Evaluation of the Mastering Maths programme: a two-arm cluster randomised trial Statistical Analysis Plan

Evaluator: NatCen Social Research
Principal investigator: Helena Takala

Education Endowment Foundation

PROJECT TITLE	Evaluation of the Mastering Maths programme: a two-arm cluster randomised trial
DEVELOPER (INSTITUTION)	University of Nottingham (UoN)
EVALUATOR (INSTITUTION)	National Centre for Social Research (NatCen)
PRINCIPAL INVESTIGATOR(S)	Helena Takala
PROTOCOL AUTHOR(S)	Tien-Li Kuo, Enes Duysak
TRIAL DESIGN	Two-arm cluster randomised trial
TRIAL TYPE	Effectiveness trial
STUDENTS AGE RANGE AND KEY STAGE	Students aged 16-19
NUMBER OF TEACHERS	161 teachers in Further Education (FE) settings, with a maximum of two teachers eligible from a single college setting
NUMBER OF STUDENTS	~10,309 students
PRIMARY OUTCOME MEASURE AND SOURCE	GCSE Maths score, collected directly from FE colleges or exam boards
SECONDARY OUTCOME MEASURE AND SOURCE	GCSE Maths grade, collected directly from FE colleges or exam boards Self-confidence in maths measured with the Attitude towards Maths Inventory (ATMI) Self-efficacy in maths measured with University of Manchester's self-efficacy questionnaire

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	02/12/2025	N/A

Table of contents

SAP version history 2
Table of contents 3
Abbreviations 4
Introduction 5
Design overview 6
Research questions 6
Recruitment 8
Randomisation 8
Primary outcome 10
Secondary outcomes 10
Sample size calculations overview 13
Planned sample size 13
Updated sample size calculations 13
Analysis 16
Primary outcome analysis 16
Secondary outcome analysis 17
Subgroup analyses 18
Compliance analysis 19
Additional analyses 21
Imbalance at baseline 24
Missing data 25
Intra-cluster correlations (ICCs) 26
Effect size calculation 26
References 28
Appendix A: Randomisation syntax 30

Abbreviations

2SLS	Two-stage Least Square
CACE	Complier Average Causal Effect
EEF	Education Endowment Foundation
FE	Further Education
GCSE	General Certificate of Secondary Education
ICCs	Intra-cluster Correlations
IMD	Index of Multiple Deprivation
ITT	Intention-to-treat
IV	Instrumental Variable
MDES	Minimum Detectable Effect Size
MoU	Memorandum of Understanding
NatCen	National Centre for Social Research
NPD	National Pupil Database
RQ	Research Question
UoN	University of Nottingham
SAP	Statistical Analysis Plan
TAU	Teaching-as-usual

Introduction

This document outlines the analyses that will be conducted for the impact evaluation of Mastering Maths. Mastering Maths is a continuing professional development (CPD) programme developed by academics at the University of Nottingham (UoN) to upskill Further Education (FE) teachers in teaching post-16 GCSE Maths resit students, addressing the limited professional development available. The programme aims to strengthen students' understanding of fundamental mathematical concepts, build fluency, and support them in achieving a grade 4 or above in their GCSE Maths resit exams.

The intervention targets teachers of GCSE Maths resit classes for 16-19-year-old students. All Mastering Maths activities took place in-person, including lead teacher training, professional development for intervention teachers, lesson study groups, and the teaching of five Mastering Maths lessons between November 2024 and March 2025. More details on the Mastering Maths programme can be found in the evaluation protocol.[1]

The evaluation was conducted as a two-arm cluster randomised controlled effectiveness trial of the Mastering Maths programme on GCSE Maths scores. The primary population included students resitting their GCSE Maths exam that did not achieve a grade 4 or higher at age 16 or in any subsequent resits. The primary outcome of interest is students' GCSE Maths scores. Secondary outcomes include a) the probability of students achieving a grade 4 or higher; b) the probability of students moving up at least one grade compared to their most recent GCSE Maths exam; c) student's self-confidence and d) self-efficacy in maths. The baseline measure for both primary and secondary outcomes is Key Stage 2 (KS2) Maths attainment in the National Pupil Database (NPD).[2]

UoN recruited eligible teachers for the trial, up to two teachers per FE setting, and the evaluation team randomised them into two groups: the Mastering Maths intervention group and the teaching-as-usual (TAU) group. Teachers in the intervention group will participate in Mastering Maths while those in the TAU group will follow their usual teaching practices without participating in the professional development or lesson study sessions during the 2024/2025 academic year. Randomisation was stratified by FE setting, defined as groups of GCSE Maths resit teachers connected due to the organisational structure and/or geographic location of their FE college (please see Randomisation section for more details).

To encourage participation and retention throughout the evaluation, UoN will provide financial incentives to settings. Settings who nominated one teacher will receive the Mastering Maths programme and £1,250 for teachers to cover classes if allocated to the intervention group, or £1,000 as a thank you for participation and facilitating data collection if allocated to the TAU group. Settings nominating two teachers will have one teacher allocated to the intervention group and one to the TAU group. Therefore, they will receive the Mastering Maths programme, £1,250 for covering the intervention teachers' participation in the professional development, and an additional £500 as a thank you for the participation of the TAU teacher and their support with the associated collection of student data.

[1] Mastering Maths evaluation protocol. Available at https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/mastering_maths_-_evaluation_protocol.pdf?v=1729167831

[2] We will use KS2 maths attainment as a baseline measure for all secondary outcomes, including attitudes towards maths. This is because the relevant surveys will only be administered at endline, to reduce burden on teachers and students. Nonetheless we believe students' prior attainment in maths is correlated with their attitudes towards the subject and can thus explain some of the variance in the model.

Design overview

The impact evaluation is designed as a two-arm cluster randomised controlled effectiveness trial to assess the effect of Mastering Maths programme on students' maths skills and GCSE Maths score.

Research questions

This impact evaluation aims to answer the following research questions (RQ):

Primary research question

RQ1: What is the impact of Mastering Maths on 16- to 19-year-old students' GCSE Maths scores relative to those students receiving teaching-as-usual? (primary outcome)

Secondary research question

RQ2: What is the impact of Mastering Maths on the probability of students achieving a grade 4 or higher on their GCSE Maths resit relative to those students receiving teaching-as-usual? (secondary outcome)

RQ3: What is the impact of Mastering Maths on the probability of students moving up at least one grade on their GCSE Maths resit relative to those students receiving teaching-as-usual? (secondary outcome)

RQ4: What is the impact of Mastering Maths on students from disadvantaged backgrounds as measured by prior FSM status relative to those students receiving teaching-as-usual? (secondary outcome)

RQ5. What is the impact of Mastering Maths on students from disadvantaged backgrounds as measured by the Index of Multiple Deprivation relative to those students receiving teaching-as-usual? (secondary outcome)[3]

RQ6. What is the impact of Mastering Maths on students with lower prior attainment relative to those students receiving teaching-as-usual? (secondary outcome)

RQ7: What is the impact of Mastering Maths on students' self-confidence in maths relative to those students receiving teaching-as-usual? (secondary outcome)

RQ8: What is the impact of Mastering Maths on students' self-efficacy in maths relative to those students receiving teaching-as-usual? (secondary outcome)

Table 1 Trial design

Trial design, including number of arms	Two-arm cluster randomised controlled effectiveness trial
Unit of randomisation	Teacher level
Stratification variables (if applicable)	Further education (FE) setting

[3] The funding formula for all providers delivering 16-19 education includes a disadvantage funding element based on students' economic deprivation (via Index of Multiple Deprivation) as well as low prior attainment. Details are covered in the Secondary outcome section.

