupil)	0.75		0.71		0.75		0.71	
	level 2 (class)	n/a		n/a		n/a		n/a	
	level 3 (school)	0.61		0.54		0.61		0.54	
Intracluster correlations (ICCs)	level 2 (class)	n/a		n/a		n/a		n/a	
	level 3 (school)	0.07		0.07		0.07		0.07	
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		15		13		15		13	
Number of schools	intervention	50	44	50	44	46	40	46	40
	control	50	44	50	44	47	41	47	41
	total	100	88	100	88	93	81	93	81
Number of pupils	intervention	15	15	13	13	15	15	13	13
	control	15	15	13	13	15	15	13	13
	total	30	30	26	26	30	30	26	26

The minimum detectable effect size (MDES) for this study has been calculated using a two-level random assignment design, reflecting the design of the trial, with randomisation occurring at the school level and analysis occurring at the individual level. MDES have been calculated for all Year 7 and Year 10 pupils, and for the respective FSM subsets. In calculating the MDES for all groups, the following assumptions have been made: randomisation at the school level with 50:50 allocation, alpha of 0.5 and power at 0.8.

Based on previous EEF trials of Year 7 maths interventions, an intra-cluster correlation of 0.07 and a pre-test/post-test correlation of 0.75 at the pupil level, 0.61 at the school level, with 15 pupils per cluster, has been used for the Year 7 pupils. Under these assumptions, the MDES reported in the protocol was 0.151. As is standard in EEF trials, subgroup analysis will be run on children from disadvantaged backgrounds, using FSM status, identified with the EVER_FSM_P variable from the National Pupil Database (NPD). Based on review of previous EEF secondary school maths trials the pre-test/post-test correlations for the FSM subgroup are 0.71 at the pupil level and 0.54 at the school level. Additionally, the average cluster size was reduced to 13, as the

trial is assuming that at least 80% of Year 7 participants will come from disadvantaged backgrounds. The MDES reported in the protocol was 0.165 for Year 7 pupils with FSM status.

Following randomisation, the overall MDES for the Year 7 participants increases to 0.153, due to under recruitment in the number of participating schools, with only 93 recruited schools at randomisation, compared to an assumption of 100 schools in the protocol. All other assumptions have been held constant. Similarly, the MDES for Year 7 pupils with FSM status has increased to 0.171 with no attrition.

Table 7. Year 10 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)		0.179	0.191	0.184	0.196	0.186	0.198	0.191	0.204
Pre-test/ post-test correlations	level 1 (pupil)	0.73		0.67		0.73		0.67	
	level 2 (class)	n/a		n/a		n/a		n/a	
	level 3 (school)	0.53		0.45		0.53		0.45	
Intracluster correlations (ICCs)	level 2 (class)	n/a		n/a		n/a		n/a	
	level 3 (school)	0.1		0.05		0.1		0.05	
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		15		8		15		8	
Number of schools	intervention	50	44	50	44	46	40	46	40
	control	50	44	50	44	47	41	47	41
	total	100	88	100	88	93	81	93	81
Number of pupils	intervention	15	15	8	8	15	15	8	8
	control	15	15	8	8	15	15	8	8
	total	30	30	16	16	30	30	16	16

The MDES for the Year 10 sample, both overall and for FSM pupils, have also been calculated. We assumed an intra-cluster correlation of 0.10, and 15 pupils per cluster, but a slightly lower pre-test/post-test correlation of 0.73 at the pupil level and 0.53 at the school level, based on a recent paper by the EEF (Singh et al., 2023). Under these assumptions, the MDES reported in the protocol was 0.179. For the FSM sub-group analysis, these assumptions are adjusted to an intra-cluster correlation of 0.05, a pre-test/post-test correlation of 0.67 at the pupil level and 0.45 at the school level, again based on the Singh et al., (2023) paper. The average cluster size has been adjusted to 8 pupils to reflect the fact that 50% of the Year 10 pupil sample will be FSM eligible.

These assumptions gave an MDES in the protocol of 0.184 for Year 10 FSM pupils. Following randomisation, the overall MDES for the Year 10 sample increased to 0.186 and to 0.191 for those with FSM status due to the slight under recruitment of schools.

The MDES for both the Year 7 and Year 10 sample, overall and FSM, for the secondary outcomes can be found in Appendix B. The same pre-test/post-test correlations and intracluster correlations have been used for the different age cohorts as explored in tables 6 and 7, however the sample sizes are reduced due to the decision to randomly assigned schools either the SMSES or the JMAI. This smaller sample size has increased the MDES, pushing it over 0.2.

Table 8. Expected attrition

	School level	Pupil level
Expected attrition (%)	12%	26%

Table 9. MDES after attrition

	School level	Pupil level	School and Pupil level
Year 7	0.168	0.178	0.190
Year 10	0.198	0.216	0.230

Finally, the impact of pupil and school level attrition on the MDES has been explored. The same assumptions listed above have been used but with the addition of pupil-level of attrition of 26% and school-level attrition of 12%. These figures are based on attrition reported in published EEF secondary maths trials (The Rise Project, Maths in Context, Fit to Study, Realistic Maths Education, and Mathematics Mastery), taking the mean value across these studies for pupil and school level attrition. The effects of these can be seen in the sample size overview tables. The attrition used in tables 6 and 7 are the school level rate of 12%. This increases the MDES of 0.168 and 0.184 for the overall and FSM Year 7 pupils, respectively, and 0.198 and 0.204 for the Year 10 pupils. Table 9 further explores the impact of attrition, including the impact of pupil level attrition and the combined attrition at pupil and school level. These MDES values are higher, with the Year 10 value exceeding 0.2, which is the conventional limit for a well-powered study.

Analysis

The statistical analysis proposed follows the most recent revised EEF Statistical Analysis Guidance available (EFF, 2022).

Primary outcome analysis

As detailed in the protocol (Speciani et al., 2025), this efficacy trial has one primary research question:

RQ1 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on mathematics attainment of Year 7 coachees, measured by the Access Mathematics Tests (AMT), compared to similar pupils in control settings receiving business-as-usual?

Our primary analysis will investigate any difference in maths attainment, as measure by AMT scores, between treatment and control schools and will be conducted on an intention-to-treat (ITT) basis, as outlined in the EEF's analysis guidance (EEF, 2022). The raw scores will be obtained for the outcome measure, as well as a Standardised Age Score (SAS), generated on sample which is representative of UK schools. Under an ITT approach, analysis will include all randomised settings and baselined children, grouped according to random assignment, regardless of programme compliance or treatment dosage. It is an inherently conservative approach, estimating the average effect of offering the intervention, and is key to ensuring an unbiased analysis of intervention effects in line with EEF's guidance (EEF, 2022).

A linear mixed effects regression model will be used to estimate the adjusted mean difference in scores, with analysis undertaken at the pupil-level. Year 7 and Year 10 pupils will be analysed separately to understand the differential effects on each group in isolation, facilitating the identification of specific impacts for each year group, with the Year 7 outcome representing the main finding of this study. This separate analysis also enhances statistical precision and power by reducing variability (Bryk & Raudenbush, 1992).

