Maths-for-Life

Evaluation report

February 2026

Patrick Taylor and Neus Torres Blas

The Education Endowment Foundation (EEF) is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, colleges, and early years settings to improve teaching and learning for 2– 9-year-olds through better use of evidence.

We do this by:

- **Summarising evidence**. Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds. Putting evidence to use.
- **Putting evidence to use**. Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were reendowed with an additional £137m from government, allowing us to continue our work until at least 2032.

This report was supported by JPMorganChase. The views expressed in this report should not be taken to reflect the official position of JPMorganChase or any of its affiliates.

For more information about the EEF or this report please contact:

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

0207 802 1653

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Table of contents

About the evaluator

This project was independently evaluated by a team at the Behavioural Insights Team. The project was led by David Nolan and Patrick Taylor. The impact evaluation was designed by David Nolan, Pantelis Solomon, and Michael Sanders. Analysis was conducted by David Nolan, Claire Cathro, and Neus Torres-Blas. Impact data collection was managed by David Nolan, Louise Jones, and Bridie Murphy. The implementation and process evaluation (IPE) was designed by Patrick Taylor and Jessica Heal. IPE field data collection was conducted by Patrick Taylor.

Contact details:

The Behavioural Insights Team
58 Victoria Embankment,
London,
EC4Y 0DS

Email: info@bi.team

Acknowledgements

Centre for Research in Mathematics Education at the University of Nottingham: Geoffrey Wake; Michael Adkins; Matt Woodford; and Sheila Evans.

The Education Endowment Foundation: Florentina Taylor; Kathryn Davies; Thomas Mackay; Fabiola Clemente; Guillermo Romero; and Daniela Alvarado.

Disclaimer

This work was undertaken in the Office for National Statistics (ONS) Secure Research Service using data from ONS and other owners and does not imply the endorsement of the ONS or other data owners.

Executive summary

The project

The Maths-for-Life programme aims to improve GCSE Maths retake outcomes for post-16 learners in further education colleges, sixth-form colleges, schools, and training providers. It uses a learner-centred classroom approach based on problem-solving and dialogic teaching in mathematics classes. Lead teachers receive six days of training consisting of: a launch; five planning/reflection days; and a closing event. Lead teachers support peers through five cycles of Lesson Study, which focus on evidence informed, discussion-based approaches to teaching GCSE Maths resit learners. The programme was created and is delivered by the Centre for Research in Mathematics Education at the University of Nottingham.

This was a two-armed randomised controlled efficacy trial, with randomisation at the setting level. A total of 100 settings were randomly assigned to receive either the Maths-for-Life programme or business as usual. The implementation and process evaluation involved observations of the Lesson Study training, observations of the lessons, and interviews with teachers. The trial took place between October 2018 to April 2025, with reporting delayed due to legal issues around data security with the National Pupil Database (NPD).

This trial was funded by the Education Endowment Foundation (EEF) as part of a joint initiative with J.P. Morgan to explore how to improve outcomes for disadvantaged 16- to 18-year-old students retaking GCSE English or Maths.

Table 1: Key conclusions

Key conclusions	
1.	Learners in Maths-for-Life settings made two months' less progress in GCSE Maths scores, on average, compared to learners in other settings. This result has a low security rating.
2.	Among learners previously eligible for free school meals (FSM), those in Maths-for-Life schools made one month's less progress in GCSE Maths scores, on average, compared to those in other settings. These results may have lower security than the overall findings because of the smaller number of learners.
3.	There is no evidence that Maths-for-Life had an impact, either positive or negative on GCSE Maths pass rate. The result is uncertain due to high attrition.
4.	Learners in the intervention group were more likely to attend their GCSE Maths exam—a key improvement, given that attendance was low across both groups. Given the importance of exam attendance for achieving qualifications, this finding is noteworthy.
5.	The intervention was delivered broadly as planned, though some teachers made changes. Some classroom activities were adapted by teachers, which may have affected how closely the approach matched the original design.

EEF security rating

These findings have a low security rating. This was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a number of settings. The trial was a well-designed, two-armed, randomised control efficacy trial that was well powered, however, 48% of learners who started the trial were not included in the final analysis. Although learners in the Maths-for-Life group and the comparison settings had similar amounts of missing data and additional analyses showed comparable outcomes, the high volume of missing learner data may have biased the results in unpredictable ways. Therefore, the primary findings need to be interpreted with caution.

Additional findings

Learners in the Maths-for-Life settings made, on average, two months' less progress in their GCSE Maths scores compared to those in the control group, with no clear difference in pass rates between the two groups. This is our best estimate of impact, which has a low security rating. It is worth noting that GCSE scores are more sensitive to changes than GCSE pass rates.

Teacher surveys suggested that while lead teachers generally delivered the Lesson Study component of the programme as intended, attendance at professional development (PD) planning sessions and observations was lower than expected. Class teachers mostly followed the lesson plans, but the quality of dialogic teaching and problem-solving varied. Teachers found it challenging to consistently facilitate high-quality dialogue and deepen their own mathematical reasoning.

Several implementation challenges may have influenced the evidence of impact on attainment. These included exposure to Lesson Study and PD for class teachers, varying levels of buy-in and preparedness, and inconsistent learner engagement, particularly when learners resisted collaborative approaches. In settings where the programme appeared more successful, key enabling factors included well-prepared teachers, strong buy-in, positive teacher–student relationships, and supportive group dynamics for both PD and classroom delivery.

Despite limited impact of evidence on attainment or self-efficacy, a notable result is the 6% increase in GCSE Maths exam attendance among Maths-for-Life students (27.8% vs 21.6%),[1] a key improvement. This suggests a positive influence on confidence and persistence, an area worth further investigation.

Cost

The average cost of implementing Maths-for-Life was £33.63 per student per year when averaged over three years. Additional running costs are minor and relate only to printing and photocopying materials. The average teacher spent 10.2 teaching days (61 hours) supporting the intervention per year, including for preparation and delivery.

Impact

Table 2: Summary of impact on primary outcome(s)

Outcome / group	Effect size (95% confidence interval)	Estimated months' progress	The EEF security rating	No. of students	P-value	EEF cost rating
GCSE Maths attainment Uniform Mark Scale (UMS) score	-0.10[2] (-0.20 – 0.01)	-2	🔒 🔒 🔒 🔒 🔒	1,631	0.09	£ £ £ £ £
GCSE Maths attainment for students eligible for FSM	-0.07 (-0.23 – 0.08)	-1	N/A	1,017	0.34	N/A

N/A=not applicable.

[1] These figures show the proportion of randomised students (i.e. those enrolled in GCSE Maths resit classes in trial colleges at the start of the 2018/2019 academic year) who sat the July 2019 exam. Some of these students dropped out of these classes before being entered into the July exam. A total of 350 students (6% of the randomised sample) dropped out because they passed their November 2018 resit.
[2] The effect size is -0.095. This value is rounded to -0.10 in the executive summary and for conversion purposes. All impact results tables and related analyses retain and reference the unrounded value (-0.095). This rounding does not affect the underlying analyses.

Introduction

Background

In 2017, the year before this trial was launched, around 30% of young people failed to attain a grade 4 in GCSE Maths during Year 11 (Murray, 2017). A significant proportion of these young people retake this exam in further education or sixth-form colleges, with the remainder retaking in school sixth forms. However, in 2017, only 28% of those resitting their GCSE Maths obtained a grade 4—the equivalent of a standard pass, indicating a minimum level of competence in the subject (Thompson, 2017). This year, in 2024, this has fallen to 17% (Camden, 2024). Mathematical literacy is a critical life skill. Despite its economic and social importance, numeracy in the United Kingdom (UK) is weak. According to the Organisation for Economic Co-operation and Development (OECD), before the trial launched, one-third of 16–19-year-olds in the UK had poor mathematical skills, three times as many as the highest performing countries (Kuczera *et al.,* 2016). For the individual, achieving a pass in GCSE Maths opens doors to better longer term educational and employment outcomes (Hayward *et al.,* 2014).

The evidence for the principles behind the programme evaluated in this trial is best described in a number of publications by Malcolm Swan, the lead designer and researcher of the materials on which the intervention is based (Swan, 2006). Fundamentally, the teaching resources draw on design principles from diagnostic teaching research. Swan (2006) reports evidence that effective use of the materials in student-centred ways in post-16 contexts leads to increases in attainment. One study used a pre-/post-test design (N=334) to assess the outcomes of students who received 'many' or 'few' of the lessons (Swan, 2006). Those who received 'many' had statistically significant gains (at the 5% level) on an algebra test compared to their peers who received 'few'. In another study (Herman *et al.,* 2015), materials were adapted for the United States (US) (and for students of all abilities, rather than the equivalent of GCSE resit students) and were evaluated using a quasi-experimental design (N=471) with a matched control group. This later study found that the intervention group made significant gains in attainment compared to the control group (0.13 Cohen's d effect size) (Herman *et al.,* 2015).

A pilot evaluation was carried out in advance of this trial. In the 2017/2018 academic year, the University of Nottingham ran a pilot version of the Maths-for-Life programme with 20 teachers from 20 colleges, with the Behavioural Insights Team (BIT) acting as the independent evaluator. This pilot study aimed to evaluate the promise, feasibility, and readiness for the trial of the programme, using a combination of qualitative case studies of pilot colleges, observations of training days, quantitative surveying, and interviews with four non-pilot colleges. The findings suggested that this was an ambitious project, due to the very low levels of student confidence, the difficult-to-master pedagogy, and the likely dilution of quality in delivery as the programme scales. However, we also found evidence of promise in terms of student outcomes, and confidence among pilot colleges that the programme is feasible to deliver and ready for trial. The details of these findings influenced the refinement of the intervention (e.g. considering the role of teaching assistants), its logic model, and the focus of the implementation and process evaluation (IPE; e.g. paying particular attention to the variation in competency of the class teachers and lead teachers as the programme is scaled up).

This current study was a two-arm randomised controlled efficacy trial, with randomisation at the setting level, stratified by setting type (further education college, sixth-form college, school, and training provider). The two arms were: i) the treatment arm, in which nominated teachers were assigned to take part in the Maths-for-Life professional development (PD) programme,[3] and implement the lessons with their students; and ii) the control arm, in which settings continued as they otherwise would have. The trial was complemented by an IPE that focused on the following domains: i) mechanisms (to contribute to our understanding of how the intervention works); ii) fidelity (to assess the extent to which the intervention was delivered as intended; and iii) context (to assess how participant characteristics and the implementation context relate to the effectiveness of the treatment). This mixed-methods design allows us to provide a robust estimate of the causal effect

[3] The Maths-for-Life programme is no longer running, but the Education Endowment Foundation (EEF) is now running an effectiveness trial of the next iteration: Mastering Maths. More information on this trial can be found on the project website, available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/mastering-mathematics-23-24-trial.

of the intervention, while providing information on implementation and process that supports proper interpretation of the impact findings and future decisions about refining and scaling the intervention.

Intervention

This intervention aims to improve GCSE Maths retake outcomes for post-16 students (Key Stage 5) in further education colleges, sixth-form colleges, schools, and training providers. It attempts to develop a more student-centred classroom approach based on problem-solving and dialogic teaching. Dialogic teaching aims to simulate learning through classroom conversation, both peer-to-peer and teacher-to-student. The intervention was designed and implemented by the Centre for Research in Mathematics Education at the University of Nottingham and a group of trained maths teachers from post-16 settings. It builds on an evidence-based corpus of classroom materials, using resources from the Standards Unit Box (Improving Learning in Mathematics) and other resources developed by Malcolm Swan and colleagues (Wake & Swan 2016; Swan, 2007).

The focus is on five key areas of the maths curriculum that are known to be challenging for GCSE Maths students: proportional reasoning; algebraic expressions; parts of a whole; contextual problems; and handling data (Swan and Swain, 2010). The materials address these key mathematical areas and concepts using contexts and problems designed to re-engage GCSE resit students in maths. Many of these students experience disaffection and disengagement after 'failure' (achieving a grade 3 or below) in their Key Stage 4 GCSE exam (Johnston-Wilder *et al.,* 2015). The intervention addresses this by introducing a problem-solving approach, adopting a student-centred focus, using discussion, and using research-informed diagnostic, and formative assessment (Swan and Green, 2002). Tasks are designed to be used with students working collaboratively. For example, students could be given a set of cards with different objects (on a tall skyscraper, the length of a fly, the distance to the moon) and asked to work in groups to match each object to their corresponding measurement.

The intervention supports teachers by providing evidence-informed materials together with a PD programme based on Wake and Swan's Lesson Study research (Swan and Swain, 2010). It takes an 'action research' approach,[4] led by a cadre of trained teacher PD leads in which teacher research groups engage in five cycles of classroom-based inquiry into effective pedagogies, supported by an online toolkit. The PD programme aims to address a skill shortage among teachers and attempts to change how maths is conceptualised by young people, moving from a binary subject where thinking is 'right' or 'wrong' to one that is debated and discussed.

The intervention begins and ends with an event for class teachers. In between these two meetings, the intervention goes through the following Lesson Study cycle (Takhashi and Wake, 2023; Wake *et al.,* 2016; Wake *et al.,* 2020):

1. Clusters of class teachers meet to learn about and plan a Maths-for-Life lesson, supported by their lead teacher.
2. Class teachers teach a Maths-for-Life lesson to their own class.
3. Class teachers meet as a cluster to observe a peer from the cluster teaching the same lesson.
4. Clusters meet again with their lead teacher to reflect on the lesson taught, and to learn about the next lesson to be taught.

This cycle is completed five times, with a new lesson being taught and studied each time. The five lessons are taught between November and April during the timetabled GCSE resit classes. They are taught by GCSE resit maths teachers. Each lesson is designed to last for one hour. The lessons are based on the belief that dialogic learning is essential to improving students' confidence and outcomes. They seek to develop five principles of dialogic learning in classrooms[5]:

1. **Purposeful.** Talk is structured with specific learning goals in view.

[4] 'Action research' is an approach that, 'generally follows a systematic and cyclical pattern of reflection, planning, action, observation, and data collection…that then repeats' (Johnson, 2020).
[5] Description taken from the teacher handbook available at: www.nottingham.ac.uk/maths-for-life/documents/teacher-resource.pdf.

2. **Reciprocal.** Participants list, share ideas, and consider alternative views.
3. **Cumulative.** Participants build on contributions to create chains of thinking.
4. **Collective.** The classroom is a site of joint learning and enquiry.
5. **Supportive.** Ideas expressed freely, without risk of embarrassment.

Five key pedagogical ideas underpin the design of each of the five lessons. Each pedagogy is studied in turn and the effect that it has on one of the principles of dialogic learning examined. These pedagogies are included in all lessons in the actions of teachers and are supported by the design of the resources:

1. **Collaborative learning**. Working together toward a common goal.
2. **Models of structure.** Representations that provide insight into mathematical structure.
3. **Formative assessment.** Assessment that provides information on what to do next.
4. **Cognitive conflict.** Being challenged by new information that contradicts prior ideas.
5. **Closure.** Drawing a lesson to a close to ensure shared understanding.

The five lesson topics are: Parts of a whole; Proportional reasoning; Algebraic expressions; Contextual problems; and Handling data.

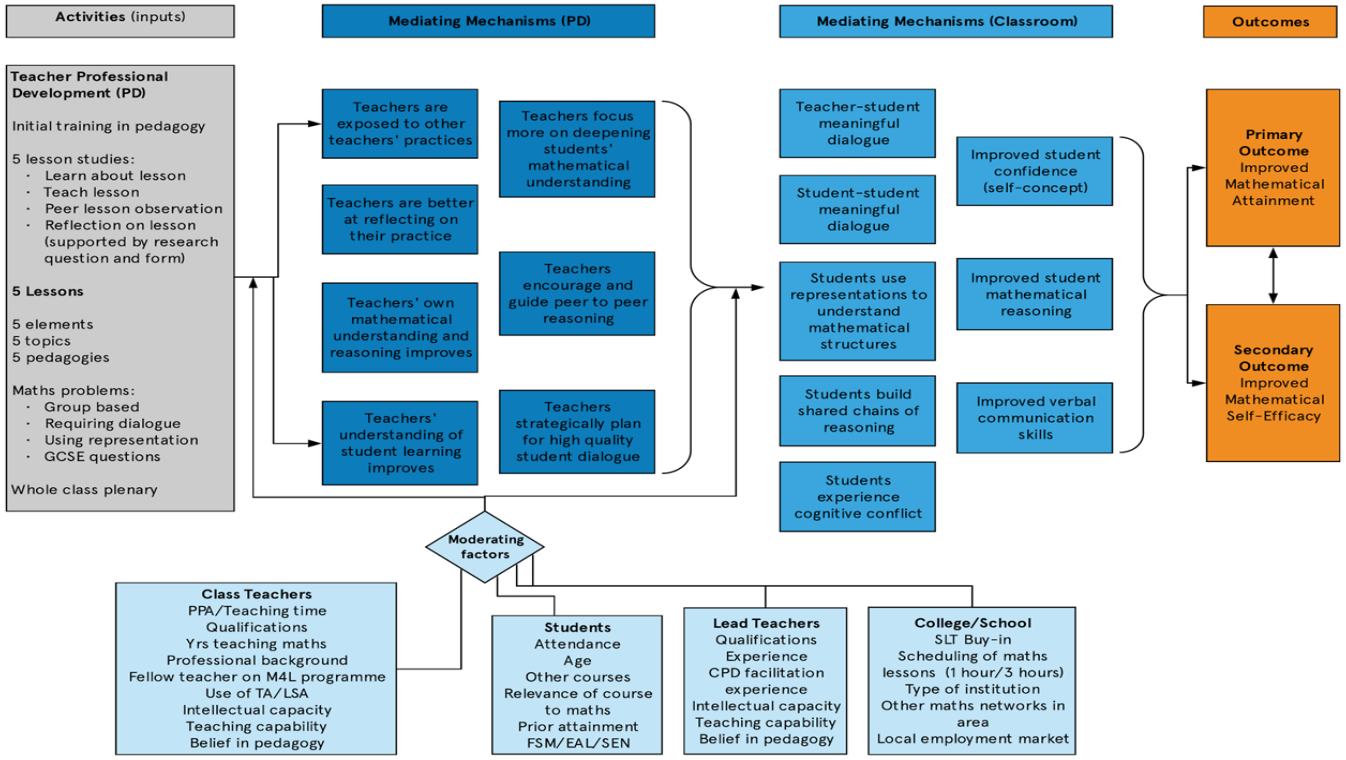
The PD requires a total of six days of teacher time and breaks down as follows:

- **Launch event.** Half a day.
- **Lesson planning and reflection.** One day per lesson (five days in total).
- **Closing event.** Half a day.

Lessons are taught in regular classrooms in schools and colleges and replace regular curriculum content. For the rest of the period, teachers teach their standard curriculum (with the expectation that teachers adopt Maths-for-Life approaches more widely in their teaching). Cluster PD sessions take place in a range of regional locations, such as participating schools and colleges. For a more detailed outline of the intervention, please refer to the Template for Intervention Description and Replication (TIDieR) framework in Appendix E. Information on actual dosage received during the trial can be found in the 'Dosage' findings section below. The programme was piloted with 20 lead teachers, across 20 colleges, in the academic year 2017/2018. In the 2018/2019 academic year, these 20 teachers supported a cohort of approximately five new teachers each through the PD programme. This process saw 84 new teachers applying the Maths-for-Life pedagogy with their GCSE Maths resit classes.

An Intervention Delivery and Evaluation Analysis (IDEA) workshop was held, using the TIDieR framework, to develop a logic model (see Figure 1 below) in collaboration with the developers at the University of Nottingham. The logic model was used to inform the impact evaluation and IPE. The intervention description that was developed can be found in Appendix E and a detailed list of further programme resources are included in Appendix F. These resources are all available at: www.nottingham.ac.uk/maths-for-life/index.aspx.

Figure 1: Logic model



Evaluation objectives

The impact evaluation had one primary research question and two secondary questions.[6]

Primary research question:

1. What is the effect of being assigned to the programme on GCSE Maths resit performance in post-16 education as measured by the moderated Uniform Mark Scale (UMS) point scores?

Secondary research questions:

1. What is the effect of being assigned to the programme on post-16 GCSE Maths resit pass rates, as measured by the percentage of students achieving a grade 4 or above?
2. What is the effect of being assigned to the programme on students' mathematical self-efficacy, as measured by Part E of the Year 10 Teleprism survey?

The IPE aimed to answer the following six questions.

1. **Fidelity:** To what extent do implementers adhere to the intended model? In particular:
   1.1 To what extent do lead teachers adhere to the programme?
   1.2 To what extent do class teachers adhere to the lesson plans?
   1.3 What are the barriers to and facilitators of adherence?

[6] These questions have been slightly refined from the trial protocol to make it clear that the primary analysis is on an intention-to-treat (ITT) basis with two-sided hypothesis tests. As a result, the phrasing refers to 'assignment to the programme' (instead of 'participation in the programme').

1.4 How does non-adherence/adaptation seem to influence outcomes?[7]

2. **Dosage:** How much of the intervention is delivered and received? In particular:
   2.1 To what extent do class teachers receive the recommended amount of PD?
   2.2 To what extent do students receive the recommended amount of Maths-for-Life lessons?
   2.3 What factors contribute to any variation in session number and length?

3. **Responsiveness:** To what extent do participants engage with the intervention? In particular:
   3.1 To what extent do class teachers engage in PD activities?
   3.2 To what extent and how is class teachers' general practice[8] altered by the programme?
   3.3 To what extent do students engage in Maths-for-Life lessons?
   3.4 Are there sufficient resources and support for class teachers (e.g. extension materials for more able students and use of teaching assistants) to allow for effective differentiation in lessons?

4. **Programme differentiation:** To what extent is the intervention distinguishable from existing practice? In particular:
   4.1 Have class teachers (both intervention and control) received PD of a similar nature (either in the past or during the intervention period)?
   4.2 Is the Maths-for-Life teaching approach significantly different from class teachers' current practice? If so, how?
   4.3 Do control group teachers receive PD of a similar nature?

5. **Quality:** How well is the intervention delivered?
   5.1 Can lead teachers effectively facilitate cluster PD sessions?
   5.2 Are the five key pedagogies used effectively by class teachers in the delivery of Maths-for-Life lessons?
   5.3 What factors contribute to variation in implementation quality?

6. **Causal mechanisms:** Are the hypothesised mediating mechanisms present? In particular:
   6.1 Are the hypothesised mechanisms that arise from the PD present?
   6.2 Are the hypothesised mechanisms that arise from the Maths-for-Life lessons present?
   6.3 Are there alternative or complementary mechanisms at play?

The trial protocol (Nolan *et al.,* 2020) is available here, and the Statistical Analysis Plan (Nolan and Taylor, 2020) is available here.

Ethics and trial registration

An independent ethical review was carried out by the ethics panel at King's College London. The review concluded that the consent procedures were appropriate for each type of data gathered and that appropriate data protection protocols were in place. There were no concerns about coercion or potential harm to students.

The developer team from the University of Nottingham recruited settings to the trial (see 'Participant selection' subsection in the 'Methods' section below). Agreement to participate was confirmed through the signing of a Memorandum of Understanding (MOU; see Appendix G).

The trial registration number is: ISRCTN14810016.

[7] See intervention description Appendix E, for hypothesised adaptations.
[8] That is, outside of Maths-for-Life lessons.

Data protection

This protocol sets out the purpose of data sharing, the data to be shared, and the way that the data was managed for the Maths-for-Life research project. BIT acted as the independent evaluator of the Maths-for-Life intervention for the EEF and the University of Nottingham acted as the intervention developer. This document was supplemented by formal data sharing agreements between BIT, the University of Nottingham, and participating settings.

See Appendices G, H, and I for the project MOU, the teacher information sheet, and the student information sheets and withdrawal form, respectively.

Purpose of data sharing

In order to deliver the Maths-for-Life pilot evaluation, it was necessary for the University of Nottingham, participating schools and colleges, and BIT to share data. This included:

- data used to identify participating schools, colleges, and students in the Department for Education's (DfE's) National Pupil Database (NPD);
- student and school-/college-level data provided by participating schools/colleges;
- outcome data (GCSE attainment data and survey data of both students and teachers); and
- interview data relating to both teachers and students.

These data were used for the purposes of the evaluation and were treated with great care to achieve high levels of security. Further information on this process is provided in Table 3 below.