Primary outcome	variable	RQ1 – GCSE Maths score
	measure (instrument, scale, source)	RQ1 – GCSE Maths standardised raw score (z-score by exam board), collected directly from FE colleges or exam boards
	variable(s)	RQ2 – Probability of achieving a grade 4 or higher in GCSE maths RQ3 – Probability of moving up at least one grade in GCSE Maths RQ4-6 – GCSE Maths score RQ7 – Self-confidence in maths RQ8 – Self-efficacy in maths
Secondary outcome(s)		RQ2 – GCSE Maths grade, collected directly from FE colleges or exam boards, a binary variable equal to 1 if GCSE Maths grade is 4 or higher and 0 if it is 3 or below RQ3 – GCSE Maths grade, collected directly from FE colleges or exam boards, a binary variable equal to 1 if student moves up at least one GCSE Maths grade and 0 if not RQ4-6 – GCSE Maths standardised raw score (z-score by exam board), collected directly from FE colleges or exam boards RQ7 – Attitude towards Maths Inventory (ATMI) anglicised version, Likert scale (1-5) RQ8 – University of Manchester's self-efficacy questionnaire, confidence segment, Likert scale (1-4)
	measure(s) (instrument, scale, source)	
Baseline for primary outcome	**variable**	Key Stage 2 (KS2) maths attainment
	measure (instrument, scale, source)	KS2 National Curriculum Test, Total marks achieved in maths test ('KS2_MATMRK' variable, 0-110)), National Pupil Database (NPD)[4]
Baseline for secondary outcome	**variable**	KS2 maths attainment
	measure (instrument, scale, source)	KS2 National Curriculum Test, Total marks achieved in maths test ('KS2_MATMRK' variable, 0-110)), National Pupil Database (NPD)

[4] At protocol stage, we planned to use the Maths scaled score (KS2_MATHSCORE_noSpeccon variable, 0-120) due to uncertainties surrounding the availability of raw maths attainment scores in the NPD data. However, after communicating with the Department for Education, we confirmed that the raw scores, labelled as 'marks', are available in the NPD. We then choose using the raw score ('KS2_MATMRK' variable) as a baseline measure, in line with the EEF's statistical guidance (2022).

Recruitment

Many FE colleges are large, often comprising of settings on different sites and maybe even in different geographical areas. GCSE Maths resit classes are administered at a setting level and taught in separate groups. There may be some interaction between groups of teachers, but it is unlikely that approaches to teaching maths would be discussed in any substantive way. For the purpose of this study, each group of teachers was considered a distinct setting, and each FE college may have more than one setting. Within a single FE college, multiple teachers may have been eligible to participate in the trial if their groups are geographically or structurally distinct.

UoN recruited up to two GCSE Maths resit teachers from each setting.[5] Settings were encouraged to have two teachers participating in the trial but were still accepted if they only signed up one teacher for the trial.

UoN distributed the invitation to participate widely and through a range of channels. 286 FE settings submitted an Expression of Interest through the EEF website or directly to the UoN. All these settings were sent a Memorandum of Understanding (MoU). Settings that did not return a signed MoU provided a range of reasons, mostly relating to capacity of the maths department. This resulted in a total of 161 teachers across 101 settings in 81 FE colleges being included for randomisation.

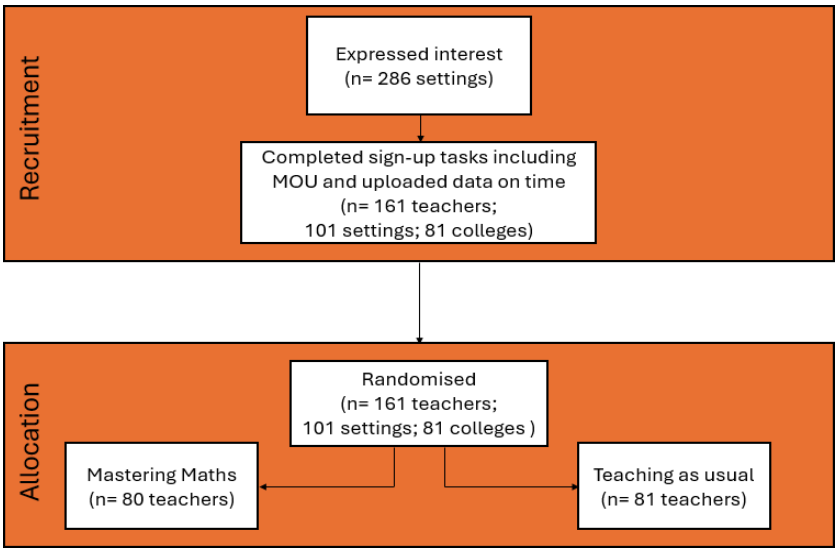
Randomisation

Randomisation was carried out blind by the Impact Evaluation team at NatCen using the ***randtreat*** command in Stata version 17 on 9th September 2024. Randomisation syntax is provided in Appendix A.

NatCen communicated the outcome of randomisation to the delivery team, who in turn notified settings and teachers.

Figure 1 presents the CONSORT diagram outlining the flow of participating teachers from the recruitment to the allocation of intervention. The diagram will be updated in the final report to reflect the flow of teachers and students from recruitment through randomisation and analysis.

Figure 1 CONSORT



[5] More information on setting and teacher eligibility criteria can be found in Mastering Maths evaluation protocol. Available at https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/mastering_maths_-_evaluation_protocol.pdf?v=1729167831

Randomisation was conducted at the teacher level and stratified by FE setting[6]. In FE settings with two teachers, one was randomly assigned to the Mastering Maths intervention group and the other to the TAU group. In FE settings with only one teacher, teachers were grouped into a single stratum and half were randomly assigned to the Mastering Maths intervention group and the other half to the TAU group.

The process of randomisation was as follows:

- Removing identifiable information and retaining only teacher ID, URN[7], and setting ID[8] as identifiers in the dataset for randomisation;
- Choosing a random number seed;
- Listing teachers in descending order by teacher ID[9];
- Using the ***randtreat*** command in Stata 17 to randomise teachers within each stratum and address misfits globally[10];
- The assignment of intervention versus TAU was then determined based on random numbers generated for each observation.

After randomisation, 80 teachers were assigned to the Mastering Maths group, with 20 teachers from one-teacher FE settings and 60 from two-teacher FE settings. Table 2 illustrates the randomisation allocation by setting's number of teachers.

Table 2 Randomisation allocation across strata

Intervention	One-teacher setting	Two-teacher setting	Total
Mastering Maths	20	60	80
Teaching-as-usual	21	60	81
Total	41	120	161

Primary outcome

The primary outcome is the GCSE Maths raw scores of students resitting the exam (RQ1).

GCSE Maths exams are administered by different exam boards in different FE colleges, at the end of the academic year, in May and June. The exams comprise multiple papers that test a range of mathematical skills, including number, algebra, ratio and proportion, geometry and measure, probability and statistics. The exams are divided into foundation and higher tiers, where the foundation tier covers grades 1-5 and the higher tier covers grades 4-9.

[6] Initially, we planned to stratify by setting and group allocation in the efficacy trial (i.e., whether in full intervention or not). Given that none of the teachers participated in the efficacy trial, we only stratified by setting.
[7] Unique Reference Number (URN) is a six-digit number assigned by the Department for Education to identify educational establishments in the UK.
[8] Setting ID is a unique ID that was created by the evaluation team to identify different settings within FE colleges.
[9] Teacher ID is a unique ID that was created by the evaluation team to identify teachers.
[10] There were 161 teachers to be randomised at this stage. Because teachers cannot be evenly distributed into treatment/control groups, we used the global method to deal with misfits. See Carril, A.
(2017). Dealing with misfits in random treatment assignment. The Stata Journal, 17(3), 652-667.

We will use GCSE Maths raw scores for the primary analysis, collected directly from FE colleges or exam boards. The scores range from 0 up to around 80 or 100 per paper, depending on the exam board or year of assessment, and are considered a continuous measure. GCSE Maths raw scores can be converted to grades on a 1-9 scale, with 9 being the highest. Grade 4 is a 'standard pass' and grade 5 is a 'strong pass'. We expect most students in our target population to sit the foundation paper only.[11] As the foundation paper goes up to a grade 5, the use of a grade scale would make it more difficult for us to distinguish between different levels of student performance, reducing the sensitivity of the measure and the power of analysis. For this reason, the grade scale will only be used for secondary analyses, where relevant.

Different exam boards use different scales, which further complicates comparison of the raw scores across the four exam boards[12]. To address this, we will first group students by exam board. Within each group, we will standardise the GCSE raw scores to have a mean of 0 and standard deviation of 1. After standardisation, we will combine these standardised scores from all exam boards to create a single primary outcome measure.