The impact will be estimated in the model outlined below in equation (1). Equation (1) is known as a random-intercept model because the school-specific intercepts for each school j ($\beta_{0j}=\beta_0+u_j$) vary randomly with the school-level residual ($\beta_{0j} \sim_{i.i.d} N(\beta_0, \sigma_u^2)$). The model will additionally control for pre-test (baseline) attainment, as measured by baseline KS2 maths scores, and estimate fixed effects for the cohort and stratification variable (region) at the school level.

$$(1)\ Y_{ij} = \beta_0 + \tau P2P_j + \beta_1 Z_j + \beta_2 X_{ij} + \beta_3 C_j + u_j + e_{ij}$$

Where:

Y_{ij} = AMT raw score for child *i* in school *j*, at endline;

β_0 = cluster-level coefficient for the slope of a predictor on number skills;

$P2P_j$ = binary variable which indicates whether the school *j* was assigned to receive the intervention [1] or was assigned to the control group [0];

Z_j = school-level characteristic, here the stratifying variable of geographical location (as used for randomisation);

X_{ij} = child-level characteristic, specifically the KS2 maths score (KS2_MATSCORE recorded in the NPD) which is used here as a baseline outcome;

C_j = Factor variable indicating whether the school *j* was part of cohort 1 or cohort 2;

u_j = setting-level residuals and

e_{ij} = child-level residuals.

The coefficient τ is the outcome of interest, as an estimate of the conditional effect of treatment on endline AMT score. The standardised effect size will be calculated using Hedges' g for τ (more in Calculation of Effect Sizes).

One important consideration is how to treat the data from the two different cohorts of pupils. Two approaches could be used. In our primary analysis model, we adopt a combined-cohort model, which effectively combines data from both cohorts into a single model. This is done by using a binary cohort variable, which indicates which cohort each school belongs to, allowing us to control for cohort-specific effects when estimating the intervention effect size. The strength of this approach is that, in making use of the full analytical sample, we increase the statistical power, reducing the risk of type II errors in analyses. However, this approach relies on the assumption that there are no significant differences in the observable and unobservable characteristics of the students or in the delivery and administration of Peer-to-Peer coaching, across the two cohorts. Observable or unobservable differences across cohorts would mean that the estimated effect of the intervention using equation 1 would be confounded by cohort. Given randomisation was conducted in one batch, any observable differences in the cohorts can only be due to random chance; however, this does not preclude differences in delivery of the intervention between cohorts. The alternative approach is to treat each cohort as a separate trial and combine the two cohorts using metanalytic techniques. We propose using this alternative approach as a sensitivity analysis on the primary estimates of treatment effects (outlined below).

Whilst we are combining cohorts, we are not proposing to combine year-groups in the primary analysis. This analysis will be conducted separately for Year 7 and Year 10 pupils. Distinct analyses should allow for clearer interpretation of results, making it easier to identify specific effects for each year group, especially given the year groups span multiple key stages and have markedly different selection criteria, particularly around previous or current mathematical achievement. However, we are aware that the two cohorts are not independent of each other, so we propose conducting a three-level (pupil, year-group, and school) hierarchical model as part of our exploratory analysis, which is outlined in further detail below.

Given we have multiple analyses, one for each year group, we will employ a Romano-Wolf correction to estimated effect sizes, to statistically correct for over-rejection of null hypotheses under multiple hypothesis testing. Given the year groups are not independent samples within the context of this trial, with the peer-mentoring dyad naturally introducing dependence between the two year groups for treatment schools, a Bonferroni correction would likely be too conservative. For this reason, we opt for the Romano-Wolf to allow for the estimation of family-wise error rate (Clarke et al., 2019), as recommended in the EEF's evaluation guidance (Education Endowment Foundation, 2022).

All analyses will be done in R or Stata, version 17 and above, using the *eefanalytics* package (Vallis et al., 2021). Multiple hypothesis testing corrections will be made using the *rwolf2* package in Stata (Clarke, 2021) or *crctStepdown* in R (Watson, 2024).

Secondary outcome analysis

As outlined in the protocol (Speciani et al., 2025), this study will answer the following secondary research questions:

RQ2 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on maths attainment of Year 10 coaches, measured by the Access to Mathematics (AMT), compared to similar pupils in control settings receiving business-as-usual?

RQ3 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on maths self-efficacy of Year 7 coachees, as measured by the Sources of Mathematics Self Efficacy Scale (SMSES), compared to similar pupils in control settings receiving business-as-usual?

RQ4 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on metacognition of Year 7 coachees, measured by the Junior Metacognitive Awareness Inventory, compared to similar pupils in control settings receiving business-as-usual?

RQ5 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on maths self-efficacy of Year 10 coaches, as measured by the Sources of Mathematics Self Efficacy Scale (SMSES), compared to similar pupils in control settings receiving business-as-usual?

RQ6 - What is the impact of the CoachBright Peer-to-Peer (P2P) coaching intervention on metacognition of Year 10 coaches, measured by the Junior Metacognitive Awareness Inventory, compared to similar pupils in control settings receiving business-as-usual?

Our secondary analysis will follow the considerations made in the primary analysis and will once again follow the ITT principle. Analysis for RQ2 will follow the same methodology and structure as in the primary analysis but use the sample of Year 10 pupils rather than the Year 7 pupils. Further secondary analysis will utilise the Sources of Mathematics Self-Efficacy Scale (SMSES) as a measure of maths self-efficacy and the Junior Metacognitive Awareness Inventory (JMAI) as a measure of metacognition, rather than AMT.

For the secondary analysis, we will use the same multi-level modelling approach as in the primary analysis. That is, more specifically, we will estimate a two-level random-intercept model (see Primary outcome analysis section for justification), where the dependent outcome variable differs in each regression. The first regression will focus on the impact of P2P coaching on maths attainment for Year 10 pupils, using equation (1). The second will look at the impact of P2P coaching on maths self-efficacy, the third on metacognition. In each regression, the independent variables will match those in the primary analysis: a treatment indicator, geographical location and type, and respective KS2 maths scores at baseline.

$$(2)\ Y_{ij} = \beta_0 + \tau P2P_j + \beta_1 Z_j + \beta_2 X_{ij} + \beta_3 C_j + u_j + e_{ij}$$

Where: Y_{ij} = Either SMES or Junior Metacognitive Awareness Inventory score for child *i* in school *j*, at endline;

β_0 = cluster-level coefficient for the slope of a predictor on number skills;

$P2P_j$ = binary variable which indicates whether the school *j* was assigned to receive the intervention [1] or was assigned to the control group [0];

Z_j = school-level characteristic, here the stratifying variable of geographical location (as used for randomisation);

X_{ij} = child-level characteristic, specifically the KS2 maths score (KS2_MATSCORE recorded in the NPD) which is used here as a baseline;

C_j = Factor variable indicating whether the school *j* was part of cohort 1 or cohort 2;

u_j = setting-level residuals and

e_{ij} = child-level residuals.

The analysis for regressions two and three will be conducted separately on Year 7 and Year 10 pupils, following the justification outlined for the primary analysis. As with the primary outcome model, the coefficient is the outcome of interest. We will use τ to calculate a standardised Hedge's g effect size (more in Calculation of Effect Sizes). Given we have multiple secondary hypotheses, we will employ a Romano-Wolf correction to the estimated effect sizes to statistically correct for over-rejection of null hypotheses.

All analyses will be done in R or Stata, 17 or above, making use of the *eefanalytics* package where possible. Multiple hypothesis testing corrections will be made using the rwolf2 package in Stata (Clarke, 2021) or crctStepdown in R (Watson, 2024).