Table 3: Data shared

Data type	Data details	Organisation (originator)	Organisation (recipient)	Date
Participating teachers, schools/colleges contact information	Name Role Work email Telephone	University of Nottingham	BIT	June 2018
Participating student data	Name Date of birth Unique pupil number (UPN) Free school meals (FSM) status	Participating schools/colleges	BIT	October 2018
Randomised school/college data	Name of setting (e.g. college campus name)	BIT	University of Nottingham	October 2018
PD attendance	Teacher names Register of attendance	University of Nottingham	BIT	June 2019
Student attainment data	GCSE Maths grades GCSE Maths raw scores Baseline Key Stage 2 Maths score Baseline Key Stage 4 Maths grade	Participating schools/colleges (for UMS scores) NPD (for baseline Key Stage 2 and Key Stage 4 data and resit grades)	BIT	September 2019
Student maths self-efficacy data	Survey responses from students	Students	BIT	June 2019

At the end of the project, anonymised student data will be made accessible to the EEF, the Fisher Family Trust (who hold the EEF data archive), and the DfE through the EEF archive.

Data processing roles

The evaluation was assessed to have the following data controllers:

- **Controller for teacher contact data.** University of Nottingham.
- **Controller for student demographic data.** Participating schools and colleges.
- **Controller for GCSE Maths grades, baseline Key Stage 2 Maths score, and baseline Key Stage 4 Maths grade.** The DfE.
- **Controller for GCSE Maths raw scores.** Participating schools and colleges.
- **Controller for survey and interview responses.** BIT.

Legal basis for data processing

Participating teacher data

In relation to teacher data that was provided to BIT by the University of Nottingham, the University of Nottingham ensured that participating schools/colleges sought clear consent from their teachers for their data to be processed for the specific purposes of the project.

The University of Nottingham sought contractual assurances from each participating school/college that consent was freely given by way of a positive opt-in by participating teachers for the processing of their data for the purposes of the project. Participating teachers were informed that their contact data would be shared with the University of Nottingham, BIT, and any other third parties.

Participating student data

It was anticipated that participating schools/colleges would process and transfer student data to BIT on the basis of legitimate interests and that they would perform an analysis to: i) identify the relevant legitimate interests; ii) identify that the processing was necessary and there was no less intrusive way to achieve the same result; and iii) undertake a balancing test so that they were confident that students' interests did not override the legitimate interests.

BIT undertook its own legitimate interest's assessment in respect of each proposed transfer of student data from participating schools/colleges. When balancing the legitimate interests against the interests of the individuals involved, we concluded that:

- Using data in this way has little to no impact on the individual students, other than to the extent that students have a positive interest in improving the efficacy of their education.
- As recommended by the Information Commissioner's Office (ICO), we gave students the ability to object to participating in the study, giving students the relevant information with which to make this decision, including the legitimate interests we were relying on.
- Student demographic and attainment data are routinely used for research purposes in a variety of ways, and hence, there is a strong argument that research of this type is something that students and parents should reasonably expect to take place provided it has little impact on them individually.

On balance, given the extent of the legitimate interest in studies of this type and the limited impact on individuals, coupled with individuals being given a right to object to inclusion in the study, we would suggest that there was a clear legitimate interest basis for processing data in this way.

Data transfer between colleges, schools, the University of Nottingham, and BIT

The project involved transferring potentially sensitive student data between the schools/colleges and evaluation team (BIT). Such data must be transferred securely, and the following processes were followed carefully.

Datasets containing personal and sensitive data were transmitted via a secure file-sharing platform, Accellion Kiteworks. This cloud-based file-sharing service encrypts the files with a unique key using 256AES encryption, both in motion and at rest. Data was also transferred physically, as part of visits to schools/colleges. Any potentially sensitive data was stored on an encrypted USB (universal serial bus) flash drive when in transit.

BIT collected initial student data (UPN, name, date of birth, and FSM) directly from participating schools/colleges. The University of Nottingham supported communication with schools/colleges for this purpose. BIT also collected GCSE Maths resit raw scores for participating students from schools/colleges.

Further GCSE attainment data (GCSE Maths resit grades) was collected from the NPD. As part of the application for access to the NPD, potentially sensitive student data had to be transmitted to the DfE. This was done using the DfE's secure service. Matched, anonymised data was deposited in the Office for National Statistics (ONS) Secure Research Service (SRS) environment for analysis.

BIT collected survey responses (on Maths self-efficacy) from students in participating schools/colleges via an online platform (Qualtrics).

These processes were reflected in the data sharing agreements between participating schools/colleges, BIT, and the University of Nottingham.

Data storage at BIT

BIT took reasonable steps to ensure that:

- personal and sensitive personal data was only accessed by those who were part of the project on a need-to-know basis;
- all BIT analysts working on this project had received a cleared Disclosure and Barring Service (DBS) check within the last year;
- unauthorised staff, contractors, and other third parties were prevented from gaining access to the data provided;
- all computer systems and other data storage devices that contain personal or sensitive personal data were password protected;
- workstations / personal computers (PCs) that contain personal or sensitive personal data were not left signed on when not in use;
- all discs, other removable media, or printouts were locked away when not in use in BIT's secure data room;
- no personal or sensitive data was transmitted via unencrypted email;
- no personal or sensitive personal data was left on public display in any form, with all staff member desks cleaned at the end of each day and sensitive material locked away safely;
- paper files were stored in BIT's secure data room (only accessible by a limited number of people in the research and evaluation team who have the door code);
- the office shredder or other contract shredding service was used to dispose of any document containing personal data (electronic or otherwise) after use;

- all datasets and do-files, for Stata or any other statistical software package, were stored on an encrypted, regularly backed up, team hard drive; and

- sufficient training was provided to all staff members to ensure they understood the importance of data security and, in particular, exercised appropriate care when handling personal and sensitive information.

Data storage at the University of Nottingham

The University of Nottingham's project data was held locally and for this purpose they have the use of a secured partition on the University of Nottingham's research network drive with access provided to the four evaluation team members and our administrative support, as well as specific information technology (IT) support personnel. Access to the network drives was tied to researchers' workstations, which were situated in private locked offices. The terminals were all password protected and authenticated through the University's centralised Information Access Management system. Remote access, where permissions allow for it only worked through a remote desktop, which had the same password and authentication requirements.

Data destruction

BIT and the University of Nottingham will destroy personally identifiable data within a period of three months of the end of the trial following their own institutional practices. For the University of Nottingham, the University currently uses KillDisk, which conforms to US Department of Defense clearing and sanitising standard DoD 5220.22-M for removal of data from active standalone computers. For networked PCs the specific partition will be locked from access, backup processes turned off, and the data overwritten.

If the computer is being decommissioned at the same time the data is deleted, then the deletion will be carried out by the University's third-party PC disposal partner. Otherwise, it will be completed by IT Services staff. Data that is stored on hard copy (paper) will be shredded on-site prior to disposal.

Use of the data

At no time will individual- or school-level data be disclosed to any third parties (with the exception of end of project transfer to the EEF or its contractors, as noted above). It will only be used for purposes connected with this evaluation including, but not specifically limited to:

- checking randomisation has worked through analysing the average characteristics of individuals and schools/colleges in the treatment and control groups; and

- calculating the estimated impact of the project by comparing the outcomes of individuals in treatment and control schools/colleges.

This includes use of the data for academic publications by the project and evaluation teams, which will follow the same standards in terms of ensuring confidentiality and anonymity of participants.

Project team

The intervention was developed and implemented by the following staff at the Centre for Research in Mathematics Education at the University of Nottingham: Geoffrey Wake; Matt Woodford; Michael Adkins; and Sheila Evans.

This evaluation was delivered by staff at BIT, led by David Nolan and Patrick Taylor. The impact evaluation was designed by David Nolan, Pantelis Solomon, and Michael Sanders. Analysis was conducted by David Nolan, Claire Cathro, and Neus Torres-Blas. Impact data collection was managed by David Nolan, Louise Jones, and Bridie Murphy. The IPE was designed by Patrick Taylor and Jessica Heal. IPE field data collection was conducted by Patrick Taylor.

Methods

Trial design

Table 4: Trial design

Trial design, including number of arms		Two-arm, cluster randomised controlled efficacy trial
Unit of randomisation		Setting level (further education college, sixth-form college, school, training provider, or a sub-site of one of these)
Stratification variable (s) (if applicable)		Setting type – two strata (further education college vs sixth-form college/school/training provider)[9]
Primary outcome	Variable	Key Stage 5 GCSE Maths retake outcomes
	Measure (instrument, scale, source)	Key Stage 5 GCSE Maths UMS score, 0 – 100, from GCSE exam raw scores provided by the settings
Secondary outcome(s)	Variable(s)	1. Key Stage 5 GCSE Maths pass rate 2. Student self-reported mathematical self-efficacy
	Measure(s) (instrument, scale, source)	1. Key Stage 5 GCSE Maths pass rate, 0 – 1, NPD 2. Mathematical self-efficacy survey score, 1 – 4, Part E of Teleprism survey
Baseline for primary outcome	Variable	Key Stage 2 SATs mathematics attainment Key Stage 4 GCSE Maths attainment
	Measure (instrument, scale, source)	Key Stage 2 Maths total marks, 0 – 100, NPD Key Stage 4 GCSE Maths grade, 0 to 9, NPD
Baseline for secondary outcome(s)	Variable	Key Stage 2 SATs mathematics attainment Key Stage 4 GCSE Maths attainment
	Measure (instrument, scale, source)	Key Stage 2 mathematics total marks, 0 – 100, NPD Key Stage 4 GCSE Maths grade, 0 to 9, NPD

The EEF Maths-for-Life trial was a two-arm randomised controlled efficacy trial, with randomisation at the setting level (settings were further education colleges, sixth-form colleges, schools, and training providers). The randomisation was stratified by a binary category of setting type (further education college or sixth-form college/school/training provider).

The two arms were:

- **Treatment arm.** Nominated teachers from settings assigned to this arm took part in the Maths-for-Life programme.
- **Control arm.** Settings continued as they otherwise would have. All settings allocated to the control group were financially reimbursed £1,000 for their continued participation in the trial.

No changes were made to the trial design specified in the trial protocol (Nolan *et al.*, 2020) and Statistical Analysis Plan (Nolan and Taylor, 2020).

The primary outcome was the overall Key Stage 5 GCSE Maths resit performance for the academic year 2018/2019, determined by the UMS score. The UMS score was calculated by BIT from the raw marks collected and provided directly by the settings. This assessment took place during Summer Term 2019.

[9] Stratification grouped sixth-form colleges, schools, and training providers to form one stratum to ensure equal numbers of students in treatment and control in both categories, given they are likely to be aligned in numbers of classes, teachers, and students.

In addition to the primary outcome, the evaluation considered two secondary outcome measures:

- **Key Stage 5 GCSE Maths pass rates.** Assessed based on a binary outcome variable indicating whether a student achieved a grade 4 or higher. This data was obtained from administrative records on GCSE Maths test grades in the Key Stage 4 dataset, Key Stage 5 dataset and the Young Person's Matched Administrative Dataset (YPMAD) from the NPD and the Individual Learner Record (ILR) database.

- **Student self-efficacy in maths.** Measured through Part E of the Year 10 Teleprism survey. Self-efficacy refers to an individual's belief in their ability to succeed in a task. The Teleprism survey was developed by the Teleprism project at the University of Manchester (Pampaka *et al.,* 2011).

The variable KS4_L2BASICS_94 from the Key Stage 4 dataset was not used to calculate the Key Stage 5 GCSE Maths pass rate as outlined in the trial protocol (Nolan *et al.,* 2020). Instead, we combined several variables from the NPD (Key Stage 4, Key Stage 5, and YPMAD datasets) and ILR (Learning Aims dataset) that included the maths grades.[10] Details on how this secondary outcome was constructed can be found in the 'Outcome measures' section below.

Participant selection

The trial was conducted with Key Stage 5 Maths resit students from schools, sixth-form colleges, further education colleges, and training providers that were in the classes of teachers from each setting who participated in the programme.

Settings were recruited by the University of Nottingham. To be eligible, settings had to have students retaking GCSE Maths in Key Stage 5. During recruitment, the settings selected teachers to be part of the programme. The University of Nottingham aimed to recruit one teacher per school and one to two teachers per college. Dosage data indicates that this teacher recruitment target was successfully achieved. The eligibility criteria for participating teachers were as follows:

- **Further education colleges.** Teachers were required to expect to teach at least 80 students (if multiple teachers were recruited, then this figure was a combined total).

- **Schools, sixth-form colleges, and training providers**. Teachers were required to expect to teach at least ten students.

In the case of large further education colleges (or college groups), the teachers could be geographically and organisationally separate. For the purposes of the trial, these teachers were considered to be teaching in distinct settings, even if they came from the same institutional group. Consequently, the risk of spillover was considered to be negligible, and one large college could be classified as two or more settings for recruitment, randomisation, and analysis.

Additional teachers from each setting beyond the specified numbers were only accepted by the programme if there was available space without limiting the capacity for more settings to participate.

Settings that wanted to participate in the trial were required to sign an MOU (see Appendix G) before enrolling in Spring Term 2018. This MOU outlined the necessary activities for both the intervention and evaluation. It included the requirement for each school to collect and provide raw GCSE Maths scores for all participating students. Settings also had to sign a Data Processing Agreement (DPA) with BIT.

Recruitment was widely open to teachers of GCSE resit classes in general further education colleges, sixth-form colleges, schools, and training providers in England. The developers worked extensively through their networks to connect with teachers in these settings through organisations such as the Association of Colleges and their regional networks, the Sixth Form Colleges Association, etc. Recruitment information events were held around the country and social and mainstream media were used to reach potential participants. Participation was not geographically restricted. Recruitment efforts resulted in 50 settings in the treatment group and 50 settings in the control group (approximately 180 class teachers across

[10] See Appendix S, for a summary and explanation of the deviations from the protocol.

both groups). They enrolled 3,071 students and 2,735 students, respectively after randomisation, who were meant to take the GCSE Maths resits in the 2018/2019 academic year. The analysis focuses only on students who did not pass the GCSE Maths resit in November 2018, immediately after intervention was launched that same month, resulting in an effective sample of 2,920 students in the treatment group and 2,536 students in the control group. This amounted to 5,456 students in total.

Outcome measures

Baseline measures

All primary and secondary analyses included two baseline measures obtained from the NPD: Key Stage 2 maths attainment, as measured by the Key Stage 2 Maths raw exam score; and Key Stage 4 Maths attainment as measured by the GCSE Maths grades.

For the Key Stage 2 Maths attainment, we used the NPD variable KS2_MATTOTMRK from the Key Stage 2 dataset. This measure ranges from 0 to 100 and represents the total marks achieved in the Key Stage 2 Maths test (the sum of Paper A, Paper B, and mental arithmetic tests).

For the Key Stage 4 baseline maths attainment, we used the NPD variable YPMAD_GCSE_GRADE_MATHS from the YPMAD. This variable is the best grade recorded in GCSE Maths. The trial was carried out in the 2018/2019 academic year, so we used the best GCSE grade recorded up to the 2017/2018 academic year or the previous non-missing record for each student.

The students that took part in the trial had a wide range of ages, so we had grade records from 2001 up to 2018 that could be either numeric (from 0 to 9) or letters (A, B, C, D, E, F, G, U, X) (GCSE Maths exams grades changed from letter to numeric in 2017). We transformed the letter grades to numeric grades following the Office of Qualifications and Examinations Regulation (Ofqual) guidelines, with some adaptations (Jadhav, 2018). The table of equivalences can be found in Table 5 below.

Table 5: Table of equivalences

Old letter grade	Assigned numerical grade
A*	9
A*	8
A	7
B	6
B	5
C	4
D	3
E	2
F	1.5
G	1
U	0
X	(missing)

The change from letter to numerical grades eliminated grade variability in the lower bound and added more grades above a standard pass. There were four letters below a standard pass (D to G), while there are now only three numbers (3 to 1). We

assigned 1.5 to those students who had an F to keep that variance in the grades. As the sample was made of resit students who had not passed their GCSE before the intervention (so they had low grades), we did not do further adaptations to account for the differences in the upper grades.

The NPD data was matched to the original student data post-randomisation. The matching was carried out using the students' name, school, and date of birth by the DfE, who linked each students' Pupil Matching Reference (PMR) to an anonymised unique identifier provided by BIT. The DfE also used the PMR to match the secondary outcome data to the baseline attainment measures, also obtained from the NPD. The BIT anonymous identifier was then used to link the NPD data to the rest of the student data collected directly from the settings.

Primary outcome

The primary outcome was the overall Key Stage 5 GCSE Maths resit performance for the academic year 2018/2019, as measured by the UMS score. The UMS is a method used to standardise the raw scores of the GCSE Maths exam across different examination boards and smooth out any variations in levels of difficulty from each exam board or exam year (Ofqual, 2013). The exam board then allocates a grade depending on the UMS score. The settings in the trial used four different examination boards, so the raw scores were converted into UMS to allow for comparisons across settings.

The score of the GCSE Maths exam was chosen as the primary outcome measure because it was the most direct way to measure changes in student performance at the GCSE resit exams and mathematical attainment, which was the main outcome of the intervention, according to the logic model (see Figure 1 above).

The GCSE raw exam scores were collected by the settings and sent directly to BIT for conversion. The formula to obtain the UMS scores was provided to BIT by the Assessment and Qualifications Alliance (AQA), one of the exam boards. The formula converts the raw score to a UMS score, which can range from 0 to 100, using the specific grade boundaries from each examination board. The mathematical formula and grade boundaries that were used to convert the scores are provided in Appendix Q.

Secondary outcomes

GCSE Maths pass rate

The first of the two secondary outcomes in the trial was the Key Stage 5 GCSE Maths pass rate, as measured by a binary outcome variable indicating whether or not a student achieved a grade 4 or higher (out of 9) in that exam. The data was obtained from the NPD and the ILR.[11] This outcome measure was chosen for two main reasons: i) because the main aim of the intervention was to improve mathematical attainment of post-16 GCSE resit students, so they can pass the resit examinations; and ii) the evaluation team opted to use data from the DfE as administrative registers are usually more reliable than data reported by the schools/colleges, and have a lower risk of attrition or missing data issues.

The data was collected from the Key Stage 4, Key Stage 5, and YPMAD datasets in the NPD, and from the Learning Aims dataset from the ILR. We used the exam grade of the GCSE Maths exam of Summer Term 2019. Relevant records were identified by filtering the exam data by season (summer), year (2019), subject code or 'mapping' (2210 – Mathematics), and type of qualification or 'sublevno' (391 – GCSE). In the case of duplicated records across different datasets, we prioritised YPMAD records, as the YPMAD contains information on students' highest educational qualification up to age 20 and links other NPD datasets, followed by ILR, Key Stage 5, and Key Stage 4, in that order.

All students who received a grade 4 or above were coded as 1, and the rest of the students were coded as 0, including those that did not sit the exam and/or did not have an exam record.

[11] The protocol specified that variable KS4_L2BASICS_94 from the Key Stage 4 dataset was to be used for this outcome, but this was not feasible. See Appendix S, for more details about the deviation from the protocol.

Mathematical self-efficacy

The other secondary outcome was mathematical self-efficacy, as measured through Part E of the Year 10 Teleprism survey. Mathematical self-efficacy was identified as a key mechanism supporting attainment in this intervention, based on findings from the pilot evaluation.

The Teleprism survey was developed by researchers at the University of Manchester. The evaluation team used a subsection of the survey, Part E, which focuses on self-efficacy. The survey had 21 questions, and asked students to rate their confidence levels on a four-point Likert scale, ranging from 'Not confident at all' (1) to 'Very confident' (4) in answering example questions related to various GCSE Maths topics: number; algebra; geometry and measures; ratio; proportion and rates of change; and statistics. The survey asked for the students' confidence in answering them but did not ask them to actually solve the mathematical problems. An example question of the survey was: 'How confident are you to solve mixed-fraction problems such as:' and then showed an example. An example question from the survey can be found in Appendix R.

A final survey score was obtained for each student from the average of all 21 Likert-scale responses. The final score ranged between 1 to 4. Although a four-point Likert scale may have limited variance, it was considered the most suitable measure for this project as the survey had been developed and validated for post-16 students and was also found to be consistently correlated to mathematical attainment for these students (Pampaka *et al.,* 2011). The survey was self-administered by students using an online survey platform in May 2019. BIT created this online version of the survey and emailed it to class teachers with instructions on how to administer it with their students. This approach was taken (as opposed to researchers visiting settings to administer the survey) to save money. For those students who submitted several entries, we used the latest completed submission for the analysis.

Sample size

Prior to recruitment (at the stage of writing the trial protocol), we used a statistical process known as simulation-based inference to conduct all power analyses, using the R statistical software package. This was because we intended to recruit different types of educational settings with substantial variation in cluster size, and large variation in cluster size tends to reduce statistical power in cluster randomised controlled trials (Lauer *et al.,* 2015).

Further education colleges are usually much bigger in size than other post-16 settings, so we expected a substantially bigger number of students taking Maths-for-Life lessons in each of the further education colleges compared to the rest of post-16 settings in the trial. We conducted thousands of hypothetical experiments with a given effect size. We then observed how many of these experiments produced a significant result. This approach is known as the Monte Carlo method for measuring experimental power. In this particular case, the combination of setting randomisation and the presence of two distinct 'types' of settings with varying size made it necessary to adopt this approach. Our simulations assumed that cluster sizes followed a Poisson distribution, with average cluster sizes of 80 students per further education college and ten students in the other types of settings.

However, post-recruitment cluster sizes were known, therefore, we deferred back to using closed-form algebraic equations to conduct our power analysis, and validated the results from the simulation-based approach taken in the trial design phase. We factored the variation in cluster sizes into the power analysis by using an inflation factor known as the coefficient of variation, which was computed with the actual cluster sizes after recruitment. We used Stata software, Version 16 (StataCorp LLC, College Station, Texas, USA) to compute the minimum detectable effect size (MDES) at randomisation and analysis stages. The calculation of the MDES at randomisation was based on the assumptions at the design stage, except for the number of clusters and cluster sizes for the overall sample and the FSM subsample. The updated cluster sizes at randomisation are shown in Table 6 below.

We present in Table 6, the sample and the MDES at the design of the trial protocol, randomisation, and analysis stages to achieve 80% statistical power for our primary outcome measure, Key Stage 5 GGSE Maths resit performance, as measured by UMS scores. The MDES is presented in terms of Hedges' g. The assumptions were updated at the analysis stage, as can be seen in Table 6, and computed using the sample of students that were included in the primary outcome analysis.