Secondary outcomes

GCSE Maths Grade

For the secondary outcomes, we will use students' GCSE Maths grades to estimate the probability of students achieving a grade 4 or higher on their GCSE Maths resit (RQ2), and the probability of students moving up at least one grade on their GCSE Maths resit (RQ3). For these outcomes, we will first convert the GCSE Maths raw scores into grades based on the relevant conversion tables provided by exam boards.

For RQ2, we will then create a binary measure which will take a value of 1 if a student achieves a grade 4 or higher, and 0 if otherwise. Similarly, for RQ3, we will create a binary measure taking a value of 1 if a student receives a higher grade than in their most recent GCSE Maths exam available via the NPD and 0 if otherwise.[13]

Defining disadvantage status

We will also use GCSE Maths raw scores to assess whether the impact of Mastering Maths differs for students from disadvantaged backgrounds, as measured by prior free school meal (FSM) status (RQ4). For students' prior FSM status, we will use "EVERFSM_6_P_[term][yy]" from the NPD, which indicates if a student has been recorded as eligible for FSM at any point in the last six years.

[11] Since the trial focuses on students resitting GCSE Maths under the maths and English condition of funding for 16-19 year olds, we anticipate most students will take the foundation paper. If students do sit the higher tier paper, we will either standardise scores within paper and exam board or exclude them from analysis, depending on the number of students in question.

[12] The exam boards are Assessment and Qualifications Alliance (AQA), Pearson Edexcel, Oxford Cambridge Recognition (OCR) and WJEC.

[13] We will also use this data to present descriptive statistics on the patterns of previous GCSE results (e.g., how many times the students attempted GCSE resit exams).

While the primary definition of students' disadvantage status is their FSM status, in our evaluation we will explore an additional definition of the disadvantage status more relevant to the 16-19 education and its funding. The funding formula for all providers delivering 16-19 education includes a disadvantage funding element based on students' economic deprivation and low prior attainment.[14] For the former block of funding, the Education and Skills Funding Agency (ESFA) increases the funding of those post-16 education providers with students living in the 27% most deprived areas. This is identified based on the Index of Multiple Deprivation 2019 (IMD). The IMD is the official measure of relative deprivation in England. It ranks areas from the most deprived (rank 1) to the least deprived[15]. IMD deciles further categorise areas on a 1-10 scale, with decile 1 indicating the most deprived areas.

We have obtained students' home postcodes from FE colleges as part of student enumeration in November 2024. Each year ESFA publishes the list of postcodes eligible for additional disadvantage funding. We will create a binary measure of the IMD status to identify students whose home postcodes fall within the 27% most deprived areas. For RQ5, we will use students' IMD status to assess whether the impact of Mastering Maths differs for students from disadvantaged backgrounds as measured by the IMD status.

Details on the analyses involving the FSM status and IMD status are provided in the Subgroup Analyses section.

Attitudes towards maths

To estimate whether Mastering Maths improves students' attitudes towards maths (RQ7-8), we will measure students' self-confidence (RQ7) and self-efficacy (RQ8) using the anglicised version of the self-confidence subscale of the Attitude towards Maths Inventory (ATMI) and the University of Manchester self-efficacy questionnaire.

The ATMI was developed by Tapia and Marsh (2004) and uses a 5-point Likert scale with responses ranging from "strongly disagree" (value=1) to "strongly agree" (value=5). The original ATMI comprises 40 items measuring four domains: self-confidence (15 items), value (10 items), enjoyment (10 items) and motivation (5 items). We will only use the self-confidence subscale to measure students' attitudes towards maths. This subscale, measuring people's confidence and anxiety toward maths, demonstrated good internal reliability (Cronbach's alpha = 0.95) and test-retest stability (r = 0.88) (Tapia and Marsh, 2004). The overall self-confidence in maths score will be calculated by summing responses. Any negative questions will be reverse scored. The overall score will take a value between 15 and 75, where higher scores indicate higher self-confidence in maths.

[14] The details of an overview of 16-19 education funding and how it is calculated can be found in here: https://www.gov.uk/guidance/16-to-19-funding-how-it-works#:~:text=We%20fund%3A,enrolled%20into%20eligible%20FE%20institutions

[15] English indices of deprivation 2019. Ministry of Housing. Available at https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

The self-efficacy questionnaire we will use was developed by academics at the University of Manchester. In this questionnaire, students are provided with 22 GSCE Maths questions. Students do not need to answer the maths questions but instead are asked to indicate how confident they feel in answering them. The questions cover topics such as dealing with numbers, fractions and ratios, interpreting graphs/coordinates/charts, and solving algebraic/quadratic equations. Each question follows a 4-point Likert scale with responses of "not confident at all" (value=1), "not very confident" (value=2), "fairly confident" (value=3) and "very confident" (value=4). The overall self-efficacy in maths score will then be calculated by summing responses. The overall score will take a value between 22 and 88, where higher scores indicate higher self-efficacy in maths.

We will collect students' self-confidence in maths and self-efficacy in maths through the endline student paper survey, which will be administered by their teacher.

Considering that validity and reliability of these measures may not hold for individual subscales, we will report Cronbach's α and McDonald's ω to assess the reliability of our measures of students' attitudes towards maths. Specifically, we will examine self-confidence (15 items) and self-efficacy (22 items) separately. These reliability coefficients will help determine how accurately our scores reflect the true construct of attitudes, taking into account that measurement errors can distort the associations among the constructs represented by the observed variables (Bland & Altman, 1997; Flora, 2020).

Cognitive testing

We agreed with EEF and UoN to conduct cognitive interviews on the self-confidence and self-efficacy scales as neither scale has previously been used with this specific student population, i.e., students in FE colleges who are retaking their GCSE Maths. We developed a cognitive testing protocol and conducted fieldwork in December 2024 and January 2025. We conducted ten cognitive interviews, each lasting up to one hour. Participants were recruited via two recruitment agencies. All the students were currently retaking their GCSE Maths, and they were aged 16-19 and based in FE colleges and sixth-form colleges in England. We screened students to make sure they fit the eligibility criteria and that they were not based in FE colleges that are part of the main Mastering Maths trial.

The main question we wanted to answer through the cognitive fieldwork was whether the questions in the self-confidence and self-efficacy scales were suitable for this student demographic. There was no evidence to suggest they were not. In the main, participants were happy with the questions and understood them clearly. They were able to choose an answer and explain how they got to it, demonstrating they understood what they were being asked to do. All participants understood that in the self-efficacy scale, they did not need to solve the maths problems but instead assess their confidence in solving the problems.

In the self-confidence scale, two words/phrases were not known to participants, though they were still able to correctly guess the meaning from the context: 'dreaded' (mentioned by one participant) and 'terrible strain' (mentioned by two participants). In the self-efficacy scale, participants raised issues about the wording and difficulty level of the 'properties of shapes' question (mentioned by seven participants) and the clarity of the 'measurement and estimation' question (mentioned by three participants). In addition, participants were not clear

on the use of the word 'congruent' (mentioned by one participant) and the meaning of the letters 'XYZ' to denote the vertices of a triangle (mentioned by four participants).

Together with the UoN, we decided to keep the scales as they were for the endline student surveys. This is because we did not find strong enough evidence to change any of the questions. While seven out of ten participants reported difficulties with the 'properties of shape' question, we were not certain whether this reflected the difficulty level of the question (i.e., that this is one of the more challenging topic areas in the CCSE Maths curriculum) or lack of clarity in the way the question was phrased. Overall, the scales were found to work well for this student population.

Baseline measure

For this evaluation, eligible students will be those aged 16-19 who will resit their GCSE Maths exam after previously scoring below a grade 4 (or not having taken the exam). Since resit students can only have KS4 GCSE Maths grades of 1, 2, or 3, there is limited variation in this measure, making it less powerful for predicting maths attainment. To address this, we will use KS2 maths attainment as the baseline for both the primary and secondary outcome measures.

For all outcome measures, we will use KS2 Maths attainment raw score ("KS2_MATMRK" variable) from the NPD, which ranges from 0 to 110, as our baseline measure. This includes attitudes towards maths. Details are covered in the Secondary outcome analysis.