Subgroup analyses

As specified in the protocol, the trial will explore the impact of Peer-to-Peer coaching on a number of subgroups. The analyses will follow the EEF analysis guidance (EEF, 2022), where the impact of the intervention on subgroups will be firstly examined by repeating the model used for primary analysis on the subgroup, and secondly through an interaction-term model on the full analytical. The subgroups to be considered are:

- **FSM eligibility:** as discussed in the Sample Size calculations section, our trial is powered to identify an effect on the sub-group of FSM pupils. To identify whether a pupil is FSM eligible, we will use the variable EVERFSM_6_P from the National Pupil Database (NPD).
- **SEND:** The second sub-group analysis undertaken will focus on pupils who have special educational needs and disabilities (SEND). To understand the impact of Peer -to-Peer coaching on SEND pupils, we will access the SENprovisionMajor variable in the NPD. This is a categorical variable with four categories that we will condense into two, denoting whether a child is considered SEND or not SEND.
- **EAL:** The third sub-group analysis undertaken in this project will focus on pupils for whom English is an additional language (EAL). This sub-group analysis provides further understanding on whether Peer-to-Peer is appropriate for pupils who need a high level of additional language scaffolding from their teachers. The variable LanguageGroupMajor from the NPD will be used to investigate this.

The first analysis will run the primary model given in equation 1 on the FSM, SEND, and EAL subgroups only. Following the approach for primary analysis, this model will be run separately on

each year group. Effect sizes and statistical uncertainty will be calculated on all subgroups following the procedure outlined in the above section on Primary Analysis. The second analysis will estimate the treatment effect on the subgroup using an interaction model, which makes use of the full analytical sample for each year group, here using FSM as an example:

$$(3)\ Y_{ij} = \beta_0 + \boldsymbol{\tau} P2P_j + \beta_1 FSM_{ij} + \beta_2 (FSM_{ij} * P2P_j) + \beta_3 Z_j + \beta_4 X_{ij} + \beta_5 C_j + \boldsymbol{u_j} + \boldsymbol{e_{ij}}$$

This is the same model specification as equation 1 and 2, with the addition of the FSM_j indicator of disadvantage and an interaction term combining FSM eligibility and treatment allocation ($FSM_j * P2P_j$). The primary coefficient of interest in the interaction model is β_2, which can be interpreted as the additional treatment effect experienced by pupils from disadvantage background: a positive β_2 is indicative of a treatment acting as a 'gap-closer' and a negative β_2 indicative of treatment acting as a 'gap-widener'. For SEND and EAL subgroups, equation 3 is repeated using SEND and EAL binary indicators instead of the FSM_i indicator.

In accordance with EEF guidelines, we will compare the two estimates of effect size of the treatment on the sub-sample: i) that generated by running the primary model given in equation 1 on the sub-sample, and ii) that calculated from the interaction model outlined in equation 3. The treatment effect from the interaction models is calculated according to the following formula:

$$\frac{\tau P2P_j + \beta_2 (P2P_j * FSM_{ij})}{sd}$$

The coefficients in the numerator come directly from equation 3, and the standard deviation used in the denominator is the unconditional standard deviation of the FSM sub-sample (both treatment and control).

Sensitivity analysis and additional analyses

As discussed above in the Primary Analysis section, we will explore alternative approaches to combining the two cohorts as a robustness check on the primary impact estimate.

The approach taken in the primary analysis includes a cohort factor variable in a 'combined cohort' two-level hierarchical model in order to explicitly account for the cohort each participant belongs to. An alternative approach is to treat each cohort as a separate trial. Under this approach, any assumptions around cohort homogeneity, either in terms of characteristics of participating schools and students or fidelity of intervention delivery, do not need to hold, and so we mitigate the risk of confounding bias from unobservable differences in the two cohorts (even if these are only present by random chance). However, separate analyses of each cohort would lower the statistical power due to the lower analytical sample size in each model. This would increase the risk of type II error.

To ensure we use the full analytical sample size of the trial, we will perform a meta-analysis to determine an aggregate weighted effect size of Peer-to-Peer coaching across both cohorts. The analysis will follow these steps:

i) Analyse each cohort separately, according to the model specified in Equation 1. This produces two separate effect sizes, one for each cohort, according to the primary outcome model.

ii) Measure heterogeneity between cohorts, either by using Cochran's Q or I^2, or testing for statistical differences in cohort-specific treatment effects from the separate cohort models. Whilst Cochran's Q or I^2, is the standard approach in metanalysis, there is a high degree of uncertainty around these statistical tests with just two separate samples (Deeks et al., 2024). For this reason, we propose additionally analysing the presence of heterogeneity through conducting statistical testing of differences between the estimated cohort-specific effect sizes from in step 1.

iii) If there is evidence of heterogeneity between cohorts, we will run a random effects metanalysis. Given the small number of independent cohorts in this analysis, we will conduct a random effects metanalysis using the Hartung-Knapp-Sidik-Jonkman approach. However, even this approach can result in confidence intervals that are too narrow when there are very small numbers of independent studies or cohorts (Rover et al, 2015). For this reason, we will conduct recommended ad hoc adjustments through the meta package in R (Schwarzer, 2025).

If there is no evidence of heterogeneity, we could consider running a fixed effects metanalysis model; however, it is unlikely that it will offer statistical benefits over the combined-cohort model conducted under primary analysis. For this reason, metanalysis will only be conducted if there is evidence of heterogeneity in treatment effect between cohorts.

Exploratory analyses

Analysis of effect of mixed-gender and mixed-ability pairings

Research indicates that the gender of peers in secondary schools significantly influences outcomes (Smith & Andersen, 2022) and progression into STEM subjects (Riegle-Crumb & Morton, 2017), particularly for girls when their female peers also have an interest in STEM (Raabe et al., 2019). To contribute to this important body of literature, further analysis will be conducted examining the differential impact of same-gender and mixed-gender pairs in Peer-to-Peer learning.

We will conduct exploratory analyses on the differences between same-gender and mixed-gender pairs of coaches and coachees, to help us understand the nuanced effects of gender dynamics in peer tutoring settings. All peer mentoring pairs will be classified as either mixed-gender or single-gendered, creating a binary mixed-gender indicator variable.[3] Given this is a binary subgroup analysis, our analysis will follow the approach outlined in the subgroup analysis section. EEF analytical guidance recommends estimating effect sizes both by estimating the primary model (equation 1) on the mixed-gender pairings only, and by estimating an interaction model using the full sample. However, this analysis can only be conducted on the treatment group. For this reason, we conduct just one model, on the full treatment sample, following equation 4.

$$(4)\ Y_{ij} = \beta_0 + \beta_1 MG_{ij} + \beta_3 Z_j + \beta_4 X_{ij} + \beta_5 C_j + \boldsymbol{u_j} + \boldsymbol{e_{ij}}$$

[3] This analysis can only be undertaken if there are at least some mixed-gender pairings. Whilst the analysis can be undertaken with large imbalances in the proportion of mixed-gender and same-gender pairings, this can lead to wide confidence intervals.

Where:

Y_{ij} = AMT score for child *i* in school *j*, at endline;

MG_{ij} = binary variable which indicates whether the child was in a mixed-gender pairing [0] or a single sex pairing [1];

Z_j = school-level characteristic, here the stratifying variable of geographical location (as used for randomisation);

X_{ij} = child-level characteristic, specifically the KS2 maths score which is used here as a baseline;

C_j = Factor variable indicating whether the school *j* was part of cohort 1 or cohort 2;

u_j = setting-level residuals and

e_{ij} = child-level residuals.

All analysis will follow that outlined for the primary outcome above.

Furthermore, literature suggests that the ability of peers can influence outcomes in secondary schools. Therefore, we will examine whether there is natural variation in Year 10 coach maths outcomes at baseline, in order to provide an understanding on whether being matched with a higher ability Year 10 coach has a larger impact on Year 7 coachee outcomes. We propose measuring variation in ability through differences in baseline KS2 SAT Maths attainment. Mixed ability pairings will be defined as those where the Year 10 coach is more than two standard deviations above their Year 7 coachee. Given the selection criteria for each year-group are based on mathematical performance, there may not be sufficient variation in ability pairings to proceed with this analysis. However, we note that the selection criteria for Year 10 coachees is based upon alternative measures of mathematical attainment to the baseline measure, predicted GCSE grades from teacher assessments and Year 9 standardised exams rather than KS2 SAT Maths attainment, so we may still have sufficient variability in ability pairings to allow for an analysis.