Table 6: MDES at different stages

		Protocol		Randomisation		Analysis	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.15	0.22	0.30	0.34	0.20	0.25
Pre-/post-test correlations	Level 1 (learner)	0.5	0.5	0.5	0.5	0.55	0.56
	Level 2 (class)[a]	N/A	N/A	N/A	N/A	N/A	N/A
	Level 3 (school)[b]	N/A	N/A	N/A	N/A	N/A	N/A
Intracluster correlation coefficients (ICCs)	Level 2 (class)[c]	N/A	N/A	N/A	N/A	0.17	0.19
	Level 3 (school)	0.2	0.2	0.2	0.2	0.08	0.09
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		80 students per college, ten per school	24 students per college, two per school	73 students per college, 30 per school/sixth-form college/ training provider	22 students per college, ten students per school/ sixth-form college/ training provider[d]	43 students per college, 16 per school/sixth-form college/ training provider	14 students per college, six per school/ sixth-form college/ training provider
Number of settings (further education college / sixth-form college / school / training provider)	Intervention	55	55	50	50	45	44[e]
	Control	55	55	50	50	46	45[f]
	Total:	110	110	100	100	91	89
No. of students	Intervention	2,300	660	3,071[g]	996	1,631	591
	Control	2,300	660	2,735[h]	800	1,401	426
	Total:	4,600	1,320	5,806[i]	1,796	3,032	1,017

[a]*This is not estimated as most students participating in the trial had moved to a new educational institution and we did not administer any pre-test assessment.*
[b]*Id. from footnote 6.*
[c]*The ICC at the class level was not used to calculate the MDES at the analysis stage, as the randomisation was clustered at the setting level.*
[d]*The cluster sizes of FSM learners at randomisation and the MDES in this report are different from the sizes reported in the Statistical Analysis Plan (Nolan and Taylor, 2020) because we used the FSM status from the NPD records, while setting records were used for the MDES calculations in the Statistical Analysis Plan (Nolan and Taylor, 2020).*
[e]*There was one school in the intervention arm and one in the control arm that did not have any FSM-eligible learners, so the number of clusters in the FSM subgroup is smaller than the number for the overall sample.*
[f]*Id. from footnote 10.*
[g]*This figure was 3,240 in the Statistical Analysis Plan (Nolan and Taylor, 2020). The difference in the number of observations is because three colleges that were in the treatment group included 166 students in their first lists used for randomisation who should not have been included in the study, as they were never taught by teachers taking part in Maths-for-Life. This was detected in the follow-up with the settings and the students were excluded from the study. Additionally, three students that changed their name were duplicated with different names in the student records and the duplicates were subsequently dropped.*
[h]*The difference in total students between treatment and control groups is driven by a higher average number of students per setting in the treatment group (65) relative to the control group (55). However, we conducted balance checks, and this difference was not statistically significant at randomisation.*
[i]*This number falls to 5,456 when ineligible learners (those who passed their November 2018 resits immediately after intervention launch) are excluded. This exclusion was pre-specified.*

Justification for the assumptions in Table 6 above are as follows:

- **Alpha and power.** We used standard assumptions of 80% statistical power and 5% significance level.
- **Pre-/post-test correlations.** In the protocol and randomisation stages, we assumed that 50% of the variance in primary outcome would be explained by previous Key Stage 2 and Key Stage 4 attainment in mathematics (for the whole sample and the FSM subgroup). Given that the literature in this domain is quite sparse, there were no formal correlations available to use. Menzies *et al.* (2021) estimate the correlation between Key Stage 2 and Key Stage 4 Maths examinations at 0.7, using 2019 data from the DfE. We made a conservative estimate of 0.5. At the analysis stage, we adjusted this assumption based on our data. We calculated the R^2 for the baseline covariates by regressing the two covariates on our outcome and obtained the correlation as the square root of the R^2. We found that the correlation coefficient was 0.55 for the full sample and 0.56 for the FSM sample, which were slightly higher than our previous assumptions.
- **One or two-sided test.** A two-sided test was performed to be conservative.
- **ICC.** We assumed 0.2 at the protocol and randomisation stages based on previous BIT further education trials (Hume *et al.,* 2018). The ICC at the setting level for the analysis stage was half the size and calculated with a one-way analysis of variance (ANOVA) at 0.08 for the full sample and 0.09 for the FSM sample.
- **Student attrition.** We assumed 20% of attrition in the protocol and randomisation. The attrition rate at the analysis stage was 47.78%, much higher than the initial assumptions. This was both from loss of students and loss of settings after randomisation and at the analysis stage (due to missing covariates). See the participant flow diagram in Figure 2 and 'Attrition' subsection in the 'Impact evaluation results' section below for a discussion on the attrition in the trial.
- **Outcome distribution.** We assumed Key Stage 4 GCSE Maths outcomes followed a normal distribution with a mean UMS score of 50, and a standard deviation (SD) of 20. This assumption is based on a number of sources, namely: Ofqual data on distribution of numerical grades (1–9) (Ofqual Analytics, 2017); the Key Stage 4 GCSE Maths pass rate of 73% (Murray, 2017); and the following average raw-score grade boundaries: grade 4 – 33; grade 3 – 17; and grade 2 – 10 (Pearson Edexcel, 2015). At the analysis stage, the mean and SD of the primary outcome in the control group were 32.60 and 7.94.
- **Homogeneous treatment effect.** For the simulations at the protocol stage, we assumed the simulated treatment effect to be uniform across all participants in the treatment group, as there was not, ex ante, any empirical reason to assume any particular different functional form.
- **Percentage of FSM students in schools and colleges.** At the protocol stage, we assumed 20% of FSM students in schools and 30% in further education colleges.[12] For the sample calculations at randomisation in the Statistical Analysis Plan (Nolan and Taylor, 2020), we collected the FSM status directly from the settings. During the collection process, the settings flagged that the FSM status was difficult to provide, and they were unsure of the accuracy of their data. Table 6 above, shows the updated calculations at the randomisation stage,[13] where we used the data on FSM status from the NPD. At randomisation, 31.77% of students across all settings were eligible for FSM. This percentage at the analysis stage was 33.54%.

Even though the number of recruited students was bigger than what was indicated at the protocol stage, the MDES was bigger at the randomisation stage (0.30 in Hedges' g) compared to the protocol stage (0.15 in Hedges' g). This was because

[12] A conservative estimate revised upwards from the standard of 20% given further education college students are drawn from disproportionately disadvantaged backgrounds.

[13] The cluster sizes of FSM learners at randomisation and the MDES in this report are different from the sizes reported in the Statistical Analysis Plan (Nolan and Taylor, 2020) because we used the FSM status from the NPD records, while setting records were used for the MDES calculations in the Statistical Analysis Plan (Nolan and Taylor, 2020).

ten fewer settings/clusters were recruited than specified in the pre-trial power analyses, and the real distribution of settings/cluster sizes was used at the randomisation stage instead of a simulated one.

The MDES at the analysis stage was 0.20 for the primary outcome analysis and 0.25 for the FSM subgroup analysis. This was 2/3 of the MDES at randomisation stage. The substantially lower ICC of the outcome at the setting level, compared to what we assumed at randomisation, may have contributed to this difference. However, an effect of 0.2 SD is still large compared to the estimated mean average effect size in the EEF trials, which is 0.04 SD (Demack *et al.,* 2021),[14] and the MDES assumed at the design stage (0.15 SD). The trial may therefore, have been underpowered for the primary analysis and the FSM subgroup analysis. However, the EEF (and other What Works Centres) still considers 0.20 SD to be a benchmark of a meaningful effect (representing three months of additional progress).

Randomisation

Randomisation was clustered at the setting level and stratified by setting type. Setting type was established as a binary variable indicating whether the setting was either a further education college or sixth-form college/school/training provider. The grouping of the setting types into two categories was done because further education colleges had a much bigger number of students than the other three setting types. Average cluster size at randomisation was 73 students per further education college, and 30 per sixth-form college/school/training provider (see Table 6 above).

Randomisation was done by BIT researchers using statistical software Stata 14, while recruitment was carried out by the University of Nottingham. After recruitment, settings sent their teacher and class lists to BIT in September 2018 and BIT randomised all settings into the treatment and control group, stratifying by type of setting as indicated above.

Statistical analysis

All analyses were carried out using the statistical software Stata 14. The analysis plan is described in the following sections.

Primary analysis

Primary analysis was done on an ITT basis in which we tested the hypothesis that having a teacher that was assigned a place on the programme had an effect on student performance. Analysis was carried out using an ordinary least squares (OLS) regression, specified below.

$$Y_{is} = \beta_0 + \beta_1 T_{is} + \alpha X_{is} + \varepsilon_{is} \quad (1)$$

where:

- Y_{is} is the outcome for the Key Stage 5 GCSE Maths resit performance for individual i, in setting s, measured by UMS score.
- T_{is} is a binary indicator for the treatment assignment for individual i in setting s (1 if the setting s is assigned to treatment and 0 if not).
- X_{is} is a vector of individual-level and setting-level covariates including a categorical variable for the type of setting (further education college, sixth-form college, school, and training provider), baseline attainment (measured through both Key Stage 2 raw maths scores and Key Stage 4 GCSE Maths grade), and a dummy variable for whether a student was a recipient of FSM.
- ε_{is} is the cluster-robust error term, for individual i, in setting s, clustered at the setting level (assuming the errors are correlated within setting and reflecting the design of the study).

[14] This analysis pooled a lot of different interventions, age groups, and subjects, so it is not directly comparable, but still provides a helpful reminder that it is difficult for an intervention to have a large effect in the English education system.

While UMS scores are bounded between 0 and 100, we assumed that the response to the treatment would be locally linear so an OLS would be appropriate (in any case OLS gives the best linear approximation).

The above model did not account for any class-level clustering effects, only for setting-level correlation. We opted to use an OLS regression as opposed to random effects specifications as random effects models require strict exogeneity of the regressors with the error term. Because of the difficulty in ensuring that all these factors are strictly exogenous, OLS was considered a safer choice as it can still provide unbiased estimates under weaker assumptions.

Additionally, in the design stage, we assumed that students could switch classes throughout the year, and class composition could change, which would dilute any clustering effect at this level. However, not accounting for class-level clustering might impact our standard errors, so we also ran a Hierarchical Linear Model (HLM) as a robustness check that incorporates the multilevel clustered nature of the data into the specification. In the 'Impact evaluation results' section below, we compare the results from our HLM against our primary analysis.

Secondary analysis

For the secondary outcome analysis, we used the same model specified for the primary outcome (an OLS on ITT, with the same covariates, and standard errors clustered at the setting level) but changed the dependent variable Y_{is} to Key Stage 5 GCSE Maths pass rate and to the self-efficacy in maths mean survey score.

We chose to fit a linear model as opposed to a logistic regression for our binary outcome measure (Key Stage 5 GCSE Maths pass or fail) as we anticipated a sample average for the pass rate close to 28%. At this point of the distribution (far from the tails), a linear model would approximate the results returned from a logistic regression and has the advantage of easier interpretation of parameter estimates. Even though the sample average for the control group was closer to 0 than expected, at 13%, the results remained qualitatively unchanged by using a logistic regression. This can be seen in Appendix J, which compares the results using logistic regression and linear regression to estimate the treatment effect on the binary GCSE Maths pass/fail rate. Given the binary nature of the data, instead of Hedges' g, we standardised the effect size using the Cox Index.[15]

Analysis in the presence of non-compliance

We employed an instrumental variable approach to estimate the Complier Average Causal Effect (CACE), to check for the potential dilution of the treatment effect caused by one-sided non-compliance, where some individuals assigned to treatment do not participate.

For this trial, a student was defined as compliant if all Maths-for-Life teachers in their setting taught at least three Maths-for-Life lessons. Due to the unavailability of identifiers to link dosage data from teachers to their students, compliance had to be defined at the setting level, which represents a limitation of this approach.[16] It is important to note that our measure of compliance is both:

- based on self-reported data,[17] and therefore, may not accurately reflect teachers' fidelity to the programme; and
- did not reflect whether a student received the treatment, as we did not have student attendance records, so we did not observe whether any given student had attended that class.

[15] The Cox Index is calculated as Cox Index = beta / 1.65, where beta is the estimated treatment coefficient in the logistic regression and represents the difference between the log-odds ratio of the control and intervention group, adjusted for covariates.
[16] This definition of compliance was a deviation from the Statistical Analysis Plan (Nolan and Taylor, 2020). See Appendix S, for more details.
[17] The data was collected by lead teachers from class teachers at cluster training days. Class teachers were asked to complete a register that logged whether or not they had delivered the lesson for that period for any given class that they teach. This data was used by the University of Nottingham to monitor delivery and sent to BIT at the end of the project for use in the CACE analysis.

However, this is the only feasible compliance indicator that has been identified.

We estimated the CACE using a two-stage least squares (2SLS) approach. We estimated the following model:

$$Z_{is} = \gamma_0 + \gamma_1 T_{is} + \delta X_{is} + u_{is} \text{ (First stage)}$$
$$Y_{is} = \beta_0 + \beta_1 \hat{Z}_{is} + \alpha X_{is} + \varepsilon_{is} \text{ (Second stage)}$$

where:

- T_{si} is a binary indicator for the treatment assignment (1 if the setting s is assignment to treatment and 0 if not).
- Z_{is} is a binary variable equal to 1 if all teachers in settings delivered at least three lessons to all their classes.
- $\hat{Z}_{is}$ is the predicted compliance indicator with the programme from the first stage.
- X_{is} is a vector of individual-level and setting-level covariates including a categorical variable for the type of setting (further education college, sixth-form college, school, and training provider[18]), baseline maths attainment (measured through both Key Stage 2 raw maths scores and Key Stage 4 GCSE Maths grade), and controlling for whether a student was a recipient of FSM.
- u_{is} are Newey-West robust standard errors.
- ε_{is} are Baum–Schaffer–Stillman 2SLS errors.
- Y_{is} is the outcome for the Key Stage 5 GCSE Maths resit performance for individual i, in setting s, measured by UMS score.

The percentage of students in compliant settings in the randomisation sample and analysis sample is reported together with the model results, as well as the unadjusted means for compliant and non-compliant students.

Missing data analysis

We started with reporting the number of complete observations (those without any data missing) for all three outcomes, the number of eligible observations with outcome data, and the rates of missing data for all covariates in the primary analysis sample. This identified the following types of missing data:

- missing pre-treatment covariates; and
- missing outcome data.

Altz and Grimes (2002) suggest that, when less than 5% of data is missing, there is likely to be little bias introduced to estimated treatment effects. We have adopted this threshold here as an indicator of potential bias in the analysis that justified conducting further missing data analysis outlined below.

Missing pre-treatment covariates

As baseline attainment covariates were missing for more than 5% of the primary outcome sample, we tried to establish which variables were predictive of the missing data. To do this, we created a new variable that is a binary indicator of

[18] The category 'training provider' was not included in the trial protocol (Nolan *et al.,* 2020) as it was not anticipated, but the developer has recruited a small number of these settings, so we have added it here.

missingness and looked for its predictors using a logistic regression model. Missing Key Stage 2 attainment data was modelled as follows:

$$M_{is} \sim binomial(p_{is});\ logit(p_{is}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (2)$$

where:

- M_{is} is the binary variable for missingness (equal to 1 if missing and 0 if not missing).
- p_{is} is the probability that a given observation is missing the Key Stage 2 Maths score.
- x_1 is the FSM eligibility variable.
- x_2 is the Key Stage 4 GCSE Maths grade.
- x_3 is the categorical variable for the setting type.
- x_4 is the binary indicator for treatment allocation.

The same model was used to model missingness for the Key Stage 4 GCSE Maths grade, substituting Key Stage 4 GCSE Maths grade for Key Stage 2 Maths score in the specification above.

Where both Key Stage 2 Maths score and Key Stage 4 GCSE Maths grade were missing, the following model was used,

$$M_{is} \sim binomial(p_{is});\ logit(p_{is}) = \beta_0 + \beta_1 x_1 + \beta_3 x + \beta_4 x_{4_3} \quad (3)$$

where:

- M_{is} is the binary variable for missingness (equal to 1 if missing and 0 if not missing).
- p_{is} is the probability that a given observation is missing both Key Stage 2 Maths score and Key Stage 4 GCSE Maths grade.
- x_1 is the EVER6_FSM_P variable.
- x_3 is the categorical variable for the setting type.[19]
- x_4 is the binary indicator for treatment allocation.

As an exploratory analysis, we re-ran the models including age as another covariate. The results of all logistic regressions are discussed in the subsection 'Missing data of pre-treatment covariates' in the 'Impact evaluation results' section below.

We conducted sensitivity analyses to assess the impact of including and excluding the baseline attainment variables with missing observations and used null imputation with the missing indicator method to analyse the full sample. Lastly, we estimated a model with no covariates apart from the stratification variable (setting type) and compared the results to the main specification.

[19] Please note: x_2 is excluded to keep consistency with the previous models specified above.

Missing outcome data

Primary outcome data was missing for 2008 out of 5,806 students randomised. Of which, 350 of these missing observations were learners who passed their GCSE Maths resits in November 2018, immediately after intervention was launched. These learners were classified as ineligible for the trial, so not missing in the same sense as the remainder. No intermediate outcomes were identified at the design stage as appropriate to use to impute primary outcome data. Observations with missing primary outcome data were therefore, dropped from the primary analysis and a complete case analysis was run.

As an exploratory analysis, in the 'Attrition' subsection in the 'Impact evaluation results' section below, we investigated the reasons for missing primary outcome data. We tabulated all the different reasons that were reported by settings when they sent the final exam data to BIT, separated by treatment and control settings, in addition to the number of students that did not sit the exam for other reasons, and the number of missing outcome observations due to setting-level attrition.

The primary outcome had an attrition rate of 48%, which poses a significant risk of bias in the effect size estimates, even though it was relatively balanced (47% in the intervention group and 49% in the control group). When the learners who passed their November 2018 resits are excluded, the rate of attrition improves to 42% (but this remains a very high proportion). To assess how this attrition might have affected the results, we repeated the analysis for the GCSE Maths pass rate (one of the secondary outcomes), using only the students who were included in the primary outcome analysis. The full sample for the pass/fail rate had a considerably lower attrition rate of 25%, as it includes students who took the exam, but whose raw scores could not be retrieved due to data collection issues (the four settings not returning the data). By comparing the results from the restricted sample (matching the primary outcome sample) with the full sample of students who sat the exam, we wished to determine whether the difference in attrition rates affects the effect sizes and the direction of potential bias.

Note that this analysis does not address attrition from students who did not take the exam but rather highlights bias arising from data collection issues or settings dropping out of the trial.

Subgroup analyses

FSM subgroup analysis

We conducted an analysis on the primary outcome for the subgroup of learners who were registered for FSM according to data from the NPD. To identify FSM eligibility, we combined two variables from the NPD to minimise the missingness: EVERFSM_6_P, the variable that was specified in the Statistical Analysis Plan (Nolan and Taylor, 2020), and EVERFSMage10to15 from the YPMAD, which were missing for 700 and 162 participants, respectively.[20] By complementing the data with YPMAD, we could collect data on FSM status for an additional 538 students.

We used the same model as our primary analysis, with the addition of an interaction term between treatment assignment and FSM status to assess whether there was a significant difference in the treatment effect between FSM students and others, captured by β_2:

$$Y_{is} = \beta_0 + \beta_1 T_i + \beta_2 T_i \times FSM_{is} + \alpha X_{is} + \varepsilon_{is} \quad (4)$$

- Y_{is} is the outcome for the Key Stage 5 GCSE Maths resit performance for individual i, in setting s, measured by UMS score.

[20] EVERFSM_6_P indicates that the student had been recorded as FSM-eligible in the last six years, while EverFSMage10to15 = 1 indicates that the student had been recorded as FSM-eligible in at least one School Census or Pupil Referral Unit (PRU) Census or Alternate Provision (AP) Census from age 10 up to age 15. EVERFSM_6_P was not available for mature learners (aged 19+) (i.e. it was always set to 0).

- T_i is a binary indicator for the treatment assignment for individual i (1 if the student is assigned to treatment and 0 if not).
- FSM_{is} is a dummy variable for whether a student was a recipient of FSM.
- X_{is} is a vector of individual-level and setting-level covariates including a categorical variable for the type of setting (further education college, sixth-form college, school, and training provider), baseline attainment (measured through both Key Stage 2 raw maths scores and Key Stage 4 GCSE Maths grade), and a dummy variable for whether a student was a recipient of FSM.
- ε_{is} is the cluster-robust error term, for individual i, in setting s, clustered at the setting level (assuming the errors are correlated within setting and reflecting the design of the study).

Additionally, we estimated the treatment effect on the primary outcome for the subsample of participants who had been eligible for FSM and compared this to the estimated treatment effect for those not eligible for FSM. This was done by estimating the regression model in the primary outcome analysis for each of these two groups.

Subgroup analysis by setting type

We conducted an analysis on the primary outcome by setting type to assess whether there was a significant difference in the treatment effect between students in different setting types. We used the same model for the FSM subgroup analysis directly above but changed the interaction between treatment assignment and FSM status to an interaction between treatment assignment and a categorical variable for the type of setting (further education college, sixth-form college, school, and training provider). The results were reported as the estimated marginal effect in Hedges' g for each category of setting.

Additional analyses and robustness checks

Estimation of a multilevel model

We conducted robustness checks for all primary and secondary analyses using a HLM, an augmented OLS specification that explicitly accounts for the hierarchical structure of the data. While the main model already clustered standard errors at the setting level, ensuring the treatment effect's confidence intervals (CIs) accounted for clustering, the HLM provided additional insights by partitioning variability at multiple levels (settings, classes, and individuals). This allowed us to quantify the contribution of each level and explore cross-level interactions, which OLS with clustered errors cannot capture.

The HLM requires strict exogeneity of random effects and fixed predictors, which can be assumed for the treatment assignment due to randomisation at the setting level. Our primary specification employed a random intercepts model, allowing average scores to vary by class and setting, rather than a random slopes model, where treatment effects would vary by class and setting. Model fit was evaluated using the Bayesian Information Criterion (BIC), which indicated that the random intercepts model provided a better fit for all three outcomes.

We estimated the following model:

$$Y_{ics} = \beta_0 + \eta_c + \tau_s + \beta_1 T_i + \alpha X_{is} + \varepsilon_{ics} \quad (5)$$

$$\eta_c \sim N(0,\ \sigma_c^2);\ \tau_s\ \sim N(0, \sigma_s^2) \quad (5)$$

where:

- Y_{ics} is the outcome for the Key Stage 5 GCSE Maths resit performance for individual i, in class c, in setting s, measured by UMS score.
- T_i is a binary indicator for the treatment assignment for individual i (1 if the student is assigned to treatment and 0 if not).

- X_{is} is a vector of individual-level and setting-level covariates including a categorical variable for the type of setting (further education college, sixth-form college, school, and training provider[21]), baseline attainment (measured through both Key Stage 2 raw maths scores and Key Stage 4 GCSE Maths grade), and a dummy variable for whether a student was a recipient of FSM.
- ε_{ics} is the idiosyncratic standard-error for individual i, clustered at the setting level.
- β_o is the average intercept.
- τ_s is the error for the setting level.
- η_c is the error for the class level.

The same model was used for the two secondary outcomes, replacing Y by the relevant outcome.

For some settings, we only had one class of students; therefore, we could not estimate distinct class and setting-level errors for those. It must also be noted that for settings that had two teachers participating in the programme, there may be an additional clustering effect at the teacher level. Unfortunately, there were too few settings for which this is the case to include this as an additional component of the estimated error term.

The ICC at the setting and class levels were obtained using post-estimation commands and are reported with the main results.

Exploratory analysis: FSM subgroup analysis with GCSE pass rate

Since the secondary outcome of GCSE pass rate was less affected by the high-attrition rates seen in the primary outcome, we conducted the FSM subgroup analysis on this measure to gather further evidence of the difference in GCSE resit exam performance between FSM and non-FSM students.

Exploratory analysis: Re-estimation of the GCSE pass rate using a logistic regression

We re-estimated the model for the GCSE pass rate, treated as a binary variable, using logistic regression. This was necessary because the baseline pass rate was lower than expected (13% vs 28%). We wanted to check if the results changed significantly due to non-linearity of the predicted outcome, which would indicate a poor fit of the OLS model.

Appendix J shows the results using logistic regression to estimate the treatment effect on the binary GCSE pass/fail rate. Given the binary nature of the data, instead of Hedges' g we standardised the effect size using the Cox Index. The Cox Index is calculated as Cox Index = beta / 1.65, where beta is the estimated treatment coefficient in the logistic regression and represents the difference between the log-odds ratio of the control and intervention group, adjusted for covariates.

The results were similar to the main results obtained by the linear regression models, showing a good fit of the OLS model.

Estimation of effect sizes

All effect sizes in this report were expressed in terms of Hedges' g, using the following formula:

$$ES = \frac{M_1 - M_2}{SD^*_{pooled}}$$

Where $M_1 - M_2$ is the difference in the adjusted mean values of the outcome between the treatment group and control group, as estimated by the relevant coefficient ($\hat{\beta}_1$) in a regression model, and the denominator is the pooled SD calculated as:

[21] The category 'training provider' was not included in the trial protocol (Nolan *et al.,* 2020) as it was not anticipated, but the developer has recruited a small number of these settings, so we have added it here.

$$SD^{*}_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

where:

- n_1 is the effective number of observations analysed in the treatment group, accounting for the design effect due to clustering.
- n_2 is the effective number of observations analysed in the control group, accounting for the design effect due to clustering.
- SD_1 is the SD of the outcome variable in the treatment group.
- SD_2 is the SD of the outcome variable in the control group.

Estimation of ICC

We estimated the ICC using a one-way ANOVA model at the school and class levels.[22] This was done post-test, as we did not collect a pre-test primary outcome but instead used Key Stage 2 scores and Key Stage 4 grades from the NPD for the baseline attainment covariates. Calculating ICCs pre-test was not valid because the trial took place in post-16 institutions, and many students had changed institutions between Key Stage 4 and Key Stage 5.