Sample size calculations overview

This trial is powered to detect a Minimum Detectable Effect Size (MDES) of 0.17 standard deviations for the primary analysis of GCSE Maths scores of students resitting the exam, using the sample size at randomisation. Details on power calculations are covered in the Updated sample size calculations section. We have used PowerUp! (Dong and Maynard, 2013) to perform all of the sample size calculations.

Planned sample size

We planned to recruit 140 teachers and 7000 students resitting the GCSE maths exam. Our power calculations were informed by the Mastering Maths efficacy trial[16]. We assumed a Type I error rate of 0.05 and a Type II error rate of 0.20 (i.e., 80% power). We used estimates found in the efficacy trial for the correlation between baseline (KS2 Maths score) and endline (GCSE Maths score) attainment at the student level (0.25) and at the teacher level (0.23), and for the intra-cluster correlation (0.14)[17]. We assumed an average of 50 eligible students per teacher, and based on the efficacy trial, we assumed that 36% of students are eligible for FSM before further education. Under these assumptions, we would achieve an MDES of 0.188 standard deviations at protocol stage.

Updated sample size calculations

Table 3 and 5 present our sample size calculations for the primary outcome in the trial, covering the stages from protocol through randomisation to analysis, including some level of attrition. These calculations indicate the smallest effect size, measured in standard deviations,

[16] Wake, G., et al. 2023. Centres for Excellence in Maths Teaching for Mastery Randomised Controlled Trial, Evaluation Report. University of Nottingham

[17] In the efficacy trial, there was one teacher per setting. We therefore assumed setting-level estimations from the efficacy trial to be applicable to the teacher-level estimations for this trial.

that the trial is able to detect with 80% probability given its sample size and a set of underlying assumptions above.

Table 3 Minimum detectable effect size across protocol and randomisation stages

		Protocol		Randomisation	
		All students	FSM eligible students (36%)	All students	FSM eligible students (36%)
Minimum Detectable Effect Size (MDES)		0.188	0.206	0.175	0.196
Pre-test / post-test correlations	Level 1 (student)	0.25			
	Level 2 (teacher)	0.23			
Intra-cluster correlations (ICCs)	Level 2 (teacher)	0.14			
Alpha		0.05			
Power		0.80			
One-sided or two-sided		2			
Number of Level 2 covariates		105	105	63[18]	63
Average cluster size		50	18	36.8	13.2
Number of teachers	Mastering Maths	70	70	80	80
	Teaching-as-usual	70	70	81	81
	Total	140	140	161	161
Number of students	Mastering Maths	3500	1260	2944	1056
	Teaching-as-usual	3500	1260	2981	1070
	Total	7000	2520	5925	2126

The revised calculations ('Randomisation' columns in Table 3) are based on the number of teachers recruited by the randomisation stage (n= 161). The calculations use the same core assumptions as those in the evaluation protocol, except the assumption for the average cluster size. After the second round of enumeration, we identified high variability in cluster sizes. We

[18] Our randomisation sample consists of 60 two-teacher settings and 41 one-teacher settings. The number of Level 2 covariates includes 60 covariates for strata fixed effects (i.e., two-teacher settings and one-teacher settings, with the latter stratum serving as the reference category) and three covariates for exam board (i.e., while there are four exam boards, one will serve as the reference category).

use harmonic average cluster size in our power calculations to accommodate the variability in cluster size (Dong and Maynard, 2013).[19]

Using the revised sample size at randomisation yields a MDES of 0.175 standard deviations for GCSE Maths scores of students resitting their exam and 0.196 standard deviations for GCSE Maths scores of FSM-eligible students resitting their exam.[20]

Table 5 Minimum detectable effect size across SAP and projected analysis stages

		At the time of writing SAP		Projected Analysis With Attrition[21]: Teacher: ~15% Student: ~25%	
		All students	FSM eligible students (36%)	All students	FSM eligible students (36%)
Minimum Detectable Effect Size (MDES)		0.187	0.210	0.193	0.218
Pre-test / post-test correlations	Level 1 (student)	0.25			
	Level 2 (teacher)	0.23			
Intra-cluster correlations (ICCs)	Level 2 (teacher)	0.14			
Alpha		0.05			
Power		0.80			
One-sided or two-sided		2			
Number of Level 2 covariates		63	63	63	63
Average cluster size		36.8	13.2	32.7	11.8
Number of teachers	Mastering Maths	71	71	68	68
	Teaching-as-usual	71	71	68	68
	Total	142	142	136	136
Number of students	Mastering Maths	2613	937	2222	802

[19] After the second round of enumeration, we have approximately 10,309 students across 142 teacher clusters. The cluster size varies between 2 and 207 students, with an average cluster size of 72.6.

[20] While we include two definitions of disadvantage status in our secondary analysis, FSM eligibility is our primary definition of coming from a disadvantaged background. Hence, the sample size calculations only include scenarios for this definition.

[21] We apply the teacher-level attrition rate to the number of teachers at randomisation (161 x 0.85 = 136 teachers) and the student-level attrition rate to the number of students at randomisation (5925 x 0.75 = 4444 students). We then calculate the average cluster size by dividing the projected number of students by the projected number of teachers after accounting for attrition (4444/136 = 32.7 students per teacher).

	Teaching-as-usual	2613	937	2222	802
	Total	5226	1874	4444	1604

We also conducted sample size calculations to reflect the current number of students in the sample.[22] As shown in Table 5, the most recent sample size would yield a MDES of 0.187 for GCSE Maths scores of students resitting their exam and 0.210 standard deviations for GCSE Maths scores of FSM-eligible students resitting their exam.

Furthermore, attrition is a risk for any study and a key consideration for this trial.[23] As shown in Table 5, we assume around 136 teachers will remain in the trial at endline (representing ~15% attrition), with an average of 32.7 students per teacher (accounting for ~25% attrition at the student level).[24] Under these assumptions, we would achieve a MDES of 0.193.

The MDES of 0.193 is lower than our initial expectation of 0.206 in the evaluation protocol. This indicates that the trial is so far on track, at least meeting the MDES assumptions outlined in the evaluation protocol, even if there might be further attrition by the time of the endline data collection. Please note that these calculations are based on projected sample sizes with assumed attrition, and the actual analysis will be based on the observed attrition.

Analysis

Primary outcome analysis

The evaluation of Mastering Maths aims to estimate the impact of the programme on maths attainment, using an intention-to-treat (ITT) approach. The effectiveness trial is designed as a two-armed cluster randomised control trial with students clustered within teachers.

We will use a two-level linear regression model with students (Level 1) clustered within teachers (Level 2). The dependent variable will be students' GCSE Maths standardised raw score. The independent variables will be a binary indicator of intervention allocation, KS2 Maths attainment raw score (as a baseline measure), exam board[25] and setting (the stratification variable used for randomisation).[26]

[22] At the time of drafting the SAP, there are 142 teachers remaining in the trial (71 Mastering Maths and 71 TAU), equivalent to 10,309 students, After the delivery team notified the outcome of randomisation to settings, six Mastering Maths teachers and three TAU teacher dropped out of the trial. A further ten teachers withdrew from the trial during student enumeration (three Mastering Maths and seven TAU).

[23] At the time of drafting the SAP, 19 teachers had withdrawn from the trial before completing student enumeration. The mean cluster size will be used to estimate the expected number of students for these teachers and to calculate the student attrition rate.

[24] More information on student and teacher attrition rates can be found in the Mastering Maths evaluation protocol. Available at https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/mastering_maths_-_evaluation_protocol.pdf?v=1729167831

[25] As different exam boards use different scales in their GCSE Maths scoring, we will standardise the GCSE raw scores within each exam board. While standardisation allows having comparable scales between exam boards, it does not account for differences between the four exam boards (such as differences in paper structure and question style) delivering the GCSE exams. Therefore, we include exam board as a covariate to account for differences between the four exam boards delivering the exams.

[26] Please note that a binary variable indicating whether a student took the November resit exam was originally included to account for variations in attendance. However, due to a high level of missing data for this variable, we decided to conduct a separate sensitivity analysis instead. See the *Additional Analyses* section for further details.