Assuming that there is sufficient variation in the baseline maths attainment for Year 10 and Year 7 students, we will aim to conduct exploratory analyses on the differences between same-ability and mixed-ability pairs of coaches and coachees. All peer mentoring pairs will be classified as either mixed-ability or same-ability, creating a binary mixed-ability indicator variable. Given this is a binary subgroup analysis, our analysis will follow the approach outlined in the subgroup analysis section. As with the mixed-gender analysis, this analysis will be carried out on the treatment group only.

$$(5)\ Y_{ij} = \beta_0 + \beta_1 MA_{ij} + \beta_2 Z_j + \beta_3 X_{ij} + \beta_4 C_j + u_j + e_{ij}$$

Where:

Y_{ij} = AMT score for child *i* in school *j*, at endline;

MA_{ij} = binary variable which indicates whether the child was in a mixed-ability pairing [0] or a same ability pairing [1];

Z_j = school-level characteristic, here the stratifying variable of geographical location (as used for randomisation);

X_{ij} = child-level characteristic, specifically the KS2 maths score which is used here as a baseline;

C_j = Factor variable indicating whether the school *j* was part of cohort 1 or cohort 2;

u_j = setting-level residuals and

e_{ij} = child-level residuals.

Analysis of a combined year-group model

We will seek to estimate the impact of Peer-to-Peer coaching at the school level by combining data from the Year 7 and Year 10 cohorts, or year groups. To estimate the overall school level effect across both year groups, we will implement a three-level hierarchical linear model. This approach follows EEF guidance (EEF, 2022) in reflecting the nested structure of the trial, in which pupils (level 1) are nested within year groups (level 2), which in turn are nested within schools (level 3). This approach can be modelled as follows:

$$(6)\ \boldsymbol{Y_{ikj} = \beta_0 + \beta_1 P2P_j + \beta_2 A_{kj} + \beta_3 (P2P_j * A_{kj}) + \beta_4 Z_j + \beta_5 X_{ij} + \beta_6 C_j + u_j + v_{jk} + e_{ikj}}$$

Where:

Y_{ikj} = AMT score for pupil i in year group *k* in school j;

β_0 = cluster-level coefficient on maths skills;

$P2P_j$, = binary indicator equalling 0 if the school is assigned to the control group and 1 if it assigned to the treatment group;

A_{kj} , = binary indicator for year group, taking on value of 1 if the pupil is in Year 7 and 0 if the pupil is in Year 10;

$(P2P_j * A_{kj})$ = interaction term that allows for the effect of the intervention to vary by year-group;

Z_j = setting-level characteristics for school *j*—specifically the stratifying variable of region used in the randomisation;

X_{ikj} = child-level characteristics for child i in setting *j*—specifically the KS2 maths score which is used here as a baseline;

C_j = Factor variable indicating whether the school *j* was part of cohort 1 or cohort 2;

u_j = setting-level residuals

v_j = cohort-level residuals

e_{ij} = pupil-level residuals.

This model in equation 6 allows for the estimation of both a pooled effect and a year-group-specific effect, while accounting for the hierarchical structure of the data and potential

correlation between year-group specific outcomes within schools, especially given the two year groups are not treated independently by the intervention. In alignment with the primary analysis, it retains a cohort indicator as well, allowing for a full reflection of the complex structure of the trial data, where students are clustered into year-groups and schools and schools are split across two treatment cohorts within the same academic year. This model allows for the estimation of effect sizes, whilst using the full statistical power of the trial and accounting for the hierarchical structure of the data and potential correlation between year group cohort-specific outcomes within schools. In accordance with EEF guidance on the use interaction models, we will calculate the year-specific treatment effect sizes from the interaction model and compare these with the effect sizes for each cohort generated in the primary analysis.

In the protocol, we suggested this would either be explored through pooling the data from the year group cohorts or combining the separate primary analyses into a single aggregated effect size using meta-analytic techniques. The three-level model is preferred as it aligns with the nested structure of the trial and provides a statistically robust framework for estimating the school-level impact of the intervention. In line with alternative approaches, such as fixed-effect meta-analysis, this model can still estimate year--specific fixed treatment effects through the interaction term. However, unlike alternative approaches, it models the full nested structure of year-group cohorts within schools, rather than assuming the year-group cohorts are independent samples. We will report the regression output, variance-covariance matrix, and intracluster correlations to allow us to analyse both the fixed effects and independence of year group cohort assumptions.

If the model encounters convergence issues or if the cohort-within-school variance component is negligible, we may consider simplifying to a two-level hierarchical model (pupils nested in schools) which broadly follows the specification in Equation 7 without the inclusion of cohort random effects, v_{kj}. This approach retains the ability to test for differential effects across year group cohorts while reducing the complexity of the random effects structure.

Imbalance at baseline

In theory, a well-conducted randomisation should create groups that are equivalent on observables at baseline, with any imbalance at baseline occurring by chance (Glennerster & Takavarasha, 2013). To check for imbalance at baseline after randomisation, we will produce cross-tabulations of background characteristics and the school and pupil level at randomisation and at analysis. The former informs whether randomisation was successful at obtaining a balanced sample, while the latter provides evidence of whether attrition might have introduced an imbalance.

At the school level, we will examine Ofsted ratings and proportion of children eligible for FSM. These will evaluate the distribution of each characteristic between the control and intervention groups. At the pupil level, balance will be assessed over: FSM, SEND, EAL, and gender status and attainment at baseline. As recommended in the EEF guidance (EEF, 2022), we will report pupil-level pre-tests means and standard deviations, and differences as effect sizes with accompanying confidence intervals. For all categorical variables, we will report counts and percentages in each category. Should there be imbalance on baseline characteristics, we will repeat all primary and secondary analysis controlling for any covariates that demonstrate imbalance at baseline as a robustness check.

Missing data

Missing data can arise through different routes, whether through non-response (a school does not provide some information about a participant) or attrition (a participant or school removes themselves from the sample through opting-out of the evaluation). This can occur at both the school and pupil level. Unfortunately, non-random missingness can introduce bias into the ITT approach outlined in the primary analysis section above. For this reason, we propose a comprehensive missingness analysis, depending on the extent and pattern of missingness, in line with EEF guidance (EEF, 2022).

We will report levels of missingness for each variable through cross-tabulations. For the primary outcome variable, we will also present an attrition flow diagram. If there is less than 5% missingness overall, we propose to only carry out a complete-case analysis, under the assumption that the data are missing completely at random (MCAR). If the missingness exceeds 5% of the sample as randomised, our approach will depend on the pattern of missingness observed, as recommended in the EEF evaluation guidance (Education Endowment Foundation, 2022).

Where missingness for the primary outcome accounts for greater than 5% of observations, we will first conduct a systematic analysis of missingness as a function of all observable characteristics contained within the original dataset and the NPD excerpt. Missingness will be modelled through logistic regression, using a binary outcome variable denoting missingness at endline (where 1=missing; 0=complete). It will mirror the multi-level structure of the models used in the main analysis, with pupils nested within settings. All available setting-level and individual-level variables will be used in this model, not just those pre-specified in equation 1.

On the basis of the logistic model, we will assess the pattern of missingness and take the following approach to the primary analysis:

- If the missing data pattern appears to be unrelated to any observables, or possible unobservable variables (for instance, solely due to pupil absences), we will presume that the data are MCAR and proceed with primary analysis based only on complete cases.