IPE

Introduction

This section details the research questions, methods, and sampling strategy that were adopted for the IPE. The research questions were developed and prioritised, guided by the pilot evaluation findings and the outputs of the IDEA workshop. This led us to focus on six domains: fidelity; dosage; responsiveness; programme differentiation; quality; and causal mechanisms. The pilot evaluation also allowed us to test a range of methods for their feasibility during this trial. Based on this, we chose a sequential mixed-methods approach, combining quantitative administrative and survey data with a set of qualitative case studies.

Methods

A mixed-methods approach was taken to data collection, combining evidence from interviews, observations, surveys, and administrative data. The approach was sequential, so that findings from each stage were used to inform the approach to subsequent stages (Teddlie and Tashakkori, 2009, p. 120). For example, themes that emerged from early observations were used to refine the design of interview and observation guides for later visits. A case study approach was taken to the qualitative data collection.

Administrative data

Dosage data was collected at two levels. The project leads from the University of Nottingham collected a session log that recorded the number and duration of Lesson Study sessions delivered by each lead teacher. Class teacher attendance rates were also recorded for each session. Lead teachers collected a lesson log that recorded the number and duration of lessons delivered by each class teacher for each of their Maths-for-Life classes. The number of lessons delivered by each class teacher was used for the compliance indicator in the CACE analysis (see above).

[22] The Statistical Analysis Plan (Nolan and Taylor, 2020) established that ICC would not be calculated at the class level as randomisation was done at the setting level. However, we have included it as it is useful information for future trials.

Online surveys

A brief quantitative survey was issued to lead teachers at the end of the intervention to provide data on fidelity and the perceived responsiveness of class teachers. A more extensive quantitative survey was issued to class teachers at the end of the intervention, covering all research themes except dosage. Intervention and control group teachers were also asked to complete a brief quantitative survey to establish whether or not they engaged in intervention-like activities during the period of the trial (for intervention group teachers, these questions were integrated into their wider survey).

Case studies

To gather in-depth qualitative insights across all themes of the IPE, six case studies were conducted, combining observations of PD sessions and Maths-for-Life lessons with interviews with lead teachers, class teachers, and students. The unit of case study was a setting, and case studies were situated within two regional clusters (three per cluster). The sampling strategy is described below.

Observations

Six semi-structured lesson observations were carried out (one per setting). These observations focused on quality, fidelity, responsiveness, and causal mechanisms, and used an observation framework based on the Maths-for-Life 'five key pedagogies' (to assess quality), the relevant lesson plan (to assess fidelity) and the logic model (to assess causal mechanisms). See Appendix K for a copy of the lesson observation guide. One PD session was observed in each of the two regional clusters, covering the same topics as the lesson observations. These observations were also semi-structured, using a set of four facilitation skills developed with the University of Nottingham (to assess quality), the lead teacher facilitation plan (to assess fidelity), and the logic model (to assess causal mechanisms). See Appendix L for a copy of the PD observation guide. The researcher who carried out these observations (of both classroom and PD sessions) has experience of designing and teaching classes that use a dialogic approach in a range of post-16 settings, as well as experience of training professionals to teach classes using this approach. This researcher also led the pilot evaluation, so had developed a good understanding of the intervention.

Interviews

Semi-structured interviews were conducted with lead teachers, class teachers, and students, covering their perceptions of fidelity, quality, responsiveness, and causal mechanisms. The two lead teachers and six class teachers who were observed were interviewed for approximately 45 minutes each, after their observation. Approximately, five students were interviewed, one to one, from each case study setting (29 students in total). See Appendices M, N, and O for copies of the interview guides.

Table 7: IPE methods overview

Research methods	Data collection methods	Participants / data sources	Data analysis methods	Research questions addressed	Implementation / logic model relevance
Case studies	Semi-structured interviews	Students (29) Class teachers (6) Lead teachers (2)	Deductive coding; inductive coding; thematic analysis	1, 3, 5, 6	Fidelity; responsiveness; quality; causal mechanisms
	Semi-structured observations	Lessons (6) PD sessions (2)	Thematic analysis	1, 3, 5, 6	Fidelity; responsiveness; quality; causal mechanisms
Surveys	Online questionnaires	Class teachers: Intervention group (64) Class teachers: Control group (46); Lead teachers (14)	Descriptive statistics	1, 3, 4, 5, 6	Fidelity; responsiveness; quality; causal mechanisms; programme differentiation
Administrative data	Session log (for class teacher attendance); Lesson log (for lesson delivery rate)	Class teachers: Intervention group (84) Lessons (five per class)	Descriptive statistics	2	Dosage

Sampling strategy

All lead teachers, class teachers, and control teachers were asked to complete the relevant online survey. For the qualitative elements of the study, sampling was purposive, with units selected for variation on characteristics that were thought to be particularly relevant to the research questions. Six settings were selected for case studying on this basis. Sampling of these settings was carried out by the researchers, aiming for variation in the following characteristics.

- Perceived lead teacher ability (according to project leads).
- Class teacher engagement (as defined by PD attendance and lesson delivery).
- Setting type (further education college, sixth-form college, school, and training provider).

Within each case study class, the class teacher selected five students for interview who varied in their level of engagement with the intervention. This yielded the following sample.

Table 8: IPE sample

Sampling unit	Sample size
Case studies	
Lead teacher	2
Class teacher	6
Setting	6
Maths-for-Life class	6
Student	29
Endpoint surveys	
Lead teacher	14
Class teacher (intervention)	64
Class teacher (control)	46

The purposive approach to sampling has important implications for the analysis and findings. The aim of this sampling method is to capture the range and diversity of experiences in relation to the research questions. Importantly, this approach is not intended to generate a sample that is statistically representative of either the study population or the wider population from which the total study sample was drawn. As such, reporting the prevalence of an experience in the qualitative findings 'tells us nothing about the prevalence within [either] population' (Ritchie *et al.,* 2014, p. 329). Furthermore, qualitative methods, by their nature, do not collect data in the structured way that is necessary for quantitative aggregation. This is in contrast, for example, to structured survey questionnaires, which collect responses in fixed categories that can be aggregated. The reporting of frequency counts in relation to qualitative findings is therefore, carefully avoided as such counts are at best uninformative and at worst misleading.

Analysis

Three types of data were analysed for the IPE: administrative data; qualitative case study data (interviews and observations); and survey data.

Administrative data analysis

Two pieces of administrative data were analysed to understand compliance and dosage: class teacher lesson logs; and lead teacher session logs. The analysis of class teacher session logs is described in the section on CACE analysis above. To explore dosage, we calculated the following descriptive statistics.

Dosage for teachers:

- % teachers who participated in all PD.
- Mean average and range number of planning sessions attended.
- Mean average and range number of lesson observations attended.
- Mean average and range planning session length.

Dosage for students:

- % teachers who taught all five lessons.
- Mean average and range of number of lessons taught.
- Mean average and range of lesson length.

Qualitative case study data analysis

Qualitative data from the interviews and observations were analysed thematically using the framework approach, which allows in-depth exploration of the data by case and by theme (Ritchie *et al.,* 2014). This consisted of creating a matrix in which to organise the data, based on the topic guides and observation proformas. Data was summarised and displayed in the matrix. This was followed by working through the managed data to draw out the range of behaviours, experiences, and views, while identifying similarities, differences, and links between them.

Survey data analysis

Survey responses were summarised with percentages of the respondent samples falling into relevant categories. For example, on the topic of quality, we calculate:

- % class teachers rating the PD as good or excellent.
- % class teachers reporting that the lead teacher was a good facilitator.
- % class teachers reporting that their lead teacher had the right professional experience for the role.
- % class teachers reporting that their lead teacher had strong understanding of Maths-for-Life pedagogy.
- % class teachers reporting that their lead teacher believed in Maths-for-Life pedagogy.

For the lead teacher survey, the sample covered the whole population of lead teachers but was small (N=14). To account for this, we calculated counts as well as percentages.

This analysis was conducted after qualitative analysis was completed and then triangulated with the qualitative findings during reporting.

Costs

Information on the cost of the intervention was collected from two sources. The University of Nottingham provided the amount that it received from the EEF to subsidise delivery of the intervention. Participating settings provided estimates of the costs that they incurred as a result of participation in the intervention. For the latter category, semi-structured interviews were conducted with six class teachers (three from further education colleges, one from a school, one from a sixth-form

college, and one from a training provider), where teachers were supported to create an itemised list of costs and how they were calculated. These interviews were carried out over the telephone after the end of the intervention. The settings varied substantially in the number of learners participating in the intervention, so the following approach was taken to calculating the cost per learner per year (over three years).

1. Calculating the average direct marginal cost to a setting per learner per year:
   a. A per learner figure was calculated for each direct marginal cost reported by each interviewed setting.
   b. A mean average was calculated for each of these per learner figures.

2. Calculating the value of the EEF subsidy per learner per year:
   a. The total value of the EEF subsidy was divided by 50 (the number of settings in the intervention group) to give a per setting figure.
   b. This figure was divided by three (as no subsidy is required in years two or three) and then by the average number of learners per setting in the interviewed sample. This gave a figure for the EEF subsidy per learner per year, over three years.

3. The costs calculated in (1) and (2) above, were then summed up, to give an average cost per setting per year, over three years.

Timeline

Table 9: Timeline

Dates	Activity	Staff responsible / leading
09 April 2018	IDEA workshop	BIT, the EEF, University of Nottingham
29 June 2018	Introductory session for all participating teachers	Geoffrey Wake, Matt Woodford
By end of July 2018	All settings (further education colleges, sixth-form colleges, schools, and training providers) were recruited	Geoffrey Wake, Matt Woodford, Sheila Evans
11 October 2018	Settings provided data for all students participating in the trial	David Nolan, Louise Jones
12 October 2018	Settings were randomly allocated to two groups by BIT, balance checks were conducted and settings were then informed of their allocation	David Nolan
W/C 15 October 2018	Intervention began (PD programme), IPE commenced	Geoffrey Wake, Matt Woodford, Sheila Evans, Patrick Taylor, Jessica Heal
W/C 13 November 2018	Settings provided updated data for students participating in the trial	David Nolan, Louise Jones
Autumn Term 2018	Project lead session log data collection	Geoffrey Wake, Matt Woodford, Sheila Evans
	Lead teacher lesson log data collection	Geoffrey Wake, Matt Woodford, Sheila Evans
	Sampling of case studies	Patrick Taylor, Jessica Heal
01 December 2018	BIT collected baseline data from the NPD for Cohort 1	David Nolan, Louise Jones
Spring Term 2019	PD observations	Patrick Taylor, Jessica Heal
	Lesson observations	Patrick Taylor, Jessica Heal
	Interviews	Patrick Taylor, Jessica Heal
	Endpoint online IPE surveys	Patrick Taylor, Jessica Heal

Dates	Activity	Staff responsible / leading
30 May 2019	BIT collected secondary outcome data (self-efficacy survey)	David Nolan, Louise Jones, Bridie Murphy
24 June 2019	Intervention ended	Geoffrey Wake, Matt Woodford, Sheila Evans
02 July 2019	Students sat GCSEs	N/A
01 September 2019 – 01 December 2019	BIT collected primary outcome data (raw scores from settings)	David Nolan, Louise Jones, Bridie Murphy
01 December 2019 – December 2024[a]	BIT collected secondary outcome data (student grades) from the DfE	David Nolan, Louise Jones, Neus Torres-Blas
10 April 2025	BIT submitted draft report to the EEF	Patrick Taylor, Neus Torres-Blas
TBC	The EEF peer review and feedback completed	The EEF, peer reviewers
TBC	The EEF and BIT published the report	The EEF

[a]*There were substantial delays in accessing NPD data from the DfE.*
N/A=not applicable; TBC=to be confirmed; W/C=week commencing.

Impact evaluation results

Participant flow including losses and exclusions

Figure 2: Participant flow diagram (two arms)

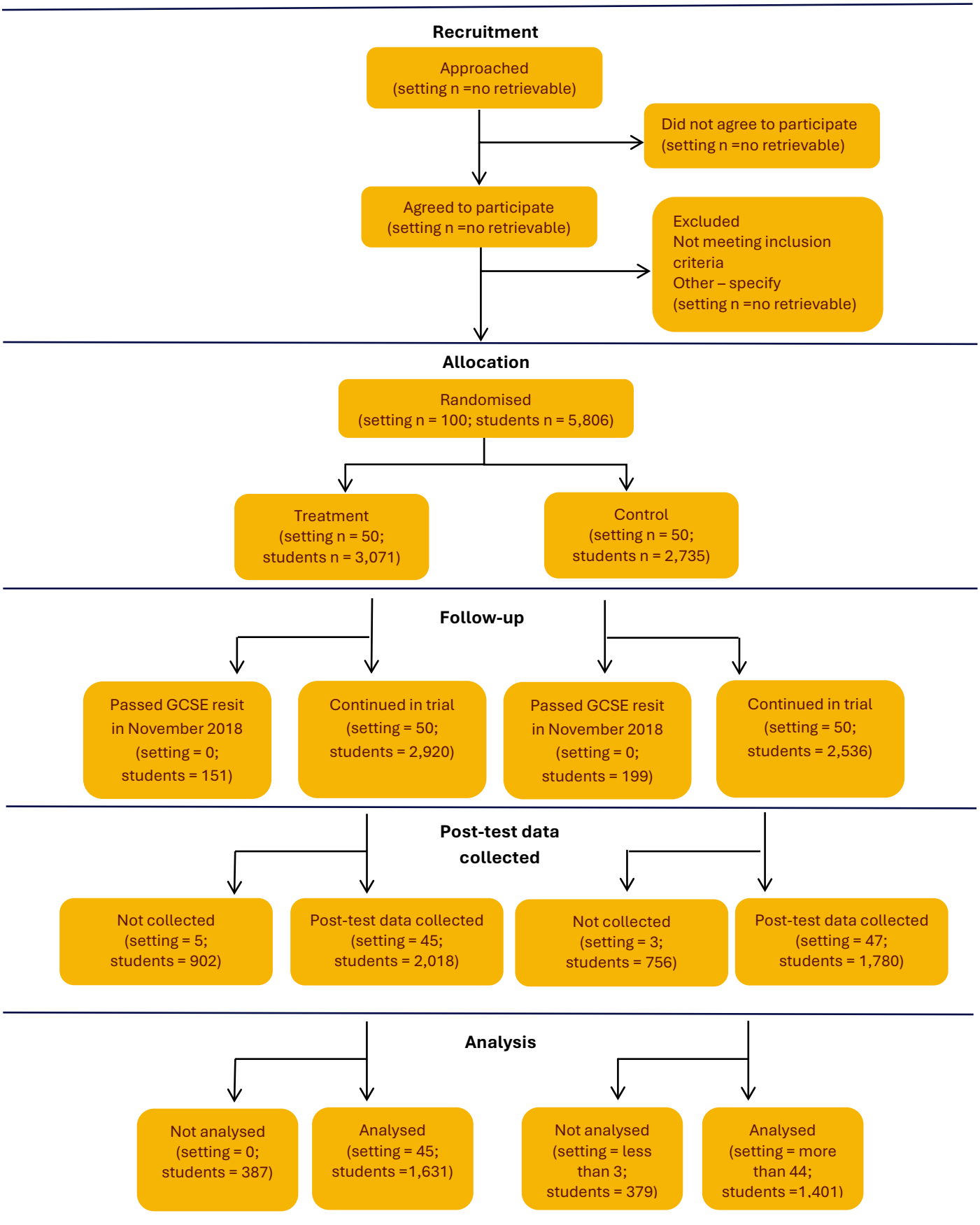


Figure 2 shows the participant flow diagram for the primary outcome analysis. Some exact numbers have been suppressed following guidelines for the extraction of data from the ONS SRS. As indicated in Figure 2, the treatment assignment was randomised in September 2018 to 100 settings. Of which, 50 settings were assigned to the control arm, and 50 other settings to the treatment arm. At that moment, the number of students in the trial was 5,806 with 3,071 in the treatment arm and 2,735 in the control arm. The difference in total students between treatment and control groups was driven by a higher average number of students per setting in the treatment group relative to the control group. However, we conducted balance checks and this difference was not significant at the 5% level.

Please note the reported numbers differ from those noted in the trial protocol (Nolan *et al.,* 2020) and Statistical Analysis Plan (Nolan and Taylor, 2020), which recorded 5,975 students at randomisation, with 3,240 in treatment settings and 2,735 in control settings. This discrepancy arises from two main factors. First, 166 students from two treatment settings were mistakenly included in the original class lists, as they belonged to classes taught by teachers not participating in Maths-for-Life. Once the updated class lists were received by researchers, these students were excluded from any post-randomisation analysis and do not appear in the sample calculations or participant diagram of this report. Second, two students were identified as duplicates in the data, recorded under different names—a discrepancy flagged by the colleges.

In November 2018, some students registered for the trial participated in the winter resit season. Since the intervention began in mid-October 2018, the 350 students who passed the exam in November 2018 did not receive the intervention and were excluded from the analysis. The exam results were released in January 2019, and updated student lists were provided by the settings in May 2019. This adjustment is shown in Figure 2 under 'Follow-up', leaving a total of 5,456 students in the trial after follow-up.

At post-test data collection, five treatment settings and three control settings did not send outcome data to BIT, without providing a reason. Out of those, one school in the control group and one in the treatment group (2% of all settings) dropped out of the trial so they did not send the data, and one setting only sent the GCSE grades instead of the raw marks, so the primary outcome (the UMS score) could not be computed for their students. All this resulted in 469 students missing outcome data for the analysis.

The rest of the primary outcome missing observations resulted from individual student attrition. Students left their college or withdrew from the study for a range of reasons, as reported by the settings. These could include employment, maternity leave, being expelled, no longer studying maths, moving to Functional Skills[23] Maths, refusing to resit the exam, or being older than 19, so they were no longer under the obligation of sitting the GCSEs, among others. See the following section on 'Attrition' for a more detailed breakdown of the reasons for attrition in the trial.

The final number of students included in the primary outcome analysis, after excluding observations because of missing covariates, was 3,032 with 1,631 in the treatment group and 1,401 in the control group. This means 52% of students were lost from randomisation to analysis of the primary outcome, which is a high level of attrition.

The resulting MDES at the randomisation and analysis stages are reported in Table 6 above. For the primary outcome, these were 0.30 SD and 0.20 SD, respectively. This contrasts with the MDES of 0.15 SD assumed at the design stage. The trial was underpowered to detect an effect with a high degree of confidence, primarily due to high attrition—48% compared to the 20% assumed at the design stage.

The first secondary outcome, the GCSE pass rate, was collected from administrative data, as opposed to the primary outcome, so attrition is lower (25%). We coded all observations that did not have a GCSE 2019 as 0 ('did not pass'), so we did not have missing data.

[23] Functional Skills are an alternative qualification to the GCSEs, which are assessed on a pass or fail basis, and which focus on more practical skills. Level 2 Functional Skills are equivalent to a C grade in GCSEs and can be taken as an alternative.

Out of the total 5,806 students randomised in the trial, 4,653 sat the GCSE Summer Term 2019 exam, representing 80.1% of the total sample at randomisation. This included 2,555 students from the treatment group and 2,098 students from the control group. Conversely, 803 students did not sit the exam, accounting for 13.8% of the total. This group comprised 365 students from the treatment arm and 438 from the control arm. The rest of students with no GCSE Summer Term 2019 exam were those that had passed the exam during the winter resit season.

Lastly, the second secondary outcome, the self-efficacy survey, was collected in May 2019 before the GCSE exams and the primary outcome. The survey response rates were low, due to the resource-light approach to data collection. Only 1,405 students (24% of the randomised sample) completed the survey. Apart from the two settings that dropped out, seven treatment settings and 12 control settings did not run the survey with any of their students.

Attrition

The number of learners analysed refers to those with both outcome and complete covariate data, as we run a complete case analysis. Given that the primary outcome (UMS score) and the other two secondary outcomes come from different sources and were collected at different times, the rate of attrition varies across analyses. Below, we present the attrition rate for each outcome individually.

Primary analysis on UMS

Table 10: Student-level attrition in terms of missing primary outcome UMS

		Intervention	Control	Total
No. of students	Randomised	3,071	2,735	5,806
	Analysed	1,631	1,401	3,032
Student attrition (from randomisation to analysis)	Number	1,440	1,334	2,774
	Percentage (no. lost / randomised)	46.89%	48.78%	47.78%

The total rate of attrition for the primary outcome was 48%, which is very high. The differential attrition (the difference in attrition rates between treatment and control groups) was small, at 1.89 percentage points. Of which, 350 of these missing observations were learners who passed their GCSE resits in November 2018, immediately after the intervention was launched. These learners were classified as ineligible for the trial, so not missing in the same sense as the remainder. When these are excluded, the rate of attrition improves to 42% but remains very high.

The reasons for student attrition and exclusion documented by colleges are presented in Table 11 below and is based on information obtained from open-text surveys from settings, except from the number of students that passed the November 2018 exam, which was obtained through NPD records. If a setting did not provide any justification for why the raw exam score was missing, it has been counted as 'Unknown'. Each setting had different ways of reporting the reason as well. For example, some settings documented students as having anxiety, while some others may have coded a similar case as illness, or said that they simply missed the exams. Each of these examples has been coded as different categories in Table 11.

Table 11: Reasons for learner attrition (primary outcome)[a]

Reasons	Control	Intervention	Total	Percentage over all attrition
a. Anxiety	<10	<10	<10	<0.5%
b. Deferral	20	0	20	1.00%

Reasons	Control	Intervention	Total	Percentage over all attrition
c. Family/childcare/employment	<10	<10	<10	<0.5%
d. Illness/bereavement/exemption	54	<10	<64	<3.19%
e. Missed exam for other reason	21	41	62	3.09%
f. Poor attendance	13	<10	<23	1.10%
g. Left for Functional Skills/Apprenticeship/Traineeship	<10	<10	<20	<1.00%
h. Passed the resit in November 2018[b]	199	151	350	17.43%
i. Left setting/course for other reason	64	164	228	11.35%
j. Missing data from setting	95	316	411	20.47%
k. Setting dropped out of trial	47	11	58	2.89%
l. Reason unknown/not reported	429	338	767	38.20%
Total	955	1,053	2,008	100.00%

[a]Counts lower than ten and the corresponding percentages have been censored to avoid identification.
[b]There were 27 students that passed the resit in November 2018 and still took the Summer Term GCSE resit exam in 2019. They have been excluded from the analysis in the 'follow-up', but have not been counted as attrition, because the school sent their primary outcome data.

The main reasons for missing a UMS score were the following:

- **Students not sitting the exams.** Some students did not attend the exams due to several reasons, ranging from having anxiety (a) to self-reported or certified illness or exceptional family circumstances (d). Some reported having childcare and employment obligations that prevented them from attending the exam (c). Around 3.09% of students missing the primary outcome did not attend the exam without providing a valid reason (e). In total, according to the administrative records, 16% of the randomisation sample did not sit the post-test GCSE exam.

- **Exam deferral.** 1% of attrition students deferred the exam.

- **Poor attendance.** Another 1% of attrition students were withdrawn from the exam or course due to low attendance records during the course.

- **Switching to vocational qualifications.** Less than 1% of students left the college or course to do apprenticeships or traineeships, or transferred to Functional Skills Maths.

- **2018 Winter Term resits.** A significant number of students took the GCSE resit exams in November 2018, which are available to post-16 students, and 350 passed the exam. These students did not continue with the course[24] so they did not attend Maths-for-Life lessons or the June 2019 resit exams, and so, they have been excluded from the analysis.

- **Leaving the college or course.** A substantial number of students left the college or withdrew from the Maths-for-Life course before the GCSE exams. Some of the reasons included moving out, switching to a class with a teacher that was not in the trial, leaving the course because it was too difficult, or behavioural issues.

[24] With the exception of 27 students who passed the November 2018 exam but also sat the Summer Term 2019 exam to increase their grade.

- **Missing data from settings.** Six settings (four in intervention and two in control) did not provide the raw exam scores, or provided the outcome data incorrectly, so the transformation to UMS scores could not be done.

- **Attrition at the setting level.** Two settings (one from each trial arm) left the trial after randomisation, so students are missing UMS and mathematical self-efficacy survey data. We did have their GCSE Maths pass data, as that outcome was collected directly from the NPD and ILR databases, which were included in the ITT analyses.

- **Unknown reasons.** Many settings did not provide justification for 38.20% of missing data.