The basic form of the primary model is:

$$GCSE\ Maths\ score_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 ExamBoard_j + \beta_4 Stratification'_i + u_j + e_{ij}$$

where students resitting the exam (i) are clustered within teachers (j). β_0 is an overall intercept, β_1 is a fixed gradient between the standardised post-test and pre-test scores and β_2 is the average effect of the intervention. Exam board information and the stratification variable used for randomisation will be included as fixed effects in this model. The term u_j is a teacher-level random effect and e_{ij} is the error term, both assumed to be normally distributed and uncorrelated with all the covariates included in the model. Teacher-level random effects will account for teacher-level variation in outcomes that is not explained by the fixed effects. In line with EEF statistical analysis guidance (EEF, 2022), other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 17 using the ***mixed*** command.[27]

The impact of the intervention will be expressed as a standardised effect size using Hedges' g with 95% confidence intervals. See the Effect size calculation section below for an explanation of how effect sizes will be calculated. Following EEF statistical analysis guidance (EEF, 2022), we will also present histograms of the pre- and post-test scores for students resitting the exam, along with a summary of means and standard deviations of pre- and post-test scores.

Secondary outcome analysis

Analysis for GCSE Maths grade

We will also estimate the impact of Mastering Maths on secondary outcomes using an ITT approach. We will use a two-level logistic regression model with students (Level 1) being clustered within teachers (Level 2) to estimate the effect of the Mastering Maths programme on students' probability of achieving a grade 4 or higher (RQ2), and their probability of moving up a grade (RQ3), on the GCSE Maths exam. The analytical approach for both research questions will be analogous to the primary outcome estimation.

The basic form of each model for RQ2 and RQ3 is:

$$logit(P(Outcome_{ij})) = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 ExamBoard_j + \beta_4 Stratification'_i + u_j$$

Where $P(Outcome_{ij} = 1)$ is the probability of achieving the outcome of interest. Students (i) are clustered within teachers (j). Other parameters are analogous to the primary model specification.

The analysis will be implemented in Stata 17 using the ***melogit*** command.

[27] As the baseline data will be supplied from NPD, the analysis will need to be conducted through the Office for National Statistics Secure Research Service (ONS SRS). To our best knowledge, Stata 17 is the most up to date version available in the SRS environment at the time of this SAP being written.

Given a binary outcome variable, we will follow EEF statistical guidance to report the impact of the intervention as relative risk ratios (RRRs) and transform RRRs into the Cox Index, which is comparable to Hedges' g standardised effect size, as a way to facilitate interpretability and improve comparability across studies. See the Effect size calculation section below for an explanation of how effect sizes will be transformed.

Analysis for attitudes towards maths

We will also estimate the impact of Mastering Maths on students' self-confidence (RQ7) and self-efficacy in maths (RQ8) following the primary analysis model. As in the primary analysis model, we will use KS2 maths attainment as a baseline measure rather than a baseline measure of students' attitudes towards maths. This is because the relevant surveys will only be administered at endline, to reduce burden on teachers and students. Nonetheless we believe students' prior attainment in maths is correlated with their attitudes towards the subject and can thus explain some of the variance in the model.

The basic form of the model is:

$$Outcome_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 ExamBoard_j + \beta_4 Stratification'_j + u_j + e_{ij}$$

The analysis will be implemented in Stata 17 using the ***mixed*** command.

Subgroup analyses

To estimate if the impact of Mastering Maths differs for students from disadvantaged backgrounds, we will conduct three subgroup analyses using both prior FSM status (RQ4), IMD status (RQ5), and prior attainment (RQ6).

First, we will follow the primary analysis model and include an interaction term on students' prior FSM status using the "EVERFSM_6_P_[term][yy]" variable from the NPD.

The basic form of the FSM model is:

$$GCSE\ Maths\ score_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 FSM_{ij} + \beta_4 Intervention_j FSM_{ij} + \beta_5 ExamBoard_j + \beta_6 Stratification'_j + u_j + e_{ij}$$

where students (i) are clustered within teachers (j). β_2 represents the impact of the intervention on non-FSM students and β_4 is the attainment gap between FSM students and their peers (i.e. the difference in average effect of the intervention between FSM and non-FSM students). The impact of the intervention on FSM students is $\beta_2 + \beta_4$. Other parameters are analogous to the primary model specification.

Second, we will use a binary measure of the IMD status to indicate students whose home postcodes fall within the 27% most deprived areas, using the list of eligible postcodes published each year by the ESFA. Similar to the FSM model, we will follow the primary analysis model and include an interaction term for students' IMD status.

The basic form of the IMD status model is:

$$GCSE\ Maths\ score_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 IMD_{ij} + \beta_4 Intervention_j IMD_{ij} + \beta_5 ExamBoard_j + \beta_6 Stratification'_j + u_j + e_{ij}$$

where students (i) are clustered within teachers (j). β_2 represents the impact of the intervention on students from less deprived areas and β_4 is the attainment gap between students from the most deprived areas and their peers (i.e., the difference in the average effect of the intervention between students from the most and less deprived areas). The impact of the intervention on students from the most deprived areas is $\beta_2 + \beta_4$. Other parameters are analogous to the primary model specification.

Finally, we will conduct a subgroup analysis to examine whether the programme impact is smaller for students with lower prior attainment. This additional analysis was agreed after publication of the evaluation protocol. The Mastering Maths programme was designed for students who previously achieved below a grade 4 in GCSE Maths and is expected to benefit all students regardless of their prior attainment. We are interested in exploring whether this is the case, or whether there is a differential impact for lower- or higher-attaining students. Consistent with the other subgroup analyses, we will use the primary analysis model and include an interaction term for students' low prior attainment status, defined as a grade 2 or below.

The basic form of the prior attainment model is:

$$GCSE\ Maths\ score_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 PriorAttainment_{ij} + \beta_4 Intervention_j PriorAttainment_{ij} + \beta_5 ExamBoard_j + \beta_6 Stratification'_j + u_j + e_{ij}$$

where students (i) are clustered within teachers (j). β_2 represents the impact of the intervention on students with higher prior attainment and β_4 is the attainment gap between students with lower prior attainment and their peers (i.e., the difference in the average effect of the intervention between students with lower and higher prior attainment). The impact of the intervention on students with lower prior attainment is $\beta_2 + \beta_4$. Other parameters are analogous to the primary model specification.

The effect of the intervention on attainment gap for all three models will be estimated. See the Effect size calculation section below for an explanation of how effect sizes will be calculated.

In line with EEF statistical analysis guidance (EEF, 2022), we will also run the primary outcome analysis model on the restricted sub-group of students eligible for FSM, students from the most deprived areas, and students with lower prior attainment. In this case, the β_2 coefficient will provide the estimated treatment effect specifically for that respective sub-group. The results from the interaction term and sub-group models will be compared as a sensitivity check to assess whether the results are robust to different model specifications. Please see the effect size calculation for continuous outcome measures section below for an explanation of how effect sizes for sub-group analyses will be calculated.

Compliance analysis

Compliance is defined as the fulfilment of a set of minimum criteria which determine whether a teacher has participated in the programme as intended. To be deemed compliant, a teacher must fulfil all three activities outlined in Table 4. Compliance will thus be a binary measure, indicating whether a teacher was compliant or not. While we define the compliance measure at the teacher level, we will conduct the compliance analysis at the student level. Data for our compliance analyses will be collected during implementation by the Mastering Maths delivery team at the UoN. Details on compliance are covered in the evaluation protocol.

Table 4 Compliance criteria table

COMPLIANCE CRITERION	DATA SOURCE	COMPLIANCE INDICATOR
Attendance in professional development	Attendance register completed by lead teacher	Teacher attends both days of professional development (this can include a mop-up day if they miss their original session)
Attendance at lesson study sessions	Attendance register completed by lead teacher	Teacher attends at least four out of five lesson study sessions
Teaching of Mastering Maths lessons	Compliance register completed by lead teacher	Teacher teaches five Mastering Maths lessons[28]

We will estimate the Complier Average Causal Effect (CACE) using a two-stage least square (2SLS) model with the intervention allocation as the instrumental variable (IV) to recover the intervention effect for compliers[29]. This model will be estimated for the primary outcome only.