- If there is evidence that missingness is correlated with observable covariates, then data is likely at least missing at random (MAR) and a complete-case analysis will be biased. We will first compare the results of a complete case primary analysis model to the results of the same model with these additional covariates included. If the results of these two models are similar, the complete case analysis is unlikely to be biased, but the interpretation of these results is conditional on the inclusion of these additional covariates. However, should there be significant differences observed between the results of these two models or there is substantial missingness in covariates in the primary analysis model, MI may be warranted. Both full-information maximum likelihood (FIML) and MI have been shown to be broadly equivalent (Lee & Shi, 2021). We will follow the guidelines for MI recommended in Jakobsen et al., (2017). This involves first generating several datasets, each containing a plausible value for the missing data, before repeating the primary analysis on each of the datasets generated, before pooling the results into one multiple-imputation analysis which provides a robust estimate of the effect size, even in the presence of data missing at random.

- If missing data seems likely to depend on unobserved variables even after controlling for all observable covariates (either because the logistic model could not predict missingness based on observable covariates or any analysis conducted under the assumption of MAR suggests some missingness may be correlated with unobservable factors), then the data is likely missing not at random (MNAR). Complete case analysis is likely to be biased, but MI will not be sufficient to correct for this. Instead, sensitivity analysis will be carried out using the approach laid out by Carpenter et al., (2007). and reported alongside the headline impact estimates.

Compliance analysis

The primary analysis captures the averaged effect of offering the Peer-to-Peer intervention on an intent to treat basis. However, we will also examine treatment effects in the presence of non-compliance. To do this, we will include analysis on the impact of compliance on our primary outcome variable. Compliance is defined at the pupil level as a binary variable. Following discussions between the EEF, CoachBright and the Evaluation team, minimum compliance has been defined as follows:

- *For Year 10 students:* at least 70% pupil attendance at coach training and 70% pupil attendance of the weekly coaching sessions
- *For Year 7 students*: at least 70% pupil attendance of the weekly coaching sessions.

This will be measured through coach and coachee attendance logs that will be collected by CoachBright and shared with RAND Europe and UoL before analysis takes place.

We will calculate the Complier Average Causal Effect (CACE) through a two-stage least squares (2SLS) instrumental variable (IV) approach. The first stage of this approach will involve regressing the compliance variable on allocation to treatment with the same pupil and school level covariates and multi-level structure used in the primary analysis. This will provide an estimate of how assignment of pupils and settings to receive P2P encourages uptake of intervention. Results from this stage will effectively provide an overall 'compliance rate' and will be estimated using the following equation:

$$(7)\ \boldsymbol{Y}_{ij} = \boldsymbol{\beta_0} + \boldsymbol{\tau}\, \boldsymbol{P2P_j} + \boldsymbol{\beta}_1 \boldsymbol{Z_j} + \boldsymbol{\beta}_2 X_{ij} + \boldsymbol{\beta}_3 C_j + \boldsymbol{u_j} + \boldsymbol{e_{ij}}$$

Where:

Y_j = Binary Compliance at the pupil level 'i' for school 'j';

β_0 = Intercept;

$P2P_j$ = Binary indicator assigned to setting 'j' indicating if it is treatment [1] or control [0];

Z_j = school-level characteristics 'j'

X_{ij} = child-level characteristic, specifically the KS2 maths score which is used here as a baseline;

C_j = Factor variable indicating whether the school *j* was part of cohort 1 or cohort 2;

u_j = setting-level residuals and

$\boldsymbol{e_{ij}}$ = pupil level residuals.

Results for the first stage, and the associated F stat, will be reported.

The second stage of the IV estimation predicts the outcome as a function of all covariates included in equation 1 but substitutes the treatment indication (P2P in equation 1) with the compliance rate estimated in the first regression. Due to ease of estimation, we will use an OLS IV approach estimate CACE, clustering the errors at the school level. This does not mimic exactly the multi-level hierarchical analysis employed throughout the rest of analysis but still controls for intracluster correlations at the setting level to ensure appropriate standard errors and confidence intervals are used. This model will be estimated for the primary outcome measure only.

However, this method uses a binary measure whereby pupils would be deemed compliant only if they met the exact attendance percentage outlined. The dichotomisation of a multi-dimensional conceptualisation of compliance poses concerns for one of the main assumptions made for CACE analysis, the exclusion restriction, as it would mean that partially compliant pupils would be deemed non-compliant. The exclusion restriction requires that treatment allocation cannot influence pupil outcomes other than through the defined compliance metric. This would mean that improved maths outcomes from pupils that were allocated treatment but did not reach 70% attendance show that treatment can still affect the final outcome and thus the exclusion restriction assumption has not been satisfied. This violation would result in a downward bias in CACE.

An alternative approach to compliance, that avoids the possible introduction of bias outlined above, is to instead use a continuous metric of compliance. To help understand the risk of bias in the analysis outlined above, we will repeat the analysis using continuous compliance measures, such as the number of coaching sessions attended and, for year 10 students only, through the number of coach training sessions attended. We will re-run the model outlined in equation 8 for each proposed continuous compliance measure. However, whilst switching to a continuous measure of compliance avoids introducing bias through dichotomisation, the multi-dimensional aspect of compliance in P2P means the exclusion restriction could still be violated for year 10 pupils. For instance, year 10 pupils who do not attend any coaching training sessions may still see improvements in their maths attainment through attendance at coaching sessions with year 7 coachees. This opens up a causal pathway by which treatment allocation can still affect maths attainment, even if there is non-compliance on one dimension. We will report the full distribution of each compliance metric and discuss the implications of this, with respect to the satisfaction of the *exclusion restriction* and the validity of a 2SLS approach. Any estimates will be caveated with a discussion of the risk of a downward bias in CACE.

Compliance and dosage analysis will be undertaken in R or Stata, version 18 or higher.

Intra-cluster correlations (ICCs)

The ICC is a crucial metric for trials involving clusters. It quantifies the fraction of variance in a specific outcome attributable to differences between clusters, rather than variance occurring within these clusters.

The ICC applied in the power calculations detailed in the section about Sample Size and Power Calculations Overview, and at the protocol stage, is set at 0.07 for all pupils and at the same value for FSM-eligible pupils.

In the final report, we will present ICCs as calculated at different stages of the evaluation: at the protocol stage, at the time of randomisation and at the analysis stage. The ICC at the analysis stage will focus on the primary outcome measure. Its calculation will involve two approaches:

(i) using the model corresponding to Equation (1), and

(ii) employing a model akin to that in Equation (1) but without any covariates, i.e. only containing the outcome and treatment indicator.

This second model accounts for the clustering of pupils in schools and is referred to as the 'empty model'.

ICCs will be estimated using Stata's *estat icc* command, or an R equivalent, using unconditional variance.

Effect size calculation

As outlined in the Analysis section, unless otherwise stated, we will calculate effect sizes (ES) for cluster-randomised trials as outlined in the EEF evaluator guidance (EEF, 2022), and adapted from (Hedges, 2007):

$$ES = \frac{(\bar{Y}_T - \bar{Y}_c)_{adjusted}}{\sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}}$$

Where $(\bar{Y}_T - \bar{Y}_c)_{adjusted}$ is the mean difference between the intervention and the control group adjusted for baseline characteristics and $\sqrt{\frac{(n_1-1)sd_1^2+(n_2-1)sd_2^2}{n_1+n_2-2}}$ is an estimate of the pooled unconditional population standard deviation.