In addition to the setting-level attrition and exclusion criteria (passing the November 2018 resit exam) presented in Table 11, 766 more students (13% of the randomised sample) were excluded from the primary analysis because they were missing at least one of the pre-treatment covariates. Missing covariate data affected 379 students in the treatment group and 387 in the control group.

Apart from attrition and missing data problems at the setting level, as well as exceptional situations such as family or work obligations or illness, most other reasons for attrition are likely to be correlated with mathematical attainment and student self-efficacy (though this is speculation). This suggests that lower-performing or less confident students may have been more likely to miss the exam or withdraw from the course. If these patterns of attrition disproportionately affected control students, it could introduce bias into the primary outcome analysis.[25] Despite this concern, attrition rates were similar between the two groups, with 47% of treatment group students and 49% of control group students missing UMS data. Furthermore, baseline characteristics of students in treatment and control groups (presented in the section 'Student and setting characteristics' below) do not present strong evidence of selection bias.

Lastly, aside from participant attrition and exclusion, failure to link to administrative records and missing pre-treatment covariates was another reason for missing observations in the primary outcome analysis. The DfE successfully matched 97.8% of student records to the NPD using personal identifiers (e.g. name, date of birth, and unique learner numbers). However, 128 students (2.2%) could not be linked to their NPD records and so were missing covariate data and GCSE pass rates, resulting in their exclusion. We present the rates of missing covariates in a later section in 'Missing data of pre-treatment covariates' below.

Secondary analysis on GCSE pass rate

Table 12: Student-level attrition from the trial (secondary outcome: GCSE pass rate)

		Intervention	Control	Total
No. of students	Randomised	3,071	2,735	5,806
	Analysed	2,357	2,015	4,372
Student attrition (from randomisation to analysis)	Number	714	720	1,434
	Percentage	23.25%	26.33%	24.70%

There were no observations with missing GCSE pass rate by construction, as we coded as 0 all participants in the sample who did not pass the GCSE exam, including those who did not sit the GCSE Maths exam in Summer Term 2019. The attrition can be explained by not meeting the eligibility criteria for the trial (296 students were excluded from this analysis because they had passed the November 2018 exam) and missing covariates. The difference in attrition between treatment and control was of 3 percentage points (compared to 2 percentage points in the primary outcome), where control students had higher rates of attrition.

[25] This is not the case for the students who passed the November 2018 resits, as they never participated in any Maths-for-Life lessons, so their attrition should be uncorrelated with their treatment assignment.

Given the way this outcome was coded, we were concerned that attrition in the GCSE pass rate would not capture differences in exam sitting rates between treatment and control students that could bias treatment effects. Specifically, if the intervention influenced participation in the exam, this could affect the GCSE resit results not only by improving students' exam performance but also by boosting their confidence to attend the exam.
Additionally, we wanted to check the actual exam participation rates to assess whether our initial assumptions about participant attrition at the design stage were accurate, as post-16 GCSE resit exam sitting rates tend to be low (Thomson, 2025). This provides a clearer measure of participant attrition rather than data collection issues in our primary outcome and is also of policy interest, as increasing exam participation could be a relevant intervention outcome.

To investigate this, we started by computing how many students in our trial sat the GCSE exam by checking the DfE administrative records. After adding those that either had a GCSE administrative record or a UMS record from the schools,[26] we calculate that only 4,711 students in the trial sat the GCSE resit in Summer Term 2019, which is 81.1% of all randomised students, close to the initial assumption of 20% attrition.

We then computed the number of observations that had a record for the exam by trial arm (Table 13) to explore any significant differences. Around 84.3% of intervention students sat the exam, compared to 77.6% of control students. An exploratory analysis[27] confirmed that control students were 6.2 percentage points more likely to miss the exam (21.6% in treatment vs 27.8% in control), after keeping baseline covariates constant (FSM status, setting type, and baseline attainment in Key Stage 2 and Key Stage 4). This represents a 22% reduction in the likelihood of missing the exam for the treatment group compared to the control group, a statistically significant difference at the 5% level, supporting the hypothesis that the intervention positively influenced participation in the GCSE resit exam. One potential explanation, aligned with the Theory of Change, is that the programme may have increased students' self-confidence, and improved their attitude towards, and relationship with, maths, which encouraged greater exam participation.

Table 13: Students that sat the GCSE Maths Summer Term exam in 2019, by trial arm

	Intervention (n/N)	Control (n/N)	Total (n/N)
No. of students	2,588 / 3,071	2,123 / 2,735	4,711 / 5,806
Overall randomised	84.3%	77.6%	81.1%

Secondary analysis on mathematical self-efficacy

Table 14: Student-level attrition from the trial (secondary outcome: mathematical self-efficacy)

		Intervention	Control	Total
No. of students	Randomised	3,071	2,735	5,806
	Analysed	564	533	1,097
Student attrition (from randomisation to analysis)	Number	2,507	2,202	4,709
	Percentage (analysed / randomised)	81.63%	80.51%	81.10%

[26] Around 4,607 learners (4,564 which had not passed the exam in November 2018) could be matched to a 2019 GCSE Maths Summer Term exam. This means only 84% of the 5,456 students eligible for analysis sat the post-test exam. Also, 104 students that had a UMS record did not have a GCSE record in the NPD. Of which, 78 out of those 104, could not be matched to the NPD or ILR datasets; four students were merged to the NPD but not the ILR.
[27] We used a logistic regression model. We regressed a binary variable indicating whether the student had missed the exam or not, on the treatment indicator and the pre-treatment covariates: FSM status; setting type; and baseline attainment in Key Stage 2 and Key Stage 4.

Overall, we had approximately 1,100 students complete the survey. The high-attrition rate (81%) affected both trial arms similarly. Fewer than ten complete cases[28] were excluded from the survey analysis because they had passed the GCSE resit exam in November 2018. The survey response rates were low, due to the resource-light approach to data collection.

Student and setting characteristics

Overall, baseline characteristics showed that intervention and control groups were similar, with small differences in FSM eligibility (slightly higher in the intervention group) and a 7-percentage-point gender imbalance favouring males in the intervention arm. While attrition increased the proportion of students eligible for FSM and altered gender composition, these changes did not create significant imbalances in baseline attainment, though male students were more likely to be missing in the final analysis.

Table 15 summarises the baseline student-level characteristics of intervention and control students as randomised. Overall, the intervention and control students were similar, but both differed in certain ways compared to the national-level figures.

Table 15: Baseline characteristics of students as randomised

		Intervention group		Control group		
Student level (categorical)		n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Eligible for FSM	32.2%[a]	996/3,071 (76)	32.4%	800/2,735 (86)	29.3%	
Gender: Male Female	47.3% 52.7%[b]	1,545/3,071 1,450/3,071 (76)	50.3% 47.2% (2.5%)	1,180/2,735 1,469/2,735 (86)	43.1% 53.7% (3.1%)	
Student level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size
Attainment in Key Stage 2 Maths	Not available[c]	2,568 (503)	51.130 (17.807)	2,264 (471)	50.828 (17.570)	0.017
Baseline Key Stage 4 GCSE Maths grade	4.6[d]	2,805 (266)	2.510 (0.748)	2,486 (249)	2.489 (0.754)	0.028
Age	Not available[e]	3,071 (0)	17.577 (3.391)	2,560 (175)	17.429 (3.181)	0.045

[a]This percentage represents Level 2 students enrolled in a Technical Certificate at state-funded schools/colleges who were classified as disadvantaged in 2018. For Level 2 vocational qualifications, this was 33.3%. Source: DfE 16-18 attainment data (DfE, 2019). Available at: https://assets.publishing.service.gov.uk/media/5c48878640f0b61704aec530/2018_revised_A_level_and_other_16-18_results_in_England.pdf (accessed 04 December 2024).
[b]Percentage of male and female students over all students aged 17 and over that sat the Summer Term GCSE Maths in 2021 in the UK. Data obtained from the Joint Council for Qualifications (2021). Available at: www.jcq.org.uk/wp-content/uploads/2021/08/GCSE-Full-Course-Results-Summer-2021.pdf (access 04 December 2024).
[c]Only scaled scores for national averages are published by the DfE. These are not comparable to the raw marks used in our analysis.
[d]This is the average GCSE Maths grade for the 2022/2023 academic year, the first year that the DfE published this metric. In previous years, the DfE only published the cumulative percentages of students that achieved a certain grade. Data obtained from Key Stage 4 performance report by the DfE (DfE, 2025). Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-4-performance (accessed 04 December 2024).
[e]The average age of GCSE resit students is not published by the DfE.

[28] The count has been censored to comply with ONS SRS clearance requirements.

The intervention group had a slightly bigger proportion of FSM-eligible students (32.4%) compared to the control group's (29.3%). Both were close to the national average of students enrolled in post-16 Level 2 qualifications, which is 32.2%. The gender split showed more difference between trial arms: the intervention students had a bigger proportion of male students (50.3%) compared to the control students (43.1%). After attrition (Table 12), the gender composition shifted, increasing the difference between the treatment and control groups. The intervention group became more gender-balanced, while the percentage of female students in the control group rose from 53.7% to 58.5%. This could influence the results, as female students generally outperform males in GCSEs (DfE, 2024).

The 7-percentage-point gender imbalance could influence results due to the gender attainment gap, but baseline attainment differences at Key Stage 2 and Key Stage 4 Maths between the groups were minimal, with Hedges' g values below 0.05. This suggests the gender split did not affect the balance in baseline attainment. Histograms of Key Stage 2 scores (Appendix P) show similar normal distributions across trial arms, both before and after attrition. However, post-attrition, the treatment group had lower kurtosis, indicating a flatter distribution with fewer scores concentrated around the average compared to the control group, suggesting greater variability in baseline attainment compared to the control. This could affect the results by making the treatment group's performance less predictable and potentially introducing noise into the analysis. In contrast, Key Stage 4 GCSE attainment showed highly similar distributions between trial arms (Appendix P), both before and after attrition, reducing the likelihood that outcome differences are influenced by Key Stage 4 baseline variations, thus supporting the validity of the group comparisons.

As an exploratory analysis, we also checked balance in age, because the age range of the sample was very wide and included students who were older than 18 years old. Older post-16 resit takers tend to have higher scores (see Table 20), and the availability of administrative data records also varies. At randomisation, the average age for intervention participants was 17.58 years old and 17.43 years old for control students; a very small difference (a Hedges' g of less than 0.05).

Table 16 presents the analogous balance characteristics for the groups in the primary outcome analysis. After attrition, FSM eligibility in the intervention arm rose to 36%, 6 percentage points higher than the control group, and higher than the national average for similar Level 2 post-16 students. While this increased disadvantage could affect results, it did not lead to a substantial imbalance in Key Stage 4 attainment, with differences remaining below 0.05 SD. At Key Stage 2, the imbalance increased, but the intervention group still had slightly higher attainment. Thus, the higher percentage of FSM students in the intervention group did not translate into lower baseline attainment compared to the control group.

Table 16: Baseline characteristics of students as analysed

		Intervention group		Control group		
Student level (categorical)		n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Eligible for FSM	32.2%	591/1,631 (0)	36.2%	426/1,401 (0)	30.4%	
Gender: Male Female	47.3% 52.7%	813/1,631 818/1,631 (0)	49.9% 50.1%	582/1,401 819/1,401 (0)	41.5% 58.5%	
Student level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size
Attainment in Key Stage 2 Maths	Not available[a]	1,631 (0)	49.986 (17.329)	1,401 (0)	49.016 (17.064)	0.056
Baseline Key Stage 4 GCSE Maths grade	4.6[b]	1,631 (0)	2.541 (0.684)	1,401 (0)	2.560 (0.651)	-0.027
Age	Not available	1,631 (0)	17.018 (0.975)	1,323 (78)	16.935 (0.923)	0.087

[a]*Only scaled scores for national averages are published by the DfE. These are not comparable to the raw marks used in our analysis.*
[b]*See footnote 32.*

Table 17 summarises the baseline setting-level characteristics at randomisation, and how they compare to national averages (see Table 18 for school characteristics as analysed). In sum, the sample of education settings was well-balanced across treatment and control groups in terms of type, urban location, and student composition, and this balance was maintained after attrition. However, the sample was moderately less representative of the national average, with an overrepresentation of urban settings, further education colleges, and the Office for Standards in Education, Children's Services and Skills (Ofsted) higher-rated institutions, which may limit the generalisability of the results to rural and lower-quality post-16 settings.

The trial primarily involved further education colleges, reflecting their majority share of post-16 education (54%), while training providers were under-represented compared to their 21% national market share. The distribution of setting types was well-balanced between treatment and control groups, and this balance was preserved after attrition.

Most settings were in urban areas (94%) across both groups, with the treatment group becoming entirely urban after attrition (100% vs 94%). This contrasts with the national distribution, where 25% of settings are in rural areas, limiting the applicability of the trial's results to rural settings.

At randomisation, control settings were more likely to be rated as 'Outstanding' (22% vs 10%), while intervention settings had a higher proportion rated as 'Good' (76% vs 64%), with similar weights of lower-rated settings in both arms. These differences in Ofsted ratings, favouring control settings, persisted after attrition. The proportion of settings with at least Good or Outstanding ratings in both trial arms was higher compared to the 2018 national average for post-16 education (86% vs 69%), which may affect the trial's replicability on a national scale.

The average size of the educational settings was similar across both trial arms before and after attrition (4,777 students in intervention settings and 5,085 students in control settings).

Regarding student composition, intervention settings had lower rates of FSM students (23.4%) compared to the control settings (27.0%), and this difference did not change after attrition. Both rates were close to the national average of disadvantaged students in post-16 settings, which sits also at 27.7%.

Table 17: Baseline characteristics of settings as randomised

Setting level (categorical)	National-level mean	Intervention group		Control group	
		n/N (missing)	Percentage (%)[a]	n/N (missing)	Count (%)
Setting type:					
Further education college	247 (54%)	33/50	66%	32/50	64%
School	1,160[b] (N/A)	9/50	18%	10/50	20%
Sixth-form college	94 (5%)	5/50	10%	5/50	10%
Training provider	546 (21%)[c]	3/50	6%	3/50	6%
Ofsted rating:					
Outstanding	2%	5	10%	11	22%
Good	67%	38	76%	32	64%
Requires improvement	28%	5	10%	4	8%
Inadequate	3%[d]	2	4%	2 (1)	4% (2%)
Location:					
Urban	75%	47	94%	47/50	94%
Rural	25%	3	6%	3/50	6%

Setting level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)
Proportion of FSM-eligible students[e]	27.7%[f]	11/50 (39)	23.4% (15.5%)	12/50 (38)	27.0% (18.8%)
No. of students per setting	Secondary schools: 948[g] further education colleges: 6,749 sixth-form colleges: 1,926 training providers: 527[h]	47/50 (3)	4,777 (4,229)	46/50 (4)	5,085 (5,767)

[a]*Percentages were calculated over all observations at randomisation and analysis stages, including non-missing observations.*
[b]*Number of schools in England offering a sixth form for 16–18-year-olds in 2014 (Hupkau and Ventura, 2017). Schools with a sixth form are not included in the percentages of total post-16 learners.*
[c]*Number of publicly funded providers that were further education colleges, sixth-form colleges, and training providers in 2014 (Hupkau and Ventura, 2017). The percentage indicates their student shares over the total number of learners of post-16 education. The 20% remaining learners in England attended other publicly funded providers. That percentage does not include schools.*
[d]*Obtained from DfE statistics on the Ofsted inspections carried out in 2017/2018. Available at: www.gov.uk/government/statistics/further-education-and-skills-inspections-and-outcomes-as-at-31-august-2018/further-education-and-skills-inspections-and-outcomes-as-at-31-august-2018-main-findings#:~:text=As%20at%2031%20August%202018%2C%20the%20proportion%20of%20general%20FE,good%20at%20inspection%20this%20year.*
[e]*Of 91 settings in total, we were only able to access data on FSM eligibility for 23 settings. Of which, 19 of these were schools, three were sixth-form colleges, and one was a further education college.*
[f]*Percentage of disadvantaged state-funded school students who were at the end of Key Stage 4 by the end of 2015/2016 academic year. Source: DfE 16-18 attainment data (DfE, 2019). Available at:* https://assets.publishing.service.gov.uk/media/5c48878640f0b61704aec530/2018_revised_A_level_and_other_16-18_results_in_England.pdf *(accessed 04 December 2024).*
[g]*Average size of state-funded secondary schools in 2018 (DfE, 2018). This does not include post-16 education providers.*
[h]*Data on student size for further education colleges, sixth-form colleges, and training providers are from 2014 (Hupkau and Ventura, 2017).*
N/A=not applicable.

Table 18: Baseline characteristics of settings as analysed

Setting level (categorical)	National-level mean	Intervention group		Control group	
		n/N (missing)	Count (%)	n/N (missing)	Count (%)
Setting type:					
Further education college	247 (54%)	30/45	66.6%	29/46	63%
School	N/A	9/45	20%	10/46	21.7%
Sixth-form college	94 (5%)	< 5[b]/45	< 11%	< 5 /46	< 11%
Training provider	546 (21%)[a]	< 5 /45	< 11%	< 5 /46	< 11%
Ofsted rating:					
Outstanding	2%	4	9%	10	22%
Good	67%	35	78%	30	65%
Requires improvement	28%	5	11%	4	9%
Inadequate	3%	1	2%	2	4%
Location:					
Urban	75%	45	100%	43	94%
Rural	24%[c]	0	0%	3	6%
School level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)
Proportion of FSM-eligible learners[d]	18.5%[e]	11/45 (34)	23.4% (15.5%)	12/46 (34)	27.0% (18.8%)
No. of learners per college/school	Secondary schools: 948 further education colleges: 6,749 sixth-form colleges: 1,926	42/45 (3)	4,728 (3,873)	44/46 (2)	5,005 (5,869)

	training providers: 527				

[a]Percentage of male and female students over all students aged 17 and over that sat the Summer Term GCSE Maths in 2021 in the UK. Data obtained from the Joint Council for Qualifications (2021). Available at: www.jcq.org.uk/wp-content/uploads/2021/08/GCSE-Full-Course-Results-Summer-2021.pdf (access 04 December 2024).
[b]Please note ONS statistical disclosure controls for DfE data prevent the reporting of cell counts lower than three at school level and lower than ten at individual (e.g. student) level, at the time of writing.
[c]Obtained from DfE statistics for 2018. Available at https://explore-education-statistics.service.gov.uk/data-tables.
[d]This is the average GCSE Maths grade for the 2022/2023 academic year, the first year that the DfE published this metric. In previous years, the DfE only published the cumulative percentages of students that achieved a certain grade. Data obtained from Key Stage 4 performance report by the DfE (DfE, 2025). Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-4-performance (accessed 01 March 2025).
[e]The average age of GCSE resit students is not published by the DfE.
N/A=not applicable.

Outcomes and analysis

Primary analysis

The primary outcome measure was the UMS score, which standardises the raw scores of the GCSE Maths exam across different examination boards. This process smooths out variations in difficulty between exam boards and exam years to allow for comparisons across settings. The UMS score of the GCSE Maths exam was selected as the primary outcome measure because it provided the most direct assessment of changes in student performance at GCSE resit exams and mathematical attainment, which was the primary focus of the intervention.

UMS scores can range from 0 to 100 for Higher tier exams (which allow for scores up to grade 9) and 0 to 50 for Foundation tier exams (an easier exam, capped at grade 5). Around 98.4% of students (2,983) sat the Foundation tier and thus, were capped at a score of 50. As a result, the average treatment effect for the full sample could, in theory, be diluted due to the ceiling effect in the Foundation tier, potentially underestimating the intervention's impact on lower-ability students.

Figure 3 shows the distribution of UMS scores in the primary analysis sample. The distribution is approximately normal but left-skewed. However, even though part of the sample was capped at 50, there is not a significant amount of bunching around that value. Instead, most students scored grades 3 (equivalent to a UMS of 30 to 39) or grade 4 (equivalent to a UMS of 40 to 49). Therefore, the ceiling effect is not a concern as it does not appear to heavily distort the distribution.

Since the vast majority of the sample was a Foundation tier, the range of the potential treatment effects for the full sample will be approximately -51 to 51, reflecting a weighted average of the two caps (2% at [-100, 100] and 98% at [-50, 50]).

Figure 3: Distribution of UMS scores

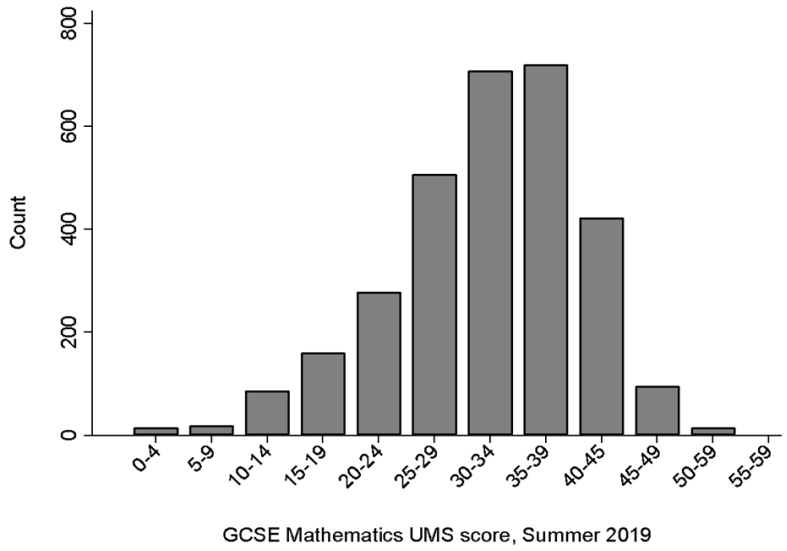


Table 19 presents the results of the analysis for the UMS score outcome. The unadjusted mean for the UMS score in the intervention group is 31.71 and 32.60 in the control group. After adjusting for covariates in the analysis model, the mean

difference between the two groups decreased from -0.888 to -0.809, corresponding to a Hedges' g effect size of -0.095 (equivalent to two months less progress). While this is the best estimate of the intervention's impact, the 95% CI suggests the true effect could range from significantly negative (three months' less progress) to negligible (zero months of progress). This indicates substantial uncertainty in the estimate, but the intervention is likely to have had little to no effect or a negative effect on the students' GCSE Maths exam scores. The wide CI reflects the study's limited statistical power to detect anything other than a somewhat large effect (relative to the average intervention evaluated by the EEF), due to the small sample size.

Table 19: Primary outcome analysis results

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (95% CI)	P-value
GCSE Maths UMS score	1,631 (1,440)	31.710 (31.271 – 32.148)	1,401 (1,334)	32.598 (32.182 – 33.014)	3,032	-0.095 (-0.203 – 0.013)	0.085

Secondary analysis

The two secondary outcome measures were the GCSE Maths pass rate, a binary indicator equal to 1 if the student passed the GCSE Maths exam with a grade 4 or more and 0 otherwise, and mathematical self-efficacy, measured by the Teleprism survey.

GCSE Maths pass rate

Figure 4 shows the distribution of GCSE pass rates. According to Figure 4, 614 out of 4,372 students (14%) passed the GCSE Maths exam at the end of the intervention with a grade 4 or higher. For reference, the national average in 2019 was 21.5% for 17-year-olds (Table 20).

Figure 4: Percentage of analysed students that passed the GCSE Maths exam with a grade 4 or higher

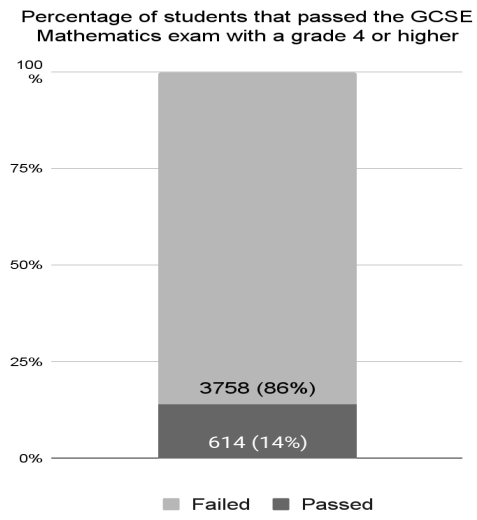


Table 20: Post-16 outcomes in maths, percentage of students with grade 4 and above, England average

Year	2018	2019
17-year-olds	22.3	21.5
18-year-olds	14.3	13.4
+19-year-olds	29.7	28.1

Source: DfE from Ofqual data.[29]

Table 21 presents the results of the analysis. The unadjusted rate of the intervention group is 14.6% and the control group is 13.4%. For the GCSE Maths pass rate, the estimated effect for the binary outcome could take theoretical values between

[29] Obtained from: www.gov.uk/government/news/guide-to-gcse-results-for-england-2019.