The first step will estimate the compliance rate, i.e., whether being assigned to Mastering Maths influences whether teachers comply. We will do this by regressing compliance on all covariates that are used in the primary outcome model and in addition, will include, as an IV, a binary variable that indicates a student's intervention allocation. The first stage equation estimate is as follows:

$$Comply_j = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 ExamBoard_j + \beta_4 Stratification_i' + e_{ij}$$

The second step will substitute the intervention indicator with the estimated compliance rate from the first step to predict the outcome. The estimation of the second stage equation is as follows:

$$GCSE\ Maths_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Com\hat{}ply_j + \beta_3 ExamBoard_j + \beta_4 Stratification_i' + e_{ij}$$

[28] Teachers are provided with 12 exemplar Mastering Maths lesson with research questions and student tasks, but they are expected to deliver at least 5 Mastering Maths lessons to their students.

[29] Imbens and Angrist, 1994. https://business.baylor.edu/scott_cunningham/teaching/imbens--angrist---late-1994.pdf.

The coefficient β_2 in the second stage equation is the estimate of the CACE, which will answer the following research question: 'To what extent does compliance with the Mastering Maths programme lead to improved GCSE Maths outcomes?'. In the event that there are no confounding factors affecting compliance and attainment the CACE estimate will be equal to the ITT estimate.

IV regression will be conducted in Stata 17, using the ***ivregress*** command and the **cluster** option to control for clustering of students within teachers.

Additional analyses

Sensitivity analysis

As discussed in the evaluation protocol (Takala et al., 2024), students who resit the GCSE Maths exam in November may have lower attendance at lessons. Given those who receive a grade 3 will again resit the exam in the summer (and thus be included in the primary analysis), their lower attendance may potentially weaken the intervention's effect. Ideally, we would explore this using data on student's attendance at lessons, but this data is impractical to collect. Instead, as part of the second round of student enumeration, we asked teachers to identify students who attempted the November 2024 GCSE Maths resit exam. We will use this information as a proxy for attendance, assuming students who attempted the November exam had lower attendance at lessons between November 2024 and January 2025.

Sensitivity analyses will be conducted to assess how November resit exams may affect the results. First, we will replicate the primary outcome analysis by adding a binary variable indicating whether a student took the November resit exam.[30] This information will be included to account for variations in attendance between students who do not attempt the November resit (whose attendance is unlikely to change) and those who do (and may stop attending lessons because they think they have achieved a grade 4).[31] Furthermore, we will assess whether excluding November resit students – who are more likely to disengage – affects the results. We will do this by replicating the primary and CACE analysis without including the November resit students.

If these sensitivity analyses yield similar results to those of the primary and CACE analyses (which include both students who do not resit their exam in November and those who do but receive a grade 3 or below), it may suggest that the intervention's effect is consistent and not significantly influenced by the November resit exam.[32]. Conversely, if these sensitivity analyses yield different results from the primary and CACE analyses, this may indicate that student attendance in lessons could influence the intervention's effect, suggesting that future trials should account for attendance when considering the intervention's overall impact.

[30] This approach ensures that the sample includes students who do not resit in November as well as those who resit but do not receive a grade 4 and subsequently enrol in the Summer resit. By capturing these groups, the study reflects real world scenarios and maintains a representative sample appropriate for an effectiveness trial.

[31] A binary variable indicating whether a student took the November resit exam was originally included to the primary analysis to account for variations in attendance. However, due to a high level of missing data for this variable, we decided to conduct a separate sensitivity analysis instead.

[32] This may indicate that the intervention's effect is not substantially influenced by changes in students' lesson attendance, or that our assumption regarding altered attendance behaviour following the November resit does not hold. It would not be possible to determine which of these explanations is correct without further analysis.

Finally, in line with the EEF's statistical analysis guidance (EEF, 2022), we will also estimate the primary outcome model including only prior attainment and the stratification variables. This analysis will be undertaken solely to support the EEF's synthesis work, and the results will be presented in an appendix. They will not be compared with those from the primary outcome analysis.

Mediation analysis

Mediation analysis is used to explore mechanisms by which an intervention affects the outcome of interest. For Mastering Maths, we propose that the intervention affects the GCSE resit outcomes by improving students' self-confidence and self-efficacy in maths. This mechanism is captured in the programme logic model, further details of which can be found in the evaluation protocol.

We will conduct an exploratory mediation analysis to decompose the ITT estimate into an indirect effect (i.e., effect of the intervention that can be attributed to changes in attitudes towards maths) and a direct effect (i.e., effect of the intervention that cannot be attributed to changes in attitudes towards maths).

The exploratory mediation analysis will enable us to understand whether the effect of Mastering Maths on the primary outcome is partially or totally mediated by changes in students' attitudes towards maths (i.e., self-confidence and self-efficacy in maths). We hypothesise that the effect will be at least partially mediated, but we do not have an expectation of the magnitude of this effect.

As detailed by Imai and Yamamoto (2013), standard causal mediation relies on the *ignorability assumption*, which requires no post-treatment confounding between the mediator and the outcome, whether observed or unobserved. Simply put, standard causal mediation analysis assumes causal independence between multiple mediators when present. Therefore, if multiple mediators are causally independent, causal mediation analysis can proceed as normal, estimating two sets of mediation effects separately for each mediator.

However, in the case of Mastering Maths, the two mediators of interest—self-confidence and self-efficacy—are likely to have a bi-directional causal relationship. Self-efficacy on specific maths questions can affect general confidence in maths, while general confidence in maths can improve self-efficacy on particular maths questions.

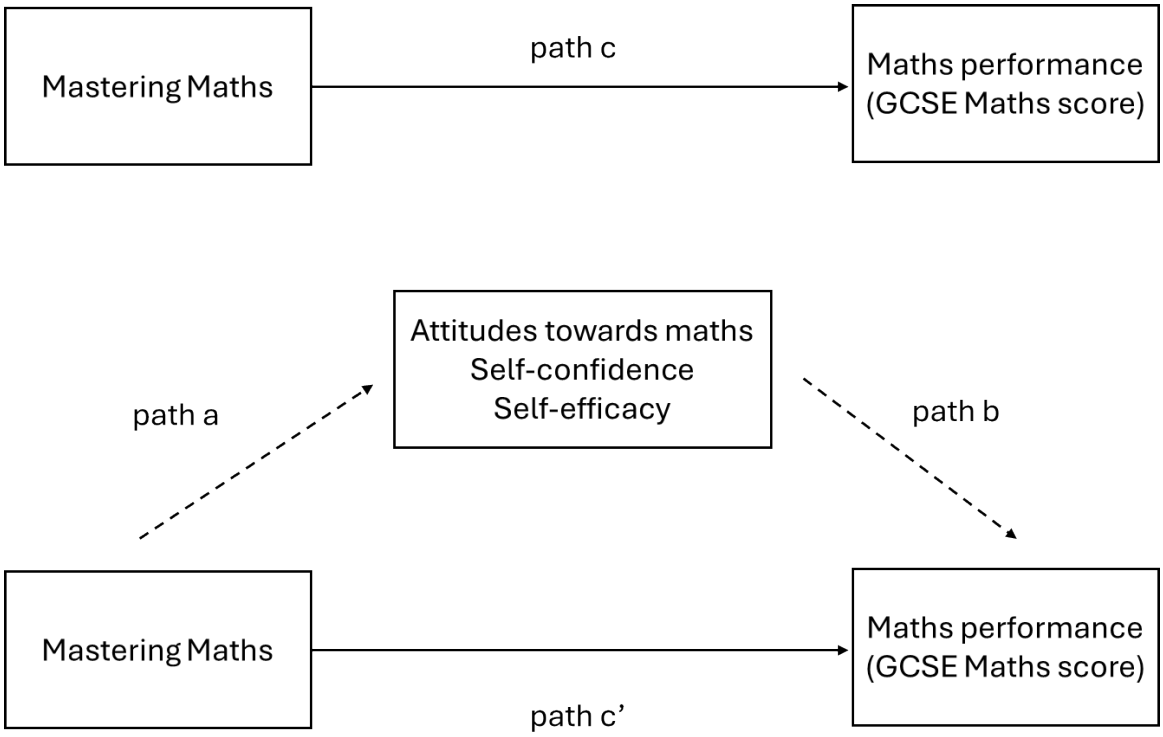
Although these outcomes are treated separately for secondary outcome analysis (as they come from different measures), they can be defined as either two closely related constructs or two approaches to measuring one common underlying construct of attitudes towards maths.