The pooled unconditional standard deviation is the weighted average of standard deviations of treatment and control (Coe, 2002). The pooled unconditional standard deviation across the two trial arms is used in the denominator, as we assume the standard deviations of both the treatment and control groups are drawn from the same underlying population distribution. If there is cause to question this assumption at the analysis stage, we will use the unconditional standard deviation of the control group, in line with EEF guidance.

From the primary outcome model, we will take each group's adjusted mean and variance to calculate the effect size. This variance will be the total variance (across both pupil and school levels, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchinson & Styles, 2010). A 95% CI for the ES, which takes into account the clustering of pupils in schools, will also be reported. Effect sizes will be calculated for each of the models estimated and converted into months progress.

References

Al Umairi, K.S.S., Salleh, U.K.M. and Zulnaidi, H (2023). Adaptation of the sources of the mathematics self-efficacy scale for Oman: A validation study. *EURASIA Journal of Mathematics, Science and Technology Education*, *19*(9).

Angrist, J. (2006). Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology 2*, 23-44. https://doi.org/10.1007/s11292-005-5126-x

Angrist, J., & Krueger, A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings?. *The Quarterly Journal of Economics*. 106. 979-1014. https://doi.org/10.2307/2937954

Aziz, T.A. and Azhar, E. (2021). The validity and reliability study of the Indonesian version of the sources of mathematics self-efficacy scale. In *AIP Conference Proceedings* (Vol. 2331, No. 1, p. 020038). AIP Publishing LLC.

Bandura, A. (1997). *Self-efficacy: The exercise of control* (Vol. 11). Freeman.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. Sage Publications, Inc.

Carpenter, J., Kenward, M., & White, I. (2007). Sensitivity analysis after multiple imputation undermissing at random: a weighting approach. *Statistical Methods in Medical Research*, 16, 259-275

Clarke, D. (2021). RWOLF2: Stata module to calculate Romano-Wolf stepdown p-values for multiple hypothesis testing. Statistical Software Components S458970, Boston College Department of Economics.

Clarke, D., Romano, J., & Wolf, M. (2019). The Romano-Wolf Multiple Hypothesis Correction in Stata. IZA Institute of Labor Economics.

Coe, R. (2002). It's the Effect Size, Stupid. What Effect Size Is and Why It Is Important. Paper Presented at the British Educational Research Association Annual Conference, Exeter, 12-14 September 2002.

Deeks, J.J., Higgins, J.P., Altman, D.G., McKenzie, J.E., Veroniki, A.A., editors (2024). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins, J.P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., et al, editors (2024). Cochrane Handbook for Systematic Reviews of Interventions version 6.5. Cochrane. https://www.cochrane.org/authors/handbooks-and-manuals/handbook

Demack, S., Culliney, M., Boyan, M., Wolstenholme, C. (2022). Realistic Maths Education: Evaluation Report. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/realistic-maths-education

Education Endowment Fund. (2022). Statistical analysis guidance for EEF evaluations. Retrieved from: Education Endowment Foundation. EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf

George, A., Goldie, S. & Dixon, T. (2024). Access Mathematics Tests Test Guidance 2025. Hachette Learning. AMT Test Guidance 2025

Glennerster, R. & Takavarasha, K. (2013) Running Randomized Evaluations: A Practical Guide. London: Princeton University Press.

Graham, J., Olchowski, A., & Gilreath, T. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, 8, 206-213.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32, 4:. 341 – 370. https://doi.org/10.3102/1076998606298043

Husain, F., Bartasevicius, V., Marshall, L., Chidley, S. & Forsyth, E. (2019). Fit to Study: Evaluation Report. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/fit-to-study

Hutchison, D., & Styles, B. (2010). A guide to running randomised controlled trials for educational researchers. Slough: NFER.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC medical research methodology*, 17(1), 1-10.

Jerrim, J., Austerberry, H., Crisan, C., Ingold, A., Morgan, C., Pratt, D., Smith, C. & Wiggins, M. (2015). Mathematics Mastery: Secondary Evaluation Report. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/mathematics-mastery-secondary

Kandemir, M.A. and Akbaş-Perkmen, R. (2017). Examining validity of sources of mathematics self-efficacy scale in Turkey. *European Journal of Education Studies*.

Kontas, H. and Özcan, B. (2017). Adapting sources of middle school Mathematics self-efficacy scale to Turkish culture. *International Journal of Evaluation and Research in Education*, 6(4), pp.288-294.

Navarro, M. Larrain, M. & Pezoa J. (2025). Internal structure of a scale of sources of self-efficacy in mathematics among middle school students. *International Journal of Educational Research,* 134. https://doi.org/10.1016/j.ijer.2025.102771

Raabe, I. J., Boda, Z., & Stadtfeld, C. (2019). The Social Pipeline: How Friend Influence and Peer Exposure Widen the STEM Gender Gap. *Sociology of Education*, 92(2), 105-123. https://doi.org/10.1177/0038040718824095

Ray, D., Muñoz, A., Zhang, M., Li, X., Chatterjee, N., Jacobson, L. P., & Lau, B. (2022). Meta-analysis under imbalance in measurement of confounders in cohort studies using only summary-level data. *BMC Medical Research Methodology*, 22, 143. https://doi.org/10.1186/s12874-022-01614-9

Riegle-Crumb, C., Farkas, G., & Muller, C. (2006). The Role of Gender and Friendship in Advanced Course Taking. *Sociology of Education*, 79(3), 206-228. https://doi.org/10.1177/003804070607900302

Riegle-Crumb, C. & Morton, K. (2017). Gendered Expectations: Examining How Peers Shape Female Students' Intent to Pursue STEM Fields. *Frontiers in Psychology,* 15(8), 329. https://doi.org/10.3389/fpsyg.2017.00329

Röver C, Knapp G, Friede T. (2015) Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol*. https://doi.org/10.1186/s12874-015-0091-1

Schwarzer, G. (2025). Package 'meta' (R package). https://cran.r-project.org/web/packages/meta/meta.pdf

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Maturi, T., Higgins, S., Einbeck, J., Culliney, M., & Demack. S. (2023). Improving Power Calculation in Education Trials.

Smith, E., & Andersen, I. G. (2022). Do Same-Gender Peers in the Classroom Have Heterogeneous Impacts on Male and Female Students? *Socius*, 8. https://doi.org/10.1177/23780231221105378

Speciani, E. R., Cardamone, I., Merewood, J., Dysart, E., and Tracey, L. (2025). Peer-to-Peer Coaching evaluation protocol. London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/peer-to-peer-coaching-trial

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2012). Metacognition and self-regulated learning constructs. *Educational Research and Evaluation*, 8(2), 117-139

Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., Wood, A., & Carpenter, J. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potentials and pitfalls. *British Medical Journal*, 338.

Tenenbaum, H. R., Winstone, N. E., Leman, P. J., & Avery, R. E. (2019). How effective is peer interaction in facilitating learning? A meta-analysis. *Journal of Educational Psychology*, 112(7), 1303-1319. https://doi.org/10.1037/edu0000436

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology,* 34(1), 89–101. https://doi.org/10.1016/j.cedpsych.2008.09.002

Vallis, D., Singh, A., Uwimpuhwe, G., Higgins, S., Xiao, Z., De Troyer, E., & Kasim, A. (2022), EEFANALYTICS: Stata module for Evaluating Educational Interventions using Randomised Controlled Trial Designs. https://EconPapers.repec.org/RePEc:boc:bocode:s458904

Van Der Stuyf, R.R. (2002). Scaffolding as a teaching strategy. Adolescent learning and development, Section 0500A.