-1 and 1. After adjusting for covariates in the analysis model, the mean difference between the two groups is 0.05, or 0.5 percentage points, which translates into a Hedges' g effect size of 0.014. This represents a negligible effect, equivalent to zero months of progress.

While this is the best estimate, the 95% CI indicates that the true effect could range from negative (one months' less progress) to moderately positive (two months of additional progress). This wide range reflects significant uncertainty, and the results do not allow for a reliable conclusion about the direction or magnitude of the effect on the students' likelihood to pass the resit exam. The very wide CI is a result of the study being underpowered.

Table 21: Secondary outcome analysis, by GCSE pass rates

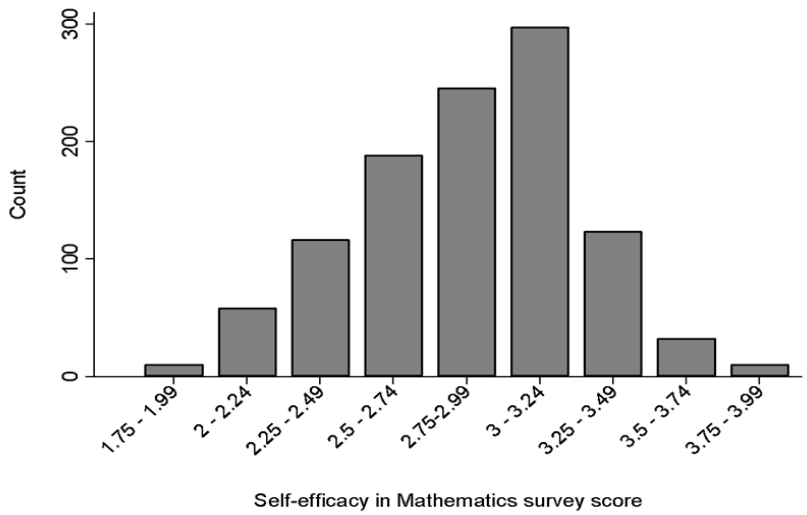
	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (95% CI)	P-value
GCSE pass rate	2,357 (714)	0.146 (0.132 – 0.160)	2,015 (720)	0.134 (0.119 – 0.149)	4,372	0.014 (-0.083 – 0.111)	0.771

Mathematical self-efficacy

This section presents the results of Maths-for-Life on the students' mathematical self-efficacy, as measured via section E of the Teleprism survey. The outcome measure was the average score of the Likert questions, which could take values from 1 (less confident) to 4 (more confident), so the estimated effect could take theoretical values between -3 and 3. Effects are also presented as Hedges' g to make it easier to compare between outcomes and with other studies.

Figure 5 shows a histogram of the survey scores. The distribution of scores is slightly left-skewed. The number of participants scoring at or near the maximum of 4 is limited, so we do not have any concerns about floor or ceiling effects in the analysis. However, given the high degree of attrition (81%), the results have a high risk of bias from unobservable student characteristics, for example, due to self-selection from the most engaged students into answering the survey.

Figure 5: Histogram of mathematical self-efficacy survey scores[a]



[a]*The histogram excludes counts below 1.75 and at 4 as they were under 10, to be compliant with ONS policies.*

Table 22 shows the results of the analysis on the mathematical self-efficacy survey score. The unadjusted mean in the intervention group is 2.876 (out of 4) and in the control group is 2.855. After adjusting for covariates in the analysis model, the mean difference between the two groups is reduced from 0.021 to 0.015. The Hedges' g effect size for the difference between groups is 0.037, equivalent to a small positive effect on the students' confidence when solving mathematical problems.

This is our best estimate of the effect but, at the 95% confidence level, the results are compatible with effects that range from -0.121, a moderate negative effect, to a large positive effect of 0.194 SD. There is therefore, a large amount of uncertainty in the estimate of the intervention's impact on the students' self-efficacy in addition to the high risk of bias, which are a result of the low survey response rates across settings.

Table 22: Secondary outcome analysis, by mathematical self-efficacy

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (95% CI)	P-value
Mathematical self-efficacy score	564 (2,507)	2.876 (2.842 – 2.911)	533 (2,202)	2.855 (2.822 – 2.889)	1,097	0.037 (-0.121 – 0.194)	0.645

Missing data analysis

Missing data of pre-treatment covariates

In the Statistical Analysis Plan for this trial, we outlined a missing data strategy (Nolan and Taylor, 2020). We noted that this strategy would be implemented if more than 5% of data for a covariate in the primary analysis sample were missing. As can be shown in Table 23, this was the case for the two baseline attainment measures. Key Stage 2 Maths attainment was missing for 18.17% of the primary outcome sample and Key Stage 4 Maths grade, for 8.4% of the sample.

Table 23: Completeness of data for pre-treatment covariates in the primary outcome sample (over all eligible participants with non-missing primary outcome)

Covariate	Intervention (n; %)	Control (n; %)	Total (n; %)
Setting type	2,018 (100%)	1,780 (100%)	3,798 (100%)
Eligible for FSM	1,977 (97.97%)	1,730 (97.19%)	3,707 (97.60%)
Attainment in Key Stage 2 Maths	1,671 (82.80%)	1,437 (80.73%)	3,108 (81.83%)
Baseline Key Stage 4 GCSE Maths grade	1,861 (92.22%)	1,618 (90.90%)	3,479 (91.60%)

Table 24 below summarises the number of complete cases, missing observations due to missing pre-treatment covariates, and missing observations due to trial eligibility for each of the three outcomes. Around 20.5% of all eligible participants with primary outcome data were dropped from the analysis for missing at least one pre-treatment covariate. The percentage of observations dropped for the same reason was very similar for the two secondary outcomes; 20.2% for GCSE pass rate, and 21.6% for the mathematical self-efficacy survey.

Table 24: Number of complete case observations and missing pre-treatment covariates

Outcome		Intervention	Control	Total
Primary outcome (UMS)	Complete cases	1,631	1,401	3,032
	Eligible participants with outcome data	2,018	1,780	3,798
	Eligible participants with outcome data, missing pre-treatment covariates	387	379	768

Outcome		Intervention	Control	Total
	Complete cases / eligible participants with outcome data	80.1%	78.7%	79.5%
	With outcome data, excluded from the analysis for passing the GCSE in November 2018	<10	20	Between 20 and 30
GCSE Maths pass rate	Complete cases	2,357	2,015	4,356
	Eligible participants with outcome data	2,920	2,536	5,456
	Eligible participants with outcome data, missing pre-treatment covariates	563	521	1,099
	Complete cases / learners with outcome data	80.7%	79.5%	79.8%
	With outcome data, excluded from the analysis for passing the GCSE in November 2018	151	199	350
Mathematics self-efficacy score	Complete cases	564	533	1,102
	Eligible participants with outcome data	725	680	1,405
	With outcome data, missing pre-treatment covariates	161	147	308
	Complete cases / learners with outcome data	77.8%	78.4%	78.4%
	With outcome data, excluded from the analysis for passing the GCSE in November 2018	<10	<10	<10

Missing data analysis was conducted for the Key Stage 2 and Key Stage 4 attainment covariates to gather suggestive evidence as to whether the absence of data followed a random pattern (missing at random [MAR]). We modelled the missingness of each variable using two logistic regressions as specified in the 'Missing data analysis' section below.

For Key Stage 2 attainment, we regressed the variable missingness on treatment assignment, FSM eligibility, Key Stage 4 attainment, and setting type. We also regressed it on these covariates and age as an exploratory analysis, which was not included in the Statistical Analysis Plan (Nolan and Taylor, 2020), because there were students in the sample that were older than 18 years old, and the higher missingness of Key Stage 2 could be related to senior students not having Key Stage 2 records in the NPD, as Key Stage 2 SATs were introduced in 1995. This analysis revealed that older students, those from sixth-form colleges, students not eligible for FSM, and those with lower Key Stage 4 attainment were more likely to have missing Key Stage 2 Maths attainment, and hence to be missing from the primary analysis. These predictors were statistically significant at the 5% level. Lastly, not having attended primary school in England could be another reason for the high rates of missing Key Stage 2 records, but we did not have available data to check whether that was the case.

We repeated the analysis for Key Stage 4 GCSE Maths attainment. A higher age was also a predictor for this covariate's missingness, as well as having a higher Key Stage 2 attainment. These were the only predictors of missingness that were statistically significant at the 5% level. Lastly, we did an equivalent analysis on the missingness of both Key Stage 2 and Key Stage 4, which showed that older students and those not eligible for FSM were statistically more likely to be missing both baseline attainment covariates. None of the logistic regressions showed any indication that treatment assignment was correlated with covariate missingness. Given that at the analysis stage the sample is balanced on all observables, this suggests that any unobservable characteristic that leads to missingness is likely to be distributed evenly between treatment and control groups, and hence, not cause a bias in the treatment estimate.

These results indicate that part of the covariate missingness was conditional upon other variables in the model. Whether this might increase or decrease the estimated treatment effect is unclear because the missingness is related to both higher and lower-performing students. Older students, those with higher Key Stage 2 attainment, and non-FSM-eligible students, who tend to perform better, were more likely to have missing data, which could lead to an overestimation of the treatment effect. On the other hand, students with lower Key Stage 4 attainment, who typically perform worse, were also more likely

to be missing, which could lead to an underestimation of the treatment effect, especially if the treatment is more effective for lower performers. Without any evidence of whether the treatment is more effective for higher or lower-performing students, the overall impact on the treatment effect is uncertain.

In this analysis, we deviated from the Statistical Analysis Plan (Nolan and Taylor, 2020) and opted not to perform multiple imputation for MAR covariate data. This decision was based on the high levels of attrition observed for both the primary and secondary outcomes, which meant that conducting multiple imputation for covariates would contribute little to the interpretation of the results. (Multiple imputation would increase the primary analysis sample from 52% to, at most, 65% of the sample at randomisation.)

Sensitivity analysis: Estimation of the model excluding covariates and using the missing indicator method

We carried out sensitivity analyses by: i) excluding all covariates from our regression specification for the primary outcome analysis; ii) excluding only baseline attainment covariates; and iii) excluding only Key Stage 2. Additionally, as an exploratory analysis, we estimated the regression using the missing indicator and null imputation.

The Hedges' g coefficients in the sensitivity analyses are very similar to the primary analysis result in all cases: the effects are negative and very small; and the 95% CIs are very wide and all of them contain the zero. The consistency of the effect estimates across sensitivity analyses indicates that the inclusion or exclusion of specific covariates does not substantially alter the conclusions, as effect estimates remain small and negative across all specifications. The wide CIs reinforce the high level of uncertainty around the estimated effects, suggesting that any true impact of the intervention, if present, is likely to be small and could be negative.

Table 25: Sensitivity analysis results

Outcome	Model	Total n (intervention; control)	Hedges' g (95% CI)	P-value	Primary analysis Hedges' g (95% CI)
GCSE Maths UMS score	No covariates	3,798 (2,018; 1,780)	-0.085 (-2.556 – 0.086)	0.328	-0.095 (-0.203 – 0.013)
GCSE Maths UMS score	All covariates except Key Stage 2 attainment	3,471 (1,856; 1,615)	-0.067 (-0.173 – 0.040)	0.217	-0.095 (-0.203 – 0.013)
GCSE Maths UMS score	All covariates except Key Stage 2 and Key Stage 4 attainment	3,707 (1,977; 1,730)	-0.082 (-0.242 – 0.078)	0.310	-0.095 (-0.203 – 0.013)
GCSE Maths UMS score	Missing indicator method[a]	3,707 (1,977; 1,730)	-0.091 (-0.203 – 0.021)	0.108	-0.095 (-0.203 – 0.013)

[a]*Imputing a missing indicator for Key Stage 2 and Key Stage 4 baseline attainment only (not FSM).*

Subgroup analyses

FSM subgroup analysis

This subgroup analysis examines whether the treatment effects differed between students eligible for FSM and those who were not. Following the EEF guidelines (EEF, 2022), these analyses were conducted in two ways: i) by including an interaction term in the model to test for differences in effects between the groups, and ii) by running separate models for each subgroup (FSM-eligible and non-FSM-eligible students).

The interaction term estimates the difference in treatment effects between FSM and non-FSM students, after adjusting for other variables. As shown in Table 26, the interaction term is small and positive, indicating that the treatment effect was slightly less negative for FSM students. The Hedges' g is 0.017, which is only 18% of the ITT effect from the primary analysis and corresponds to less than a month of additional progress between FSM and non-FSM students. However, the results are

highly uncertain, as the CI is wider than that in the primary analysis, ranging from -2 months to +2 months of additional progress.

When estimating the effects separately for FSM and non-FSM students, both groups saw a similar impact of -1 month of progress. The 95% CI for non-FSM students was [-3 months, 0 months], and for FSM students, it was [-3 months, +1 month].

While the intervention appeared to have a slightly less negative impact on FSM students than non-FSM students, both groups still experienced a negative effect overall (though the CIs are very broad again). The small positive interaction term suggests the programme may have been somewhat less ineffective for FSM students, but it did not improve their attainment. However, the level of uncertainty around the estimates makes the results inconclusive. The study was likely underpowered to detect these heterogeneous effects; it had a sample of FSM students to detect an MDES of 0.30 for this subgroup. This limitation in statistical power means that the observed differences between the two groups should be interpreted with caution, and the differences are also extremely small.

Table 26: Subgroup analysis for FSM-eligible students

Outcome	Model	Total n (intervention; control)	Coefficient (SE)	Hedges' g (95% CI)	P-value	Primary analysis Hedges' g (95% CI)
UMS score	Interaction effect	3,032 (1,631; 1,401)	0.149 (0.614)	0.017 (-0.124 – 0.159)	0.809	-0.095 (-0.203 – 0.013)
UMS score	Subgroup (non-FSM)	2,015 (1,040; 975)	-0.828 (0.445)	-0.098 (-0.202 – 0.007)	0.066	-0.095 (-0.203 – 0.013)
UMS score	Subgroup (FSM)	1,017 (591; 426)	-0.637 (0.668)	-0.074 (-0.228 – 0.080)	0.343	-0.095 (-0.203 – 0.013)

Subgroup analysis by setting type

The intervention was delivered in four different types of settings: schools; sixth-form colleges; further education colleges; and training providers. The four types of post-16 education settings differ in terms of focus, with schools and sixth-form colleges offering more academic courses, further education colleges providing a broader range of academic and vocational options, and training providers specialising in vocational training. Additionally, the size of their student bodies varies substantially, leading to further education colleges having much larger cluster sizes in this trial.

We conducted a subgroup analysis by type of setting to see if the treatment effect on the primary outcome varied by type of education provider. This is done by including in the primary outcome model an interaction term with the categorical variable of setting type and the treatment.

Table 27 presents the marginal treatment effects for each type of setting, revealing notable differences in how the intervention worked across settings. In further education colleges, the largest subgroup, Maths-for-Life had a moderate negative impact of two months less progress on students' scores (95% CI: [-3 months, 0 months]). Similarly, students in sixth-form colleges experience a negative impact of two months less progress, with a wide range of possible effects from -7 months to +3 months. In schools, the estimated impact was small and positive, equivalent to less than a month of progress (95% CI: [-3 months, +4 months]). For training providers, the estimated treatment effect was five months of additional progress, but this estimate is less reliable due to the small sample size, with a wide range of possible effects from -2 months to +11 months of additional progress.

These results suggest that the Maths-for-Life intervention may have had varying effects depending on the type of provider, but the evidence is not strong enough to confirm that the treatment worked differently across settings. The small sample sizes mean the estimates are imprecise.

Table 27: Subgroup analysis on UMS score by setting type (further education college is the reference category)

Setting type	Total n (intervention; control)	Coefficient (SE)	Hedges' g (95% CI)	P-value	Primary analysis Hedges' g (95% CI)
Further education college	2,519 (1,406; 1,113)	-0.852 (0.508)	-0.100 (-0.218 – 0.018)	0.097	-0.095 (-0.203 – 0.013)
School	197 (124, 73)	0.323 (1.050)	0.041 (-0.225 – 0.308)	0.759	-0.095 (-0.203 – 0.013)
Sixth-form college	284 (148; 136)	-1.357 (1.781)	-0.155 (-0.555 – 0.246)	0.445	-0.095 (-0.203 – 0.013)
Training provider	32 (N/A; N/A)[a]	3.206 (2.165)	0.370 (-0.126 – 0.865)	0.142	-0.095 (-0.203 – 0.013)

[a]*The counts in the intervention and control group cannot be disclosed as per the ONS disclosure rules because some are under ten.*

Additional analyses and robustness checks

Analysis in the presence of non-compliance

A setting was considered to be compliant if all teachers participating in the trial delivered at least three out of five Maths-for-Life lessons to their classes. Based on the collected data from the settings, 37 out of 50 settings assigned to the treatment group were considered as compliant.

Table 28 shows how learners were distributed across compliant and non-compliant settings. Around 69% of all students in the treatment arm were in compliant settings and therefore, likely to have received the minimum recommended dosage of the intervention.[30]

Table 28: Compliance frequencies and unadjusted mean and SD of the primary outcome

Sample	Control group	Students in compliant settings		Students in non-compliant settings	
	UMS mean (SD)	n (% of intervention group)	Unadjusted mean (SD)	n (% of intervention group)	Unadjusted mean (SD)
At randomisation	32.927 (9.086)	2,122 (69%)	32.095 (9.154)	949 (31%)	31.943 (10.855)
At CACE analysis	32.598 (7.938)	1,171 (72%)	32.038 (8.890)	460 (28%)	30.873 (9.320)

Compliance data were available for the full primary analysis sample (n=3,032). The first-stage regression produced an F-statistic of 20.93, well above the threshold of 10 commonly used to identify a strong instrument (Table 29). This confirms that treatment allocation is a reliable predictor of compliance, supporting the robustness of the instrumental variable design and the reliability of the CACE estimates.

Table 29 presents the results of the CACE analysis. The estimated difference in UMS scores between the intervention and control groups is -1.124, larger (in absolute magnitude) than the ITT estimate of -0.809. The corresponding Hedges' g for the CACE analysis is -0.134, equivalent to -2 months of progress (95% CI: [-4 months, 0 months]). While this remains our best estimate of the effect, the wide CI highlights considerable uncertainty, similar to the ITT analysis.

[30] We did not have attendance data from students, so this is our best guess of the students' dosage.

For comparison, the ITT analysis produced a Hedges' g of -0.095, equivalent to -1 month of progress (95% CI: [-3 months, 0 months]). The compliance analysis, which uses data on compliance at the teacher and setting level, provides additional evidence that the intervention likely had a negative effect on students' GCSE Maths scores. The broader CIs in both analyses reflect the study's limited statistical power, driven by a small sample size, which restricts its ability to detect anything other than a large effect. The CACE analysis, while slightly stronger, still points to a high level of uncertainty and does not significantly alter the overall interpretation of the results.

Table 29: CACE analysis

Model	Total n (intervention; control)	Coefficient (SE)	First stage F-statistic	Partial R-squared	Hedges' g (95% CI)	P-value	Primary analysis Hedges' g (95% CI)
Compliance analysis	3,032 (1,631; 1,401)	-1.124 (0.641)	F(7, 3024) = 20.93	0.540	-0.134 (-0.284 – 0.016)	0.080	-0.095 (-0.203 – 0.013)

Estimation of a multilevel model as a robustness check

As a robustness check, we estimate an HLM with the primary and secondary outcomes, which accounts for the hierarchical structure of the data—specifically, the nesting of students within classes and settings, by including random intercepts at the school and class level. While our primary analysis uses OLS with clustered standard errors at the setting level to handle the clustered nature of the trial, the HLM allows us to further test whether the hierarchical structure of classes within settings affects the results, providing an additional comparison to our main estimates.

For this analysis 952 out of 5,806 students were missing a class identifier (16.40% of all randomised students). However, most of them were also missing the primary outcome or pre-treatment covariates. As a result, only 0.36% of the sample for the primary analysis is missing the class identifier. This ensures that the sample used to estimate the HLM closely aligns with the primary analysis sample, allowing for a reliable comparison.

In the case of the robustness checks for the secondary outcomes, 10.20% (592 observations) of the sample for the GCSE pass rate is missing a class identifier and is therefore, excluded from this analysis. This results in a weaker overlap with the analysis using OLS, which should be considered when making comparisons. In contrast, the robustness check for the survey analysis provides a strong basis for comparison, as fewer than ten students are missing a class identifier in that dataset.

Additionally, 24 schools had only one class identifier and two schools did not send any class identifiers, so we cannot estimate distinct class and setting level errors for those schools.

Table 30: HLM models

Outcome	Total n (intervention; control)	Coefficient (SE)	Hedges' g (95% CI)	Post-estimation ICC: setting and class levels	P-value	OLS analysis: Hedges' g (95% CI)
UMS	3,021 (1,620; 1,401)	-0.676 (0.458)	-0.079 (-0.184 – 0.026)	0.031 (3.1%), 0.100 (10%)	0.140	-0.095 (-0.203 – 0.013)
GCSE Maths pass rate	3,780 (2,023; 1,757)	-0.039 (0.156)	-0.108 (-0.951 – 0.735)	0.063 (6.3%), 0.07 (7%)	0.801	0.014 (-0.083 – 0.111)
Mathematical self-efficacy score	1,092 (560; 532)	0.018 (0.032)	0.045 (-0.110 – 0.200)	0.042 (4.3%), 0.042 (4.2%)	0.568	0.037 (-0.121 – 0.194)

Table 30 presents the estimated effect sizes of the HLMs. For the primary outcome, the HLM estimate aligns closely with the OLS model, indicating a regression equivalent to approximately one month of progress. The 95% CI spans from -2

months to no progress, identical to the interval observed in the OLS model. This consistency suggests that the OLS approach captured the treatment effect accurately.

A similar pattern is observed for the mathematical self-efficacy outcome. The HLM estimates a small, positive effect size of 0.045 in Hedges' g (95% CI: [-0.110, 0.200]), which is very close in size and CI to the OLS results (Table 30).

However, the results for the GCSE pass rate differ from the OLS findings in both magnitude and direction. The HLM estimates an effect size of -0.108 in Hedges' g, equivalent to a regression of about two months of progress. The 95% CI is much wider, ranging from -11 months to +9 months of progress, indicating greater uncertainty in the estimate. This difference should be interpreted cautiously, as the sample used for this analysis excludes 10.20% of observations in the main analysis for this outcome due to missing class identifiers, and the wider CI reflects the reduced precision of the HLM model compared to the OLS model.

Overall, the HLM analysis largely confirms the OLS findings for the primary outcome and mathematical self-efficacy. By accounting for the hierarchical structure of the data (students nested within classes and settings), the HLM demonstrates that the nested structure does not meaningfully change the effect size or CI for the primary outcome.

Checking for potential attrition bias using the GCSE Maths pass rate

The trial's primary outcome had an attrition rate of 48%, which introduces a high risk of bias in the effect size estimates. To explore how this attrition might have influenced the results, we repeated the analysis for the GCSE Maths pass rate using only the students who were part of the primary outcome analysis.

The full sample for the pass rate had a much lower attrition rate of 25%, as it includes students who sat the exam, but whose raw scores could not be collected from the settings. By comparing the results from the restricted sample (which matches the primary outcome sample) to the full sample of students that sat the exam, we can see if the difference in attrition rates leads to different effect sizes and what the direction of the potential bias is. Note that this analysis cannot give information on the attrition coming from students who did not sit the exam, it only informs our understanding of the bias stemming from data collection issues or from settings dropping out of the trial.

Table 31 presents the results of this subsample analysis. After restricting the sample to those same individuals included in the primary outcome analysis, the effect size moves closer to zero, indicating no additional months of progress. The lower bound of the 95% CI also decreases, shifting from -1 month of progress to -2 months, while the upper bound remains unchanged. This suggests a potential selection bias in the composition of the sample used to estimate the effect on the primary outcome. This may have caused a downward bias in the estimated treatment effect for that outcome.