For mediation analysis, we propose to define self-confidence scores and self-efficacy scores as a common mediator (i.e., attitudes towards maths). This will be done by summing z-score standardised self-confidence and self-efficacy in maths scores, given the different scales used for the two measures.

The assumed causal model for this analysis is shown in Figure 2. The mediation analysis will follow the causal steps approach (Baron & Kenny, 1986), involving the steps below:

1. Regressing students' attitudes towards maths score (the mediator) on allocation to the Mastering Maths programme. The effect of the programme on the mediator is conventionally referred to as path a, as shown in **Error! Reference source not found.**.
2. Regressing students' GCSE Maths scores on the Mastering Maths programme and on the mediator. The effect of the mediator on the outcome is conventionally referred to as path b, while the average direct effect (ADE) of the programme (with the mediator accounted for) is referred to as path c'.[33]
3. Estimating the average causal mediated effect (ACME) and the proportion mediated (i.e., the magnitude of the ACME effect relative to the total effect).

Figure 2 Casual Mediation Model



[33] The total effect of the programme on the outcome (direct and indirect) is referred to as path c.

To estimate path a in step one of the causal mediation analysis, we will use a two-level linear mixed effects regression model, predicting student's attitudes towards maths from programme allocation, with baseline measures, exam boards, and randomisation strata as covariates.[34] The model will include a teacher-level random intercept.

$$Mediator_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 ExamBoard_j + \beta_4 Stratification'_i + u_j + e_{ij}$$

The coefficient β_2 represents the difference in attitudes towards maths between students who were allocated to the Mastering Maths programme and those who were not, i.e., the effect of the programme on self-confidence and self-efficacy.

As step two of the causal mediation analysis, we will estimate a two-level linear mixed effects regression model, predicting GCSE Maths raw scores from programme allocation and attitudes towards maths, with baseline measures, November resit, exam boards, and randomisation strata included as covariates. The model will include a teacher-level random intercept.

$$GCSE\ Maths\ score_{ij} = \alpha_0 + \alpha_1 Baseline_{ij} + \alpha_2 Intervention_j + \alpha_3 NovemberResit_{ij} + \alpha_4 Stratification'_i + \alpha_5 Mediator_{ij} + u'_j + e'_{ij}$$

The coefficient α_2 provides an estimate of path c' or the ADE of Mastering Maths on students' performance in GCSE Maths exams, while α_5 represents the effect of the mediator on students' GCSE maths performance or path b.

Drawing on the two models, $\beta_2\alpha_6$ provides the ACME.

$$Proportion\ mediated = \frac{ACME}{Total\ effect} = \frac{\beta_2\alpha_5}{\beta_2\alpha_{5\,+\,\alpha_2}}$$

The ACME, ADE and the proportion mediated effect will be estimated using the **mediation** package in R (Imai et al., 2010).

For all steps, we will present unstandardised model coefficients, p values and 95% confidence intervals obtained using bootstrapping (or quasi-Bayesian estimation whichever is more appropriate) with 1,000 simulations. The primary effect size interpreted will be the proportion mediated effect and its confidence interval.

[34] This is an identical statistical model for RQs 6 and 7, except for the fact that the dependent variable is the aggregate 'attitudes towards maths' variable.

Imbalance at baseline

To check for and monitor imbalance at baseline, we will undertake descriptive analysis at student level.[35]

Specifically, we will assess imbalance between intervention and TAU group on the following characteristics:

- Ever received FSM in the past six years
- Whether the home postcode of a student fall within the 27% most deprived areas
- Quartile distribution of the IMD rank of students' home postcode
- KS2 maths attainment
- Previous GCSE grade
- Whether a student attempted the November 2024 GCSE Maths resit exam
- Exam board

Categorical variables will be explored by conducting cross-tabulations, including counts and percentages in each category. Continuous variables will be summarised with descriptive statistics (n, mean, standard deviation and effect sizes) by group allocation. We will report standardised mean differences in baseline characteristics as Hedges' g effect sizes. An effect size of greater than 0.05 will be considered as an indication of possible imbalance. Note that the analyses will be performed through ONS SRS workspace, the outputs will thus have to follow SRS rules on statistical disclosure control (SDC).[36] The SDC will apply to all outputs, irrespective of their origin.

If imbalances are indicated, a sensitivity analysis will be estimated for the primary outcome. This model will include the unbalanced variables (i.e., where Hedges' g is greater than 0.05) in addition to those in the main model.

Missing data

As a first step, we will explore the extent of missing data on the outcome and baseline covariates descriptively, with cross-tabulations, including counts and percentages in each category.

To assess whether missingness is systematic, we will estimate multilevel logistic regression models. The outcome will be binary, reflecting whether (a) the primary outcome variable and (b) any covariates from the primary analysis (i.e. baseline attainment and exam board) are missing for each individual at follow-up. These models will include all covariates outlined in the Imbalance at baseline section, in addition to a random effect for teachers to account for clustering. Missing data for these covariates will be coded up as separate binary variables in the model. Models will be estimated using the glmer() function from the **lme4** package in R.

[35] We also originally planned to assess imbalance at baseline at the teacher level. However, the DfE does not allow student and teacher data (even anonymised or pseudonymised) to be accessed in the same SRS project area. Therefore, we will not be able to assess imbalance at baseline at the teacher level.

[36] More details on statistical disclosure control can be found from this link: https://assets.publishing.service.gov.uk/media/660d8798758315001a4a49d2/DfE_ONS_statistical_disclosure_control_policy.pdf

We will follow the protocol for missing data suggested by EEF statistical analysis guidance (EEF, 2022). For less than 5% missingness overall from randomisation to final analysis, a complete case analysis will be employed. For more than 5% missing primary data overall from randomisation to final analysis, our approach will depend on pattern of missingness:

- MCAR (missing completely at random): If the pattern of missingness is not correlated with either observable and/or unobservable variables, then missing data will be assumed missing completely at random (MCAR) and we will continue with a complete case analysis.
- MAR (missing at random):
    - If **only** the primary outcome variable is missing in a way that is related to observed variables, we will assume MAR and conduct a sensitivity analysis by replicating the primary model and including those observed predictors of missingness to aid interpretation.
    - If covariates are missing in a way that is correlated with observable variables, we will assume MAR and use Multiple Imputation (MI) by Chained Equations (MICE) to impute missing values. The MI model will include all variables used in the missingness models to satisfy the congeniality assumption (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren & Oudshoorn, 2000; Woods et al., 2024). Depending on the extent of missingness, we may impute both outcome and covariates to retain sample size, while acknowledging the potential risks of bias associated with imputing outcomes.

      Multiple imputation will be conducted using the **mice** package in R, accounting for the multilevel structure of the data. The minimum number of imputed datasets will depend on the fraction of missing information, as suggested by Graham et. al., (2007). The imputed datasets will be used to replicate the primary analysis, and we will compare the results with the complete case analysis as part of sensitivity analyses.

      To further address the congeniality assumption and mitigate potential bias introduced by imputing the outcome variable, we will conduct two complementary analyses using the imputed datasets:

      1. We will replicate the primary analysis model, which includes only the core variables (i.e., baseline attainment, intervention indicator, exam board, and stratification variables), and compare the results from the imputed datasets with those from the complete case analysis.

      2. We will estimate an expanded analysis model that includes all variables used in the imputation process. This ensures that the analysis model is congenial with the imputation model. Results from both models will be reported and compared to assess the robustness of findings under different assumptions and model specifications.

- MNAR (Missing Not at Random): If the pattern of missingness depends on an unobserved variable, even after considering all the information in the observed variables, we will consider the missing observations are missing not at random (MNAR) and follow EEF statistical analysis guidance (EEF, 2022) to carry out a weighting approach after MI, as suggested by Carpenter et al. (2007).

Note that missing data analysis will only be possible in cases where we have data from the NPD (i.e. baseline attainment).