Watson, J. A. (2024). cectStepdown: Stepdown Procedure for Cluster Randomised Trials (R package). https://cran.r-project.org/web/packages/crctStepdown/index.html

Wake, G., Hodgen, J. Adkins, M., Ainsworth, S., & Evans, S. (2022). Young Enterprise: Mathematics in Context Evaluation Report. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/maths-in-context

Webb, N. M., & Mastergeorge, A. (2003). Promoting effective helping behaviour in peer-directed groups. *International Journal of Education Research*, 39(1-2), pp. 73-97. https://doi.org/10.1016/S0883-0355(03)00074-0

Wiggins, M., Jerrim, J., Tripney, J., Khatwa, M., & Gough, D. (2019). The Rise Project: Evaluation Report. The Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/the-rise-project-evidence-informed-school-improvement

Yaşlıoğlu, M. and Yaşlıoğlu, D.T. (2020). How and when to use which fit indices? A practical and critical review of the methodology. *Istanbul Management Journal*, (88), pp.1-20.

Zakariya, Y. (2022). Improving students' mathematics self-efficacy: A systematic review of intervention studies. *Frontiers in Psychology,* 13. https://doi.org/10.3389/fpsyg.2022.986622

Appendix A: SMSES pilot

A pilot was conducted before endline to ensure that the Sources of Middle School Maths Self Efficacy Scale (SMSES) measure was valid and reliable with our chosen population. The SMSES was refined and validated through a multi-phase study by Usher and Pajares (2009), involving US middle school students. The final version consists of 24 items that are divided across four subscales, each corresponding to one of Bandura's sources – mastery experience (ME), vicarious experience (VE), social persuasions (P), physiological state (PH) (Bandura, 1997).

Considering that the original scale was created through testing on US children, there have been a number of studies assessing its effectiveness in other country contexts. This includes Kandemir and Akbaş-Perkmen (2017), who were examining the validity of the scale in Turkey. They examined the scale on 616 middle-school students, with questions adapted to better fit the Turkish context instead of a direct translation from the English version. This study found the original four-factor model to be a valid and reliable measure. Alternatively, Navarro et al. (2025) constructed the model for Chile, determining that the four-factor model with significant changes to the number of items included produced a better fit. This used 20 items, excluding one item from each source with the lowest factor loading. One item that was highlighted as problematic was ME4 (*Even when I study hard, I do poorly in math*), which was excluded from Navarro et al. due to its high factor loading on physiological states, the reasoning suggested being that unsuccessful effort was interpreted as frustration. This was also observed in Oman (Al Umairi et al., 2023). Similarly, some studies have had difficulties with item P18 (*My classmates like to work with me in math because they think I'm good at it*), where students have interpreted this as evidence of ability (mastery experience) rather than the belief of their classmates (social persuasion) (Navarro et al., 2025; Al Umairi et al., 2023; Kandemir and Akbaş-Perkmen, 2017). Without a study looking at UK schools, it is difficult to determine if these same issues, or different cultural differences, will challenge the validity and efficacy of the SMSES.

The University of Leeds ran the pilot in three schools for 66 children (17 Year 7 pupils and 49 Year 10 pupils). The version of the test featured minor spelling adaptations to suit the UK without altering the scale – for instance changing the wording from "math" to "maths".

To collect evidence on the structure of the scale, confirmatory factor analysis (CFA) was conducted using the structure proposed by Usher and Pajares (2009). The success of this structure was then tested through EFA on the same sample. Bartlet's test and the Kaiser-Meyer-Olkin (KMO) measure was estimated to ensure that the sample was adequate for EFA. Although it is usually recommended that CFA and EFA be performed on different samples, the current sample size suggestion for multivariate data analysis is 20 subjects per item (Navarro et al., 2025). This means that the pilot sample size is already underpowered and further dividing it would compromise the outcome further.

Various fit indices were used to assess the goodness of fit for both CFA and EFA. The first of these is the SRMR index which is the standardised root mean squared error, measuring the average distance between the observed correlations and the model-predicted correlations. This value is considered within a good range if it is less than 0.08. The next is the RMSEA, root mean square error of approximation showing how well the model would fit the population covariance matrix, considered good if it is less than 0.05. The CFI, which is the comparative fit index, should be above 0.90 to be considered an acceptable fit. (Yaşlıoğlu, and Yaşlıoğlu, 2020) Finally, we will test the

chi-square/df measure, where a smaller value indicates better fit through less residual misfit per parameter.

The internal consistency of the scale by source was evaluated using Cronbach's Alpha. All analysis was conducted in STATA 19.

Table 10. Means, standard deviations, and correlations for sources of self-efficacy items

Item	M	SD	ME1	ME2	ME3	ME4	ME5	ME6	VE7	VE8	VE9	VE10	VE11	VE12	P13	P14	P15	P16	P17	P18	PH19	PH20	PH21	PH22	PH23	PH24
ME1	3.3	1.1	**.92**																							
ME2	3.1	1.3	.74	**.87**																						
ME3	3.6	1.3	.62	.53	**.73**																					
ME4	3.8	1.1	.49	.51	.25	**.64**																				
ME5	3.4	1.1	.90	.76	.64	.54	**.94**																			
ME6	2.9	1.2	.86	.81	.60	.47	.86	**.92**																		
VE7	2.7	1.3	.49	.55	.29	.41	.49	.55	**.78**																	
VE8	3.5	1.2	.54	.51	.48	.46	.51	.49	.61	**.82**																
VE9	3.4	1.2	.55	.58	.40	.63	.56	.56	.60	.66	**.86**															
VE10	3.4	1.1	.48	.53	.33	.55	.49	.51	.49	.66	.73	**.82**														
VE11	3.3	1.2	.61	.68	.41	.50	.62	.73	.60	.49	.65	.62	**.81**													
VE12	3.5	1.2	.51	.43	.33	.40	.52	.49	.44	.53	.58	.50	.58	**.75**												
P13	3.6	1.1	.65	.74	.40	.61	.60	.61	.55	.54	.65	.55	.71	.43	**.85**											
P14	3.1	1.4	.69	.77	.48	.45	.71	.70	.49	.37	.51	.39	.61	.38	.70	**.91**										
P15	3.4	1.4	.59	.65	.43	.42	.60	.60	.47	.40	.48	.35	.64	.47	.67	.77	**.83**									
P16	3.3	1.3	.74	.77	.61	.47	.79	.74	.52	.52	.57	.49	.69	.48	.75	.79	.73	**.90**								
P17	3.1	1.3	.60	.65	.38	.47	.63	.54	.57	.43	.48	.40	.55	.43	.67	.75	.61	.71	**.86**							
P18	3.1	1.2	.53	.67	.39	.36	.51	.51	.56	.46	.48	.39	.50	.32	.71	.67	.52	.69	.79	**.83**						
PH19	3.4	1.4	.39	.31	.56	.13	.34	.42	.19	.33	.36	.17	.42	.24	.26	.33	.46	.36	.21	.20	**.92**					
PH20	3.2	1.4	.48	.36	.62	.13	.39	.49	.21	.38	.32	.31	.41	.33	.35	.38	.51	.41	.27	.20	.81	**.89**				
PH21	3.6	1.3	.33	.26	.50	.15	.27	.32	.12	.36	.37	.15	.38	.31	.35	.34	.49	.38	.24	.19	.86	.81	**.93**			
PH22	3.5	1.3	.38	.41	.61	.37	.36	.43	.24	.45	.40	.26	.43	.32	.38	.46	.47	.42	.34	.23	.79	.76	.77	**.88**		
PH23	3.5	1.5	.26	.17	.53	.11	.23	.30	.34	.34	.34	.15	.39	.26	.31	.21	.42	.34	.17	.12	.74	.72	.75	.70	**.87**	
PH24	3.9	1.4	.26	.25	.56	.17	.27	.27	.16	.36	.37	.15	.37	.25	.33	.32	.42	.40	.26	.21	.83	.73	.87	.79	.79	**.92**

N = 66. Item-total correlations between each item and its subscale counterparts appear on diagonal. Items within each given subscale appear in greyscale.