Looking at the unadjusted means for both samples, the percentage of students who passed the exam in each treatment arm are higher in the restricted sample (17.7% and 17.6% for treatment and control) than in the full sample for the secondary outcome analysis (14.6% and 13.4%). This indicates that the students in the sample used in the primary outcome analysis were of higher ability on average compared to the full cohort of students in the trial who sat the GCSE exam. For these higher-ability students, the effect of treatment on the primary outcome was small and negative. The true average treatment effect, which includes a larger proportion of lower-ability students, might therefore, be closer to zero. However, these results do not suggest that the programme was positive either, as the GCSE Maths pass rate results do not provide evidence that the intervention improved the likelihood of passing the resit exam to a significant degree.

As with all other treatment effect analyses in this study, the results should be interpreted with caution due to the very wide CIs.

Table 31: Estimation of potential attrition bias on GCSE Maths pass rate

Outcome	Sample	Unadjusted means				Effect size		
		Intervention group		Control group				
		n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (95% CI)	P-value
GCSE Maths pass rate	Full GCSE Maths pass rate sample	2,357 (714)	0.146 (0.132 – 0.160)	2,015 (720)	0.134 (0.119 – 0.149)	4,372	0.014 (-0.083 – 0.111)	0.771
GCSE Maths pass rate	Primary analysis sample	1,631 (1,440)	0.177 (0.158 – 0.195)	1,401 (1,334)	0.176 (0.156 - 0.196)	3,032	0.003 (-0.105 – 0.111)	0.956

Exploratory analysis: Subgroup analysis with the GCSE Maths pass rate

As the secondary outcome of GCSE Maths pass rate was not subject to the high-attrition rates of the primary outcome, we repeated the same FSM subgroup analysis on this outcome to collect additional and hopefully more robust evidence of the difference in GCSE resit exam performance between FSM and non-FSM students. This was an exploratory analysis.

As shown in Table 32, the results differ from those of the subgroup analysis on UMS scores. The interaction term for the GCSE Maths pass rate outcome is -0.007 in Hedges' g, which is very small (close to null) and negative, indicating no significant difference in progress between FSM and non-FSM students (95% CI: [-2 months, +2 months]). When estimating the effect separately for each subgroup, both FSM and non-FSM students saw similar, negligible impacts (zero months of progress), with wide CIs. For non-FSM students, the 95% CI ranged from -1 month to +2 months, and for FSM students, it ranged from -2 months to +2 months.

Although the attrition rate for this outcome was lower than for the primary outcome, this binary measure (likelihood of passing the exam) is less sensitive to small changes in performance compared to the continuous UMS score. This likely explains the wider CIs in spite of the bigger sample size.

Both subgroup analyses (on UMS and GCSE Maths pass rates) show no evidence that the intervention improved GCSE resit performance for FSM students or that it had a different impact on them compared to non-FSM students.

Table 32: Exploratory subgroup analysis for FSM-eligible students

Outcome	Model	Total n (intervention; control)	Coefficient (SE)	Hedges' g (95% CI)	P-value	Secondary analysis Hedges' g (95% CI)
GCSE Maths pass rate	Interaction effect	4,372 (2,357; 2,015)	-0.002 (0.025)	-0.007 (-0.157 – 0.143)	0.928	0.014 (-0.083 – 0.111)
GCSE Maths pass rate	Subgroup (non-FSM)	2,857 (1,512; 1,345)	0.007 (0.017)	0.019 (-0.078 – 0.117)	0.692	0.014 (-0.083 – 0.111)
GCSE Maths pass rate	Subgroup (FSM)	1,515 (845; 670)	0.000 (0.026)	0.001 (-0.153 – 0.155)	0.986	0.014 (-0.083 – 0.111)

Estimation of ICC

The ICC at the setting level for the primary outcome (GCSE UMS score) was 0.08, and 0.09 for the subsample of FSM-eligible students, which were lower than initial predictions at the analysis stage (0.20). The ICC at the class level were 0.17 and 0.19, respectively.

Implementation and Process Evaluation results

Fidelity

The intervention had two levels of implementation; lead teachers implemented a PD programme to class teachers, and class teachers then taught a programme of lessons to their students. The plans provided for lead teachers and class teachers were designed to include some flexibility. The types of adaptation that were 'permitted' at both levels are specified in the intervention description (see Appendix E), which states that three things should not change in any setting: i) the key pedagogical principles; ii) the Lesson Study structure of the PD; and iii) the lesson materials provided to students.

Table 33 below shows the results of the fidelity survey conducted with the lead teachers. Lead teachers were asked to report the extent to which they adhered to the plans in the lead teacher pack, and to give their view on the extent to which the class teachers in their cohort adhered to the Maths-for-Life lesson plans.

Table 33: Fidelity survey results

Item	n (%) who always stuck to the plan	n (%) who made small changes	n (%) who made big changes
PD activities delivered by lead teachers:			
Lesson planning	4 (29)	10 (71)	0 (0)
Lesson observations	7 (50)	7 (50)	0 (0)
Reflection sessions	4 (29)	10 (71)	0 (0)
Lessons taught by class teachers	3 (21)	10 (71)	1 (7)

N (lead teachers) = 14. This is the whole population of lead teachers in the trial. Counts are presented as well as percentages due to the small sample.

These results suggest that the intervention plans were fairly closely followed. According to the survey, most lead teachers and class teachers made only small changes to the plans provided, with a minority of class teachers making what lead teachers considered to be big changes.

PD programme

Adaptations and their effects

The Lesson Study structure of the PD—which involved preparing for a lesson, teaching a lesson, peer observation of a lesson, and group reflection on each lesson—was followed by all lead teachers. At a more granular level, two types of adaptation were observed in implementation of the PD programme activities, one during the lesson planning sessions and one during the peer lesson observations.

During observations of **lesson planning sessions**, lead teachers followed the order of activities quite closely, but one main type of adaptation was observed. These sessions contained some activities that started with a stimulus (such as a video of a Maths-for-Life lesson) and then led into a group discussion. Each stimulus was designed, in part, to address a key concept in the Maths-for-Life approach to teaching. The guidance for lead teachers specified this concept in each case and provided a question to focus the PD discussion accordingly. In some cases, this part of the guidance was not followed by lead teachers. In one case, for example, the lead teacher followed an instructive video with a completely open-ended discussion, led by the participating teachers. The intervention description (Appendix E) does specify that some variation in discussion topics is encouraged so that the PD sessions are responsive to the needs of the participants, but it also specifies that key pedagogical principles of Maths-for-Life should not be altered. As a result of this deviation from the plan, the group's conversation moved quickly between a range of topics, without in-depth exploration and without any discussion of the intended topic (in this case, 'reciprocal dialogue'). This approach seemed to limit the value of the discussion for the

teachers, as no clear insights that might influence their practice were produced and the main 'lesson' of the activity was not addressed.

Observations of **teachers taking part in lesson observations** also revealed one type of deviation from the lesson observation guidance. Class teachers playing the role of observer during a lesson were asked not to intervene with students as a teacher would, i.e. not to help the students with their work. Some teachers found this difficult to do and were observed asking students questions to help them solve problems. The approach to helping students that was taken by these teachers was broadly aligned with the Maths-for-Life pedagogy—asking questions, rather than giving answers—and did not seem to have a negative effect on the students. However, by intervening in this way, these teachers did not have the opportunity to see how students were responding to the lesson under 'normal' conditions (without their intervention). Seeing this was one of the ways that the PD programme aimed to deepen teachers' understanding of how students learn. This adaptation may therefore, have resulted in a less effective PD experience for these teachers.

Factors affecting adherence

Three factors that affected adherence to the PD plan were identified from observations and interviews with lead teachers: i) the complexity and presentation of PD resources; ii) the preparation and attention to detail of the lead teachers; and iii) the class teachers' attitude towards and understanding of the PD process. These factors are discussed in turn below.

First, each lesson planning session included a range of different resources for the lead teacher to manage, including the lesson plan and slides that class teachers would be using with their students, instructive videos, handouts for class teachers, explanatory notes for lead teachers on elements of the theory, and an A3 sheet for the lead teacher that aimed to summarise the activities and resources for that session. Lead teachers valued the A3 sheet as a tool that summarised the key activities and resources to be used during each session, but the **complexity and slightly fragmented presentation of these resources** seemed to lead to some of the subtle adaptations described above in terms of discussion topics (Appendix F). In moving between activities and resources, lead teachers sometimes missed the detail of instruction. One lead teacher described this happening because they did not want to disrupt the flow of the session by having to refer too closely to the notes. Another suggested that a slideshow for lead teachers would have helped them to manage the flow and details of the activities. While this teacher saw a rationale for avoiding slides—as this may take away from the discussion-based nature of the sessions—they described finding it difficult to ask the key session questions and ensure that the key topics were discussed in the absence of a tool like this.

Second, the **preparation and attention to detail of the lead teacher** also played a role here. One lead teacher described not taking enough time to prepare for the early PD sessions, resulting in less flow and some details being missed. The other lead teacher that was interviewed also highlighted the importance of detailed preparation given the complexity of each session.

Third, the fact that some class teachers intervened with students when they were supposed to be observing seemed to be in part down to their understanding of the PD process. During one reflection session that was observed by the researcher, teachers who had intervened to help students did not seem to understand why it was important for them to hold back and what they may have missed as a result. The lead teachers interviewed described having detailed discussions about this in their first PD sessions but reported some class teachers struggling with the idea throughout the programme.

Lessons

Adaptations and their effects

As with the PD, the survey results suggest that the five Maths-for-Life lesson plans were fairly closely followed by the majority of class teachers, with a large majority reporting making 'small changes'. However, lesson observations and interviews revealed three types of adaptation made by class teachers that may have influenced the effectiveness of the intervention: i) removing planned dialogue; ii) dropping key activities; and iii) adding new content. These are discussed in turn below.

In some cases, teachers **removed planned dialogue** from the lessons. This was the case for instances of both student-to-student and teacher-to-student dialogue. In the former type, students were sometimes encouraged to work alone because they were substantially ahead of the rest of the group in their mathematical understanding. Alternatively, some students were allowed to work alone if they refused to participate in paired and group work. In less extreme cases, teachers ran an activity that required student-to-student dialogue but did not enforce the nuances of the activity's design. For example, some paired activities stipulated that students should take on specific roles (the scribe and the reader) and swap these roles for each new problem, as a way of encouraging both students to speak, but some teachers did not enforce this rule. The result of these deviations was that some students did not interact with their peers at all, and some had more limited interactions than were intended. Some teachers also removed planned elements of teacher-to-student dialogue; for example, by running whole-class plenaries with didactic demonstrations, rather than by asking the class questions that were built into the lesson plans. Beyond this, teachers also sometimes found it difficult to adopt a dialogic approach when offering one to one and small group support during activities.

> *I think a lot of ours are so low in confidence that sometimes it feels like if you don't help and talk them through it really quickly then they just get totally lost and worried that they are not going to get [it] at all.* (Class teacher 1)

Student-to-student and teacher-to-student dialogue are both central parts of the intervention's logic model, and the findings about causal mechanisms below suggest that both were effective mechanisms for some students. Class teachers that removed some opportunities for planned dialogue from the lessons may therefore, have reduced the effectiveness of their lessons. The case of more advanced students was not clear cut on this point, however. Some teachers saw value for these students in explaining concepts to their peers, and deepening their own understanding, or consolidating their knowledge in the process. Where the gap was very wide, however, it seemed more productive for more advanced students to work on different problems that were beyond the understanding of the rest of the group.

> *It's just that my classes are…such a wide level of ability from barely scraping, maybe grade 2/3 up to grade 7. [So] I use different tactics… [On] that particular occasion [one student] had advanced and I knew exactly where her algebra skills were from her previous education… [S]he immediately could go straight to the end [by] using algebraic expressions. She doesn't come for the full time each week because she works, and for her to spend an hour and a half peer teaching it wasn't appropriate.* (Class teacher 2)

In this class, which was an extreme case in terms of the range of student ability, the whole-class plenary discussions were also removed by the teacher because even those students who were working in groups made progress at very different rates. This meant that some of the more advanced students were left to progress to the next activity on their own, without the teacher being able to prepare them for it. This resulted in some confusion, with students reverting to simpler methods, rather than those encouraged in the lesson plan. However, the class teacher in this case was conscious of this issue, quickly circulating around groups to bring them back on track and using support staff to help with this. This seemed to be effective but required a high level of awareness and skill from the teacher, which was beyond the level of skill observed in other classrooms; particularly those where the dialogic approach was new to the teacher.

As well as changing the approach to teaching, some teachers were also observed **dropping key activities** from lessons, including those that were designed to induce cognitive conflict, representations that were designed to deepen student understanding, and the final part of lessons that were designed to bring closure. Of the first type, there was one activity that teachers found particularly difficult to use effectively. In this activity, students were asked to review two different example responses to a problem, work out what had been done differently in the two cases and, which response was correct. The intention of this activity was to induce cognitive conflict and, in so doing, reveal two important algebraic concepts. For some teachers, this activity induced too much confusion among their students, whose basic grasp of the key concepts was not strong enough to grapple with a problem like this.

> *[A] lot of people were getting confused…It was trying to show them that the boxes [variables] can [represent different values]…But, I felt it was just that slight jump too far to get to this. I don't think they were solid enough in what that meant and why [the two examples produced two different answers].* (Class teacher 1)

Some of the representations that were designed to expose the mathematical structure of a concept and deepen understanding were also seen as confusing or superfluous by some teachers. In this category, there were some objections to specific representations—for example, a number line—but there were also teachers who saw dialogue as the core component of the intervention, and reported using the representations only occasionally.

> *I think it's about the dialogue and the doing, and for some students I think the [representations were] really important and for others [they weren't]. I liked giving lots of different options. I think you know your students and you have to adapt everything to your students.* (Class teacher 3)

During a lesson observation, this teacher also decided not to complete the 'closure' part of the lesson. This also seemed to be based on a decision to focus on student-to-student dialogue, in part because the group was so lively and very difficult to engage in the whole-class elements of the lesson. This class did engage in problem-solving in pairs and groups, but this lack of emphasis on the representation and closure components of the lesson may have left students with a less secure sense of understanding.

In one observation, a **teacher added a starter activity** that was not in the Maths-for-Life lesson plan. This activity was a worksheet with questions to be done on an individual basis, covering a range of topics, all of which were different to the main topic of the lesson. In interviews, students described this approach as confusing but characteristic of their 'normal' (non-Maths-for-Life) lessons, which often skipped between multiple topics. The discussion of causal mechanisms below picks up on this issue in more detail.

Factors affecting adherence

Five factors were identified from interviews and observations that seemed to affect class teacher adherence to the lesson plans and approach: i) teacher preparation; ii) flexibility in the scheme of work; iii) the quality of the teaching resources; iv) teacher attitudes; and v) student attitudes and understanding. These factors are discussed in turn below.

Class teachers described two elements of the intervention that helped them prepare in detail for each lesson, which in turn supported their adherence to the lesson plans. First, the lesson planning component of the PD, which involved an in-depth discussion of each lesson before it was taught, was identified by lead teachers and class teachers as particularly important. The fact that these sessions encouraged teachers to question the rationale of the lesson design and included videos of elements of the lessons being taught, seemed to contribute to their effectiveness. Second, class teachers observed their lead teacher teaching Lesson 1 before attempting it themselves. This was described as building confidence in the approach that made it more likely that class teachers would be faithful to it in their own classroom.

> *I let them observe me first and…I think that was a good thing to do, because it allowed them to see that it's okay to teach like this and that actually, the students generally enjoy it. You get quite a nice vibe going on in the classroom, and it also gave them the experience of seeing the questions the teacher needs to ask in action, as well as the videos. In the feedback after that session they were quite surprised…about how I was during the lesson.* (Lead teacher 2)

Third, in some cases class teachers tried out Maths-for-Life lessons with other classes; some that were in the trial and some that were not. Sometimes these practice lessons resulted in teachers reducing their adherence (because they felt that an activity did not work) but sometimes increased adherence (because they became more comfortable with the approach). Both class teachers and lead teachers suggested that the depth of preparation that Maths-for-Life supported was considerably greater than in their normal practice. Teachers did not have the time to prepare for 'normal' lessons in this way, but this level of preparation was seen to be very valuable. Related to this, teachers who felt that they had the **flexibility in their scheme of work** to spread a Maths-for-Life lesson over multiple sessions were able to cover the content more comprehensively, for example, ensuring that the 'closure' part of the lesson was not missed, as described above.

The high **quality of the teaching resources** was highlighted by both lead teachers and class teachers. Lessons were seen to be well-structured and supported by good slides (which supported the facilitation of activities and discussions, rather than acting as a script for teachers).

> *All the materials are done for you, they're all differentiated. There's a clear plan attached as to how to use them in what circumstance and how to conduct it well, how to ask questions, how to prompt learners to construct their own understanding through the activities, and how to end the lesson. So it's extremely comprehensive in that sense, and I would recommend using it because not only is it great for that particular lesson you're teaching but also is a springboard as it was with me, ideas how to develop other material for different topics.* (Class teacher 2)

This quality was seen to affect adherence by giving teachers confidence in the approach and by providing a clear framework within which they could implement the more difficult parts of the pedagogy.

> *You'll get teachers saying, 'I couldn't do that with my students. My students struggle too much', and it's not the case. It's just that they lack [the] confidence to…[let] them have a go. I think the design and structure of [the] lesson[s] enables that discussion to take place.* (Lead teacher 1)

Teachers' attitudes towards the intervention seemed also to affect adherence. Attitudinally, teachers who were committed to faithful implementation and perceived the intervention as effective, seemed more likely to follow the lesson plans and the dialogic approach more closely. The former attitude seemed more important when it came to adherence. Some teachers were agnostic about the latter (even leaning to more didactic practice in their day-to-day teaching), but were still observed following the plans extremely closely, because they were committed to engaging in the details of the design and trying the intervention as it was intended to be delivered. However, others—who reported strongly believing in the dialogic approach—quickly dropped or adapted activities when they perceived them not to be working for their students. One teacher in this second group, who reported being very experienced and comfortable with the dialogic approach, seemed to make the decision to drop key representations quite quickly.

> *[I]t was just all these things going everywhere. They were just like, 'this is too confusing,' so I said, 'okay if you all have different methods just put it on the wall,' and we took it from there.* (Class teacher 3)

It was not clear whether this decision was based on sound judgement, or a lack of engagement with the details of the intervention's design, in this case, the rationale for the representation. In a lesson where the teacher took a very different approach—sticking closely to the representations and their rationale—it did seem to benefit the students. For example, in a lesson on algebra, some students in this latter teacher's class wanted to skip the use of the pictorial representation (a series of boxes) and move directly to using algebraic expressions. In an effort to maintain fidelity, the teacher prevented them from doing this, insisting that they progress through the lesson as it was designed. These students were observed making mistakes in constructing algebraic expressions at the beginning of the lesson and using them effectively by the end of the lesson, in conjunction with the box-based representation that they had been taught.

The final factor that seemed to influence adherence was **students' attitudes and level of mathematical understanding**. Some students refused to participate in paired and group work; a core part of every Maths-for-Life lesson. Students in this category who were interviewed described finding it difficult to have conversations with students that they did not know. During lesson observations, teachers were seen effectively encouraging students like this who were reluctant, but potentially willing. In some cases, this resulted in sustained group work, but in others, these students reverted to individual work as soon as the teacher left them. In extreme cases, where students were highly resistant, teachers did not attempt to place them in pairs or groups at all. Other students, who were happy with the group work, refused instead to engage with some of the representations—another core part of the intervention—sometimes preferring methods that they had been taught at school or that their peers had shared with them. This was explained by some class teachers as a desire on the part of the students to find the most efficient method. However, lesson observations revealed a slightly more complex picture. While some students were observed rejecting Maths-for-Life representations on the grounds of efficiency, it sometimes seemed that a lack of understanding (and a lack of desire to develop understanding) was the real reason. As with the attitude towards group work, the teacher influenced whether or not this attitude led to a lack of adherence. Some teachers strictly enforced the use of each representation when it was specified in the lesson plan, where others did not.

I think it's about the dialogue and the doing, and for some students I think the [representation] was really important and for others it wasn't. I liked giving lots of different options. I think you know your students and you have to adapt everything to your students. (Class teacher 3)

Dosage

Table 34 summarises the amount of the PD programme received by class teachers and the number of Maths-for-Life lessons received by students. Less than half of teachers (44%) participated in all core PD activities. However, on average, they attended almost four out five of the intended planning sessions and lesson observations. There was very high variation in the length of the planning and reflection sessions that were run by lead teachers for class teachers; an average of two hours, with a range of one to five hours. The plan was for each planning and reflection session to run for approximately three hours (roughly one hour for reflection and two hours for planning the next lesson). The dosage here is therefore, broadly in line with the intervention specification.

The picture is better for students, with 75% of classes receiving all five Maths-for-Life lessons, and an average of four per class. As with the class teacher PD sessions, lesson length varied widely; between 40 minutes and nearly 120 mins. The average lesson length of 90 mins is 30 mins longer than the intended 60 mins. However, the intervention was designed to allow for this wide range, explicitly specifying that teachers might take between 60 mins and 120 mins for each lesson.

Table 1: Summary statistics on dosage

Item	Statistic
Dosage received by class teachers:	
% teachers who participated in all PD activities (five planning sessions and five research/observation lessons)	37/84 teachers (44.1%)
Mean number of planning sessions attended	3.77 sessions (0–5)
Mean number of lesson observations attended	3.69 observations (0–5)
Mean length of planning sessions	124 mins (~120 mins) (67–300 mins)
Dosage received by students:	
% classes that received all five lessons	187 classes (75.4%)
Mean number of lessons delivered per class	4.34 lessons (1–5)
Mean length of lessons	88 mins (~90 mins) (40–218 mins)

N (class teachers) = 84. This is the whole population of class teachers in the treatment group. Figures in brackets are the range for that statistic.

Factors affecting dosage

Three factors were identified as affecting how much of the intervention was received by students. First, class size seemed to play a role, with one teacher suggesting that having a small class (12 students) made it easy to teach lessons within the specified time (and sometimes in less time). Second, student understanding and progress was identified as an important factor. One lead teacher suggested that some Maths-for-Life lessons could be doubled in length, depending on the level of progress made by the class, and this was seen to happen during lesson observations. For example, an observed lesson that covered expanding and factorising was split over two 60-minute-sessions (one for each sub-topic) based on the teacher's judgement of the class's level of understanding. Third, teachers' views about an acceptable amount of progress in terms of topic coverage (in line with the wider scheme of topics to be covered during the year) also seemed to affect how much time was given for each Maths-for-Life lesson. One lead teacher suggested that some teachers went into lessons with a certain amount of content that needed to be taught in mind.

I think [the class teacher] is probably wanting to get through to where he wants to get through to…[H]e could be a bit more relaxed and just a bit more able to listen to what the students are saying, reflect on it and just have a slightly more context-dependent response. (Lead teacher 2)

Programme differentiation

Control group experience

Teachers in the control group were asked to complete a survey, reporting their participation in PD programmes that could be similar to Maths-for-Life in the past five years. Following the guidance of the developers, six questions were asked in this survey to represent the six key characteristics of the Maths-for-Life PD. Table 35 below shows the findings.

Table 35: Control teacher programme differentiation survey results

Item	% (% in past year)
Teachers in the control group who reported taking part in PD in the last five years on…	
maths teaching	93 (48)
teaching specific maths topics	83 (44)
how to ask questions to deepen students' understanding	74 (39)
how to use group work to support student learning	70 (33)
formative assessment	78 (24)
student-centred approaches to teaching	74 (30)
Teachers in the control group who reported taking part in all six elements in the past five years	48
Teachers in the control group who reported taking part in three or more elements in the past five years	89

N (class teachers in control) = 46. This is approximately half of the total population of control group class teachers in the trial.

These results suggest that the majority of teachers in the control group had participated in PD programmes with similar elements in the past five years, with almost half taking part in programmes that covered all six elements specified. A substantial minority also reported taking part in PD with similar elements during the intervention period ('in the past year'). This might mean that the observed effects of the intervention are smaller than they otherwise would have been. There may have been an element of selection bias in the trial sample here, with settings and teachers that are more amenable to the Maths-for-Life approach signing up to the trial, though we do not have any evidence to verify this hypothesis. However, these findings should be read with caution for two reasons. First, it is very difficult to accurately assess by survey whether the control group teachers' experiences really were similar to those in the intervention group. While efforts were made to specify the different elements of the Maths-for-Life PD and lesson study approach, these brief descriptions have a range of different but reasonable interpretations and the survey tells us nothing about the quality of the PD received by each teacher.[31] Second, we cannot say from these results whether these potentially similar PD experiences in the control group transferred into similar practices in the classroom (the ultimate driver of any effects on students).