Intra-cluster correlations (ICCs)

The intra-cluster correlations (ICCs) will be estimated directly from the primary analysis model, using the variance estimates for each level of clustering. The ICC for teachers $\boldsymbol{\rho S}$ will be estimated with the post-estimation command ***estat icc*** in Stata 17, using the following formula based on Hedges (2011):

$$\rho_S = \frac{\sigma^2_{BS}}{\sigma^2_{BS} + \sigma^2_{WS}} = \frac{\sigma^2_{BS}}{\sigma^2_{WT}}$$

where σ^2_{BS} is the between-teacher variance, σ^2_{WS} is the within-teacher variance and σ^2_{WT} is the total variance.

Effect size calculation

Effects size calculation for continuous outcome analyses (e.g., GCSE scores)

For primary and secondary outcome analyses and sensitivity analyses involving continuous outcome measures, we will use the effect sizes for cluster-randomised trials, as adapted from Hedges (2007):

$$ES = \frac{(\bar{Y}_T - \bar{\bar{Y}})_{C\ adjusted}}{\sqrt{\sigma^2_u + \sigma^2_e}}$$

Where $(\bar{Y}_T - \bar{\bar{Y}})_{C\ adjusted}$ is the mean difference between the intervention and TAU group

adjusted for baseline characteristics, while $\sqrt{\sigma^2_u + \sigma^2_e}$ is an estimate of the population standard deviation. σ^2_u is the variance of school level intercept and σ^2_e is variance of residuals.

From these models involving a continuous outcome measure, we will take each group's adjusted mean and variance to calculate the effect size. The variance will be the total variance (across both students and teachers, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). A 95% CI for the effect sizes, that takes into account the clustering, will also be reported.

From these models involving a continuous outcome measure, we will take each group's adjusted mean and variance to calculate the effect size. The variance will be the total variance (across both students and teachers, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). A 95% CI for the effect sizes, that takes into account the clustering, will also be reported.

Effect size calculation for binary outcome analyses (e.g., GCSE Maths grade)

For secondary outcome analyses involving binary outcome measures, we will report effect sizes as relative risk ratios (RRRs):

$$RRR = \frac{P(GCSE\ Maths\ exam\ grade \mid Intervention, X)}{P(GCSE\ Maths\ exam\ grade \mid Control, X)}$$

Where the numerator is the probability of achieving/moving up a specific GCSE Maths exam grade for the intervention group conditional on covariates (denoted X in the formula), and the denominator is the probability of achieving/moving up a specific GCSE Maths exam grade for the TAU group conditional on the same set of covariates.

We will calculate the conditional probabilities from the fitted coefficients of the multilevel logistic regression models by holding the covariates constant at their means. We will then calculate relative risk ratios using the ***nlcom*** command in Stata, which returns the standard errors and confidence intervals of each ratio.

Following the EEF statistical guidance, the effect size measure can be transformed into that comparable to Hedges' g by using the Cox Index as presented below (EEF, 2022, p.8; What Works Clearinghouse, 2017, p13).

$$d_{Cox} = \frac{[\ln(\frac{P_t}{1-P_t}) - \ln(\frac{P_c}{1-P_c})]}{1.65}$$

Where P_t is the probability of achieving/moving up a specific GCSE Maths exam grade in the intervention group and P_c the probability of achieving/moving up a specific GCSE Maths exam grade in the TAU group.

Effects size calculation for subgroup analysis using an interaction term

Following EEF statistical analysis guidance (EEF, 2022), we will calculate the effect sizes for subgroup analyses using an interaction term using the following equation:

$$ES_{subgroup} = \frac{\beta_2 Intervention_j + \beta_4 Intervention_j Subgroup_{ij}}{sd_{subgroup}}$$

Where β_2 and β_4 correspond to the specifications outlined in the *Subgroup analyses* section above, and $sd_{subgroup}$ is unconditional standard deviation of the relevant student subsample. A 95% CI for the effect sizes, that takes into account the clustering, will also be reported.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj*, *314*(7080), 572.

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

Carril, A. (2017). *Dealing with misfits in random treatment assignment*. The Stata Journal, 17(3), 652-667.

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. Journal of Research on Educational Effectiveness, 6(1), 24-67.

Education Endowment Foundation [EEF]. (2022) *Statistical analysis guidance for EEF evaluations*. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1696581237

Flora (2020) Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. *Advances in Methods and Practices in Psychological Science*. *2020; 3(4):484-501*. Doi:10.1177/2515245920951747

GOV.UK (2024). *16 to 19 funding: how it works*. https://www.gov.uk/guidance/16-to-19-funding-how-it-works#:~:text=We%20fund%3A,enrolled%20into%20eligible%20FE%20institutions

GOV.UK (2019). *English indices of deprivation 2019*. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*(1), 549-576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (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.

Imai K, & Yamamoto T (2013). Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments. *Political Analysis, 21*(2), 141–171. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=32fa962ebd7580dc5d6a08688d02b9a532403703

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25, 51–71.

Imbens, G., Angrist, J. (1994) *Identification and Estimation of Local Average Treatment Effects*. https://business.baylor.edu/scott_cunningham/teaching/imbens--angrist --- late-1994.pdf

Luedtke, O., Robitzsch, A., & Grund, S. (2017). Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies. *Psychological Methods*, 22(1), 141–165. https://doi.org/10.1037/met0000096

Office for National Statistics (2024). *Statistical Disclosure Control Policy for DfE data*. https://assets.publishing.service.gov.uk/media/660d8798758315001a4a49d2/DfE_ONS_statistical_disclosure_control_policy.pdf

Takala, H., Duysak, E., Rennick, A., Damodaran, A., Kuo, T-L., Tomlinson, C. (2024). *Independent evaluation of the Mastering Maths programme: a two-arm cluster randomised trial Evaluation Protocol*. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/mastering_maths_-_evaluation_protocol.pdf?v=1729167831

Tapia, M., & Marsh, G. E. II (2004). An Instrument to Measure Mathematics Attitudes. *Academic Exchange Quarterly*, 8, 16-21.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate imputation by chained equations. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.1177/0962280206074463

van Buuren, S., & Oudshoorn, C. G. M. (2000). Multivariate imputation by chained equations: MICE V1.0 users's manual. TNO Prevention and Health, Public Health.

Wake, G., Adkins, M., Dalby, D., Hall, J., Joubert, M., Lee, G., & Noyes, A. (2023). *Centres for Excellence in Maths Teaching for Mastery Randomised Controlled Trial Evaluation Report*. https://www.nottingham.ac.uk/research/groups/crme/documents/cfem-tfm-report.pdf

What Works Clearinghouse (n.d). *Procedures and Standards Handbook*, Version 4.0, https://ies.ed.gov/ncee/wwc/docs/referenceresources/wwc_procedures_handbook_v4.pdf

Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (2024). Best Practices for Addressing Missing Data Through Multiple Imputation. Infant and Child Development, 33(1), e2407. https://doi.org/10.1002/icd.2407

Appendix A: Randomisation syntax

```
*********************************************************************
***  Mastering Maths - Randomisation ***
*********************************************************************
* 09/09/2024

clear all
capture log close
set more off

*********************************************************************
*********************************************************************

* Import the randomisation data-blinded

use "[datapath]\MM_randomisation_data.dta", clear

*****
* Set up a log-file
*****

log using ""[datapath]\Randomisation for MM.smcl", replace

****
* Setting the Stata version and seed number for replicating the results
* Set the Stata version for replicating the results
version 17.0
```

```
* First choose a seed number

set seed 5939572


sort teacher_id


*ssc install  randtreat


***RANDOMISATION FOR Mastering Maths ***

randtreat, generate(treatment) replace strata(campuscode) misfits(global) setseed(5939572)


* Label treatment variable

lab def treat 0 "Control" 1 "Treatment", replace lab

val treatment treat


* Saving the output data-blinded

sort teacher_id

save ""[datapath]\MM_allocation_blinded_data.dta", replace

*******************************************************************************

*******************************************************************************

*******************************************************************************

* Merging back with the original data and save in dta and excel formats

sort teacher_id

merge 1:1 teacher_id urn campuscode using ""[datapath]\MM_recruitment_data.dta"

* Exporting excel and dta datasets


export excel using ""[datapath]\MM_teacher_allocation.xlsx", firstrow(variables) replace save

""[datapath]\MM_teacher_allocation.dta", replace


log close
```

```
exit
```