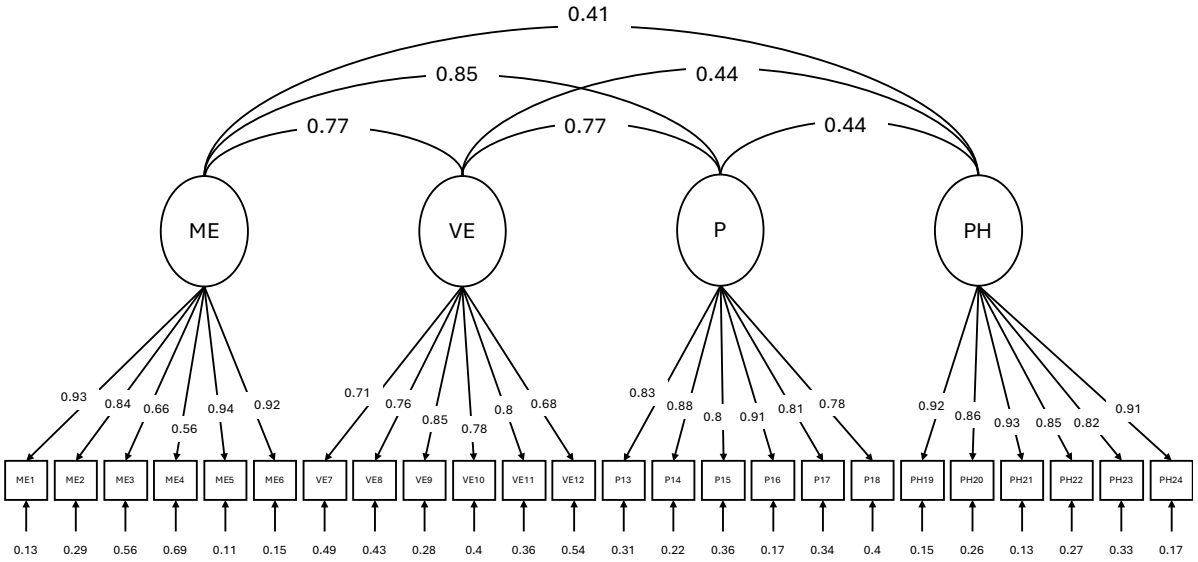
Table 10 presents the correlation matrix and item-total correlations for the dependent variables in the model. Correlations between the six items in each of the source sub-scales ranging from 0.25 to 0.90. Each subscale showed a high level of internal consistency, with Cronbach's alpha coefficients of 0.91 for Mastery Experience, 0.89 for Vicarious Experience, 0.93 for social persuasions, and 0.95 for physiological state. These figures are all above those provided by Usher and Pajares (2009) in the original validation study.

Figure 1 shows the results for the 4 factor, 24 item, measurement model. All standardised factor loadings in the model were significant at the $p < 0.001$ level and ranged in magnitude between 0.56 and 0.94. Our results are consistent with other studies, showing lower scoring on ME4 (*Even when I study hard, I do poorly in math(s)*), which is the 0.56 factor loading. While this is lower than the magnitude on other items, it is above the 0.55 cutoff suggested by Usher and Pajares (2009). The pilot does indicate that while P18 (*My classmates like to work with me in math because they think I'm good at it*), that has been problematic in other studies, has the weakest loading in its subgroup, it is a strong representation of social persuasions at 0.78.

The four sources showed intercorrelations ranging between 0.41 and 0.85. The largest correlation is seen between mastery experience and social persuasions; the same as reported by Usher and Pajares (2009), which is described as unsurprising given that students who perceive their last

performances in maths as successful are likely to receive frequent praise on those performances.

Figure 1. Measurement Model for 24 item SMSES



However, confirmatory factor analysis (CFA) suggested that this four-factor model did not show an acceptable fit for these data. As can be seen in 10, while the Chi-square/df value for this model is low, as is CFI at only 0.866. Both the RMSEA and SRMR are too high with values that are over double the cutoff for acceptable fit. To determine the cause of the fit issues, exploratory factor analysis (EFA) was conducted. First, the KMO sample adequacy test was performed, with the value of 0.871 – any value above 0.8 is considered to indicate adequate sampling (Navarro et al., 2025). Bartlett's test of sphericity additionally showed that the correlation matrix was significantly different from an identity matrix, allowing for EFA.

Table 10. Goodness of Fit Measures

Model	**N**	**ChiSquare/df**	**CFI**	**RMSEA**	**SRMR**
Usher and Pajares (2009)	803	2.44	0.96	0.04	0.04
Al Umairi et al. (2023)	700	3.43	0.928	0.059	0.0478
Kandemir and Akbaş-Perkmen (2017)	616	3.38	0.98	0.06	0.05
Current Study					
Four Factor	66	1.8	0.866	0.115	0.088
EFA Three Factor	66	1.97	0.836	0.127	0.086
ME only	66	1.2	0.993	0.06	0.031
VE only	66	1.65	0.971	0.101	0.04
P only	66	3.08	0.941	0.182	0.046
PH only	66	1.35	0.992	0.07	0.02

EFA indicated that a three-factor model would be a better fit but was borderline on the suggestion of a four-factor model, with the eigenvalue for a fourth factor ranging between 0.9 and 1.07. The three-factor model largely kept the sources together, combining mastery experience and social persuasions, with ME4 shifting to vicarious experience. This loading could be expected considering the strength of the intercorrelation between the mastery experience and social persuasions factors in the four-factor model. The goodness of fit measures, however, indicate that this structure is weaker than the four-factor model. The RMSEA is higher at 0.127 compared to the 0.115 of the previous model. Additionally, the CFI is even lower at only 0.836 and a higher chi-square/df value. There are some minor improvements in the SRMR, dropping 0.002 lower.

The lack of improvement following the EFA adjustment could mean that it is the small sample size that is causing fit issues. With the generally accepted minimum size of 20 observations for each question, a sample size of at least 480 would be required for a fully powered test. Table 11 shows the goodness of fit measures for a selection of other studies that opted to use the four-factor model of the SMSES. All are far above this minimum value required for a well powered calculation. As such, a better, although still underpowered, measurement of the sources could be obtained when each factor was run independently. All showed a strong goodness of fit in their CFI and SRMR values. The chi-squared/df measure is even lower for all but social persuasions. The RSMEAs do appear to be too high, but all have a notable probability of being below the acceptable 0.06 measurement with their confidence intervals.

The CFA for this pilot suggests that the SMSES, estimated through a four-factor model, is likely both a valid and reliable measure to use in the UK context. Minimal changes are required for the scale other than the rewording of items to change spelling differences.

Appendix B: Secondary outcomes sample size calculations

Table 12. Secondary outcomes sample size calculations

		Self-Efficacy				Metacognition			
		OVERALL		FSM		OVERALL		FSM	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Year 7		0.226	0.241	0.247	0.263	0.223	0.238	0.244	0.260
Year 10		0.267	0.285	0.284	0.303	0.264	0.281	0.281	0.300
Number of schools	intervention	23	20	23	20	23	20	23	20
	control	23	20	23	20	24	21	24	21
	total	46	40	46	40	47	41	47	41
Number of pupils	intervention	15	15	13	13	15	15	13	13
	control	15	15	13	13	15	15	13	13
	total	30	30	26	26	30	30	26	26