Intervention group experience

The results of the survey with teachers in the intervention group suggests a slightly different picture. Here, only 34% reported having taken part in PD that was similar to Maths-for-Life in the past or during the trial period.[32] Around 58% reported that

[31] The developers at the University of Nottingham believe that it is very unlikely that teachers had participated in similar programmes, particularly ones with a Lesson Study approach, given their knowledge of the available provision in England.
[32] Given the random assignment, this difference between intervention and control group is unusual, and more likely an artefact of the surveying, rather than a real difference. Teachers in the intervention group were asked to directly compare their Maths-for-Life PD

the Maths-for-Life approach to teaching was different to their normal approach, with 17% saying that it was very different. This means that nearly half of the intervention group (42%) considered the Maths-for-Life approach to be similar to their normal practice. Given the random assignment of settings to the intervention, a similar proportion of teachers in the control group would be expected to say the same if presented with the Maths-for-Life intervention. If these teachers are right, this is further reason to expect a smaller effect from the intervention, as nearly half the teachers in the control group will have been teaching in the same way. Having said this, the case study findings above, suggest that teachers found many core elements of the lessons different and, in some cases, difficult to implement. The qualitative data—especially from student interviews—revealed five clear points of difference between the Maths-for-Life approach and some teachers' day-to-day practice.

First, all Maths-for-Life lessons focused on a single topic. In interviews, students identified this as a valuable feature of the intervention's approach.

> *In his usual lessons, what he does he puts objectives onto the side of the board, but he'll go from say, ratios and fractions and jump to algebra and Pythagoras…[The Maths-for-Life lessons are] easier to understand because it's just one thing. He's not telling us like fifty different things on questions we're not even on, because some of us work faster in the group and with this we're all on the same page…and not working ahead of each other and getting confused about what he's saying.* (Student 16)

Second, Maths-for-Life lessons were described by students as more interactive in three senses: they involved interaction with peers through paired and group work; they involved activities other than exam questions (such as card matching); and they led to more interaction between the teacher and the students.

> *[In the lessons that focus on] exam questions, [the teacher] just gives it to us and he sits there sort of thing. So he gives it to us for half an hour, 40 minutes and will go through it on the board. Then it will just be like, 'some got it wrong, and some got it right.' But when it's a Maths-for-Life [lesson], he is walking around and always engaging and talking to people and that sort of thing, rather than just leaving it with us.* (Student 18)

Third, students described non-intervention lessons as less focused on developing their understanding.

> *[Normal lessons are] like being spoon fed. [The teacher] spoon feeds us and we don't want to do that, we want to learn why is it that. So that when we come to it next time we're not just expecting him to put the answer down.* (Student 16)

Fourth, both students and teachers identified a difference in the quality of planning and design between Maths-for-Life and other lessons. Students described Maths-for-Life lessons as having a clearer structure and progression from one activity to the next. Teachers described basing some other lessons around exam questions, with no activities that were designed to develop a broader understanding of the topics in question. One class teacher in the case study cohort was teaching with no scheme of work at all and expressed demand for double the number of Maths-for-Life lessons to fill this gap.

Responsiveness

Class teacher engagement

Class teachers were asked by survey to report how engaged they felt in the three parts of the PD programme. Table 36 below, shows the findings. These results suggest that teacher engagement was high. Interviews with class teachers suggested that where general practice was changed, this took place in one or more of three areas: they became less focused on exam

---

experience with previous experiences, whereas teachers in the control group had no such direct experience of the intervention to use as a point of comparison. It is perhaps not surprising then that the class teacher survey suggests a greater differentiation between the intervention and other experiences that the class teachers may have had.

questions; they increased the amount of teacher–student dialogue; and they adapted some of the Maths-for-Life activities (like card matching) and representations for other topics.

Table 36: Teacher responsiveness survey results

Item	% (% strongly agree)
% class teachers who reported being actively engaged in the lesson planning sessions	84 (48)
% class teachers who reported being actively engaged in the lesson observation sessions	89 (3)
% class teachers who reported being actively engaged in the reflection sessions	83 (45)
% class teachers reporting a change in their general teaching practice (outside of Maths-for-Life lessons)	83 (14)
% class teachers reporting resources supported differentiation	70 (13)

N (class teachers in treatment group) = 64. In total, there were 84 class teachers in the treatment group. Response options were: Strongly agree; Agree; Neither agree nor disagree; Disagree; Strongly disagree; Don't know; and Don't want to answer.

Factors affecting teacher engagement

From interviews with teachers and observations of PD sessions, three factors were identified as influencing teacher engagement in the intervention: i) the style of the PD; ii) the perceived robustness of the intervention; and iii) the fact that teachers chose to participate.

Lead teachers believed that the **style of the PD was particularly** appealing to teachers. They liked the depth of engagement with a single lesson that the Lesson Study process required, and the interactive nature of the programme.

> *I think it is the process; it's the lesson study…[T]hey get a chance to talk and think about a lesson, do the lesson, see the lesson and then talk about it afterwards…I think it's them involved with it rather than something being done to them…I think they like the time involved.* (Lead teacher 1)

Conversely, when PD sessions became less interactive and more didactic (which happened rarely due to the design), the attention of some teachers was observed to wane. Class teachers also described the **robustness of the approach**—both in terms of the intervention and the evaluation—as an important factor that motivated their engagement. Related to this was an expectation that the intervention was likely to produce benefits for students.

> *I have been on the EEF [website], and I knew it was going to be a randomised control study…It wasn't just something that you know, somebody fancied having a go at. I knew it was Malcolm Swan, it was based on his work and I've used a lot of his stuff before, so I just figured it would be something robust.* (Class teacher 6)

Finally, **where teachers chose to participate** this also seemed to contribute to the high levels of engagement reported in the surveys and seen in observations and interviews. This included teachers who were keen to try something new, as well as some teachers who were very familiar with dialogic teaching and the Standards Unit Box.

Student engagement

The class teacher survey asked teachers to report how engaged they believed their students to be in Maths-for-Life lessons. The results suggest that the majority of students were engaged in Maths-for-Life lessons, but a minority of teachers suggested that this engagement was higher than in other lessons and a low proportion strongly agreed with the survey statements; perhaps reflecting a general difficulty in engaging students in any GCSE resit lessons. The levels of student engagement seen during lesson observations varied substantially, especially with the elements of the lessons that required student-to-student dialogue. In some observed classes, the interaction between students during paired and group activities was extremely limited, requiring almost constant teacher input to encourage and maintain discussions. At the other end of the spectrum, engagement in student-to-student dialogue was extremely high. In this case, students were so lively in their

conversations (which were on topic) that other elements of the intervention became difficult to deliver (described in more detail below).

Table 37: Student responsiveness survey results

Item	% (% strongly agree)
% class teachers reporting that students engage in Maths-for-Life lessons	84 (17)
% class teachers reporting that student engagement is higher in Maths-for-Life lessons than others	39 (6)
% class teachers reporting that students enjoy Maths-for-Life lessons	66 (9)
% class teachers reporting that students enjoy Maths-for-Life lessons more than others	28 (0)

N (class teachers in treatment group) = 64. In total, there were 84 class teachers in the treatment group. Response options were: Strongly agree; Agree; Neither agree nor disagree; Disagree; Strongly disagree; Don't know; and Don't want to answer.

Factors affecting student engagement

Interviews with teachers and students revealed five factors that seemed to influence student engagement in the intervention: i) the depth of learning; ii) enjoyment of group work; iii) students' maths confidence; iv) the relating of maths to 'real life'; and v) the difference between Maths-for-Life and other lessons.

Students described Maths-for-Life lessons as encouraging a **greater depth of engagement with the subject** by basing each lesson around a single topic and asking students to focus on their understanding (as opposed to the application of techniques to get to the right answers). This focus on depth was supported by the interactivity of the lessons, the use of representations, and more focus from the teacher on student understanding. Some students described this approach as motivating their engagement in the lessons (as well as supporting their learning).

> *[I]n a normal lesson you write, you leave, you forget. But these sorts of lessons you go in and you have to actually apply it instead of just doing questions and answering. So, when you are having to apply it you're learning and you are recognising where you are going wrong and where you are going right. So then, you just start understanding the algebra, proportions or anything…[N]ormally you will get a few questions, but it's mostly just the teacher talking to you, do this, do this, do this…But in these sorts of lessons it seems more like we will give you this, try it. He [teacher] will obviously help us if we don't understand it, so then we will be going through it and start getting it and start getting better…. It makes more sense, because the whole point in doing maths is to learn maths.* (Student 19)

However, for some students, this focus on depth and understanding led to them disengaging with the intervention. Students in this category were exam-focused and were sometimes attached to methods or techniques that they had learned at school and considered to be reliable.

> *For me it's quite contradictory because we've already got methods, so instead of learning methods that we've already got and perfecting them. It's like going back to square one and learning methods that you might not particularly use so for me it's just time wasting.* (Student 14)

Related to these attitudes was a focus on getting the right answer (rather than understanding why it was right) and, sometimes, a lack of interest in deepening their maths understanding and, for some, a fear of not being able to answer specific exam questions without step-by-step processes to follow in each case.

> *I literally watched a quick five or ten minute video on YouTube and that worked better for me than the [Maths-for-Life] cards if I'm honest with you. [T]hat video literally went step-by-step and how to work out this question…It has to relate to exam questions because if it doesn't, then the exam question could be something completely different to what you have learnt and it's really difficult to get your head around*

> *it…[T]here's numbers literally everywhere, and I need to learn how to focus on narrowing it down. It sounds silly, but it can be quite intimidating for some students.* (Student 19)

The **extent to which students enjoyed group work** seemed to be an important determinant of their engagement. Some students described being engaged in the intervention lessons as a result of the group work elements. For some, this was because the dialogue required by the lessons made them feel valued.

> *I like this class because everyone gets a chance. Everyone gets a chance to say an answer, and everyone gets a chance to say this is my way of doing it, and I like that. I'm learning from you, and you are helping me.* (Student 4)

This effect of group work on engagement seemed to be moderated by the quality of the relationships between students. Even students who were very low in confidence enjoyed group work if these relationships were positive.

> *I struggled at school, and I was bullied at school, so…I kind of sat at the back of the class and thought 'oh god'…[W]hen I was asked something I couldn't even speak. [I]t's…completely different now to what it was back in the school days…I think it's because we're all grown-up…There's a lot of positivity.* (Student 1)

However, both teachers and students believed that, where these conditions were not present, students with low communication confidence found it difficult to engage with their peers. Beyond this, there were some students for whom confidence was not an apparent issue, but who just preferred working alone. However, as noted above in the findings on fidelity, high engagement in group work and discussion did not necessarily mean better learning. In the lesson observation where student engagement in group work was highest, it was difficult for the teacher to hold the class's attention for long enough to explain important representations and to deliver the 'closure' part of the lesson.

Students' general **confidence in maths** also seemed to influence their engagement. Some students described not wanting to engage in discussions unless they felt confident in the topic. However, both teachers and students suggested that the problems set in Maths-for-Life lessons were of a type that most students could engage in some way.

> *I think the [intervention] is just trying to encourage [the students] to explore a bit and not panic about 'we are not going to know this' or 'we are not going to get the answer.' It's just doing some maths, and generally, the tasks are set-up where they can do something, even if it's not necessarily the right thing. They can have a go.* (Class teacher 1)

Finally, some students were engaged by the fact that the Maths-for-Life problems were **related to 'real life'** (e.g. paint prices), and others described enjoying the lessons simply because they were **different to their normal lessons**.

> *It's refreshing for us. It's nice to do different.* (Student 1)

Quality

Quality of the PD programme

The class teacher survey results suggest that teachers rated their PD experience highly, believing that their lead teacher had the right combination of skills, experiences, and attitudes for the job.

Table 38: Quality of PD programme according to class teachers

Item	%
% class teachers rating the PD as good or excellent	78
% class teachers reporting that the lead teacher was a good facilitator	77
% class teachers reporting that their lead teacher had the right professional experience for the role	80

% class teachers reporting that their lead teacher had strong understanding of Maths-for-Life pedagogy	81
% class teachers reporting that their lead teacher believed in Maths-for-Life pedagogy	81

N (class teachers in treatment group) = 64. In total, there were 84 class teachers in the treatment group. Response options for the first item were: Excellent; Good; Average; Poor; Very poor; Don't know; and Don't want to answer. Response options for the other items were: Strongly agree; Agree; Neither agree nor disagree; Disagree; Strongly disagree; Don't know; and Don't want to answer. For the latter set, we report the percentages of teachers who Agreed or Strongly agreed.

Factors affecting the quality of PD

PD observations and interviews with lead and class teachers identified five factors that influenced the quality of the PD programme: i) the structure and management of discussions; ii) the complexity and presentation of PD resources; iii) the lead teacher's understanding of the Maths-for-Life theory and lesson designs; iv) the quality of the cohort; and v) the structure of the programme.

The PD programme involved a lot of discussion between class teachers, both in the lesson planning sessions and the reflection sessions that followed the lesson observations. **The way that lead teachers structured and managed these discussions** seemed to have a strong effect on the quality of the PD experience. During the lesson planning sessions observed, lead teachers were observed asking questions that facilitated productive learning conversations. For example, after showing a video clip of a Maths-for-Life lesson, the lead teacher began by asking: 'What happened?', which sparked a good discussion about the students' understanding and the teacher's approach. The lead teacher then followed up with: 'What would you do to move things on?', which led to a specific and productive discussion between class teachers on possible approaches that could be taken in the classroom. Lead teachers also asked questions that facilitated productive learning conversations during the reflection sessions that followed the peer-to-peer lesson observations, but these sessions were sometimes less successful overall. In the reflection sessions observed, there was no obvious structure to the discussions, and class teachers offered a range of observations, which were often unrelated to each other and were not always probed in sufficient detail to result in concrete ideas that teachers could take into their classrooms. During these sessions, lead teachers were seen to give better quality feedback than class teachers, giving some detailed descriptions of students' experiences and linking these descriptions back to the Maths-for-Life pedagogy. However, lead teachers also reported struggling to give effective feedback, even when they had identified specific things that a class teacher could do to improve.

> *I think the reason why I didn't say anything was because I didn't know what words to say.* (Lead teacher 2)

The **complexity and presentation of PD resources** has been described above as affecting fidelity but was also identified as a factor affecting quality. As described above, managing discussions effectively required skill and effort from the lead teachers, and when elements of the PD instructions became too complex, the quality of discussions sometimes suffered. In one observation, for example, the facilitation guide contained the instruction to discuss 'how class teachers actions can facilitate and inhibit cognitive conflict'. The lead teacher struggled to ask a question that met this aim and, consequently, the discussion did not address the topic at all. More generally, lead teachers sometimes described being unclear on substantial details of session plans. In one case, for example, the lead teacher did not know whether they should be showing the slide show for the lesson that was being discussed, as this detail was unclear in the facilitation guide. The effects of this complexity in the resources was related to the **lead teachers' understanding of the Maths-for-Life theory and lesson designs**. In the pilot year, the PD was delivered by the developers, with one facilitator who was focused on the design of each lesson and one that had a deep knowledge of the theory of dialogic learning. One lead teacher identified gaps in his knowledge and understanding relative to these facilitators.

> *[Developer 1] was totally on top of the material…[Developer 2] would…give us a little bit of theory, some literature. So you've got [Developer 2] as an absolute expert, which is inspirational in itself and because [Developer 1] created the materials he did know them inside out, and that probably wasn't there with me at the start.* (Lead teacher 2)

As the programme relied on class teachers sharing ideas, knowledge, and experience, the dynamics of the cohort of class teachers in each cluster also seemed to influence the quality of the PD. Two factors were identified as important here. First, class teachers valued the creation of an environment where they felt safe to have open discussions about their practice.

> *[I]t's a safe environment and you have got to have trust in each other…[It's important to know] that you can talk to each other freely and it's not a personal criticism and it's not going back anywhere saying that they do this and they do that.* (Class teacher 5)

Second, the **quality of the cohort**, in terms of experience and commitment, was also seen as important by some.

> *They were all really experienced teachers which was fantastic, so everybody discussed what they thought might be issues. It was a very open and frank discussion, and after you delivered a lesson people were saying that 'this went well,' and 'that didn't go well,' and 'you want to try something different.' We always met, and everybody was always there, and it was really great.* (Class teacher 3)

Finally, while the findings above focus on the way that the PD was facilitated by lead teachers and the makeup of the teacher cohorts, the overall **structure of the programme** was a factor that seemed to ensure a certain level of quality, in the sense that it encouraged a useful process of doing, observing, reflecting, and planning.

> *I like it all. There is observing and then reflecting afterwards and…discussing…what was to come.* (Class teacher 3)

Quality of Maths-for-Life lessons

The framework used to assess the quality of delivery of lessons was the Maths-for-Life 'five key pedagogies': cognitive conflict; formative assessment; collaborative learning; models of structure (or 'representations'); and closure.

Three levels of teacher engagement with **cognitive conflict** as a tool for teaching and learning were identified through lesson observations. At the first level, teachers tried to minimise the amount of confusion experienced by students by intervening quickly and directly. When students were working in small groups, this intervention included correcting errors while students were in the middle of a problem, offering unsolicited suggestions for changes in approach, and even taking the students' pen and writing solutions out on the groups' mini whiteboards; all before asking any questions of the students. For example, after a couple of minutes of starting the first problem of the lesson, a pair of students asked a teacher for help. Rather than asking the students a question, the teacher began with a direct instruction to, 'Start with x = Jelly Snakes'. In doing so, the opportunity seemed to be missed to assess what the students did or did not understand, and to allow them to use their confusion to deepen their understanding. When teachers were operating at this level, they allowed almost no cognitive conflict to be expressed during whole-class discussions, preferring instead to either take input from students who obviously understood, or to bypass student discussion and demonstrate how to apply a method 'correctly' themselves.

> *I think a lot of [our students] are so low in confidence that sometimes it feels like if you don't help and talk them through it really quickly then they just get totally lost and worried that they are not going to get [it] at all. Obviously, you don't want them having that feeling.* (Class teacher 1)

At the second level, teachers allowed students to struggle with problems and concepts during group work, using questions instead of explanations as the starting point for any teacher–student dialogue. As with teaching in the first level, however, when it came to the whole-class elements of the lesson, the teacher avoided allowing students to express cognitive conflict and reverted to a more didactic approach. At the third level, teaching actively sought out opportunities to use cognitive conflict to deepen students' understanding, both in small group and whole-class work. For example, in one of the intervention's lessons, there is an activity that asks the class to review example student responses, try to work out which one is correct, and try to establish the misunderstanding behind the incorrect response. Teachers were observed finding this activity very difficult to run due to the level of confusion it engendered in the class, with some choosing to drop it entirely. In one case however, a teacher did implement it fully, encouraging a wide range of students to give 'incorrect' answers and to explain their (often confused) reasoning. This teacher used extensive questioning to probe the responses given and, as

the discussion developed, some students were observed having moments of realisation; eventually offering correct and comprehensive explanations of the problem. Working with cognitive conflict at this level was challenging and time consuming, as is discussed in more detail in the subsection 'Factors affecting the quality of Maths-for-Life lessons' below.

Teachers' use of **collaborative learning**, **formative assessment**, **models of structure**, and **closure** related closely to these three levels of engagement with cognitive conflict. A greater engagement with cognitive conflict was accompanied by a greater focus on collaborative learning and formative assessment. For example, the teacher that is described above as being less comfortable allowing students to grapple with confusion, facilitated less discussion (both student-to-student and teacher-to-student), asked fewer questions and, consequently, carried out less formative assessment during the lesson. This teacher had a lively style of communication, which kept students engaged but, because they seemed to want to maintain the pace of the lesson and avoid silences and disengagement, few students were encouraged to speak. Conversely, the teacher described above that fully committed to the use of cognitive conflict as it was designed into the lesson, had to facilitate more discussion, and ask more questions. In one exchange with a student, for example, the questions: 'How do you know you're right?'; 'Is that the only thing you need to do to know you're right?'; and 'How could you check that that's right?', helped a student to realise and resolve a mistake, and supported the teacher's assessment of that student's understanding. This teacher was also more committed to the models of structure that were built into the lessons, as these were designed to help overcome students' confusion. In this lesson, closure seemed easier to establish, as students had engaged more deeply with the problems and had moments of realisation during the lesson, which could be drawn together in a plenary discussion at the end.

Factors affecting the quality of Maths-for-Life lessons

Four factors were identified from interviews with teachers and students as affecting the quality of Maths-for-Life lessons: i) the quality of the lesson and resource design; ii) the quality of planning and preparation; iii) the teacher's ability to manage a tension between collaborative learning and closure; and iv) the teacher's ability to form positive relationships and support students emotionally.

Teachers and lead teachers described the **quality of lesson plans and resources** as very high. The lessons were described as having a logical progression in terms of students' learning, with starter activities at each stage that prepared students well for the next step in the lesson. The activities for students were seen to facilitate the pedagogical approach, for example, a card matching activity on ratios and fractions provided an easy beginning to a student-to-student dialogue ('Where shall we put the first card?'), led to a predictable bit of cognitive conflict ('1:4 is the same as ¼') and gave a framework for teachers to use to help resolve that conflict. Teachers also described finding it helpful that all problems were placed in real-world contexts but simplified to an appropriate degree. Finally, the slideshows were seen as well-thought through in terms of their role in the lesson, supporting the dialogic approach rather than acting as a distraction.

> *All the materials are done for you, they're all differentiated. There's a clear plan attached as to how to use them, in what circumstance and how to conduct it well, how to ask questions, how to prompt learners to construct their own understanding through the activities, and how to end the lesson. So it's extremely comprehensive in that sense, and I would recommend using it because not only is it great for that particular lesson you're teaching but also is a springboard as it was with me, ideas how to develop other material for different topics.* (Class teacher 2)

The **quality of planning and preparation** was also identified as a key factor affecting the quality of the intervention lessons. The intervention supported high-quality preparation through in-depth planning sessions with other teachers (as described above). Class teachers also appreciated being able to see their lead teacher teach the first lesson before trying it themselves. Some teachers also benefited from being able to teach each intervention lesson multiple times, either because they had more than one class in the trial, or because they practised the lessons with classes that were not in the trial. On the other hand, the fact that there were only five topics covered by the intervention, and that the lessons had to be taught in a certain order, made it difficult for some teachers to integrate the intervention into their scheme of work effectively. Both teachers and students identified this as an issue.

> *It's like one week we will come in and do a Maths-for-Life lesson and the other day we will come in and do a booklet based on triangles…It's like all over the place.* (Student 14)

Teachers judged **the balance between collaborative learning and closure** differently, and this was related to the amount of cognitive conflict that they were comfortable with encouraging. The quality of this judgement seemed to affect the quality of student learning. Where teachers intervened quickly and didactically, students seemed to take less from the lessons. This was particularly because they were designed to be taught dialogically, so even if a didactic approach could be effective, it was difficult to execute in the context of the intervention. Teachers who were observed struggling with this balance in class were aware of this and sometimes described the Maths-for-Life approach as very different to their normal practice and difficult to implement within the constraints of a large class.

> *The idea is to try and let [the students] solve things, and as a teacher it's really hard to say, 'Work it out for yourself, talk to each other, I'm not going to help.' You're wanting them to talk to each other and work it out and that's difficult. As a teacher you're often the first port of call and when you've got a big group to get around you haven't time, so say, just think about it and I'll come back in a minute…[But] it could be my fault as a teacher, because as a teacher we do try to be helpful.* (Class teacher 5)

As described in the findings on fidelity above, however, some teachers also seemed to misjudge this balance in the opposite direction. In one case, the teacher allowed the whole lesson to be based on group work, skipping the plenaries, teacher-led explanations, and closure element of the lesson. In this instance, the students were highly engaged throughout, but many left with unresolved misconceptions. The teacher described this as a conscious decision, but it was not clear that the right balance had been struck.

> *I just decided to let them go because I felt in that moment that there were so many misconceptions, and they were really talking them through. I could see that the weaker ones were really engaged and I just thought I would let it go.* (Class teacher 3)

The final factor identified as affecting the quality of the lessons was **teachers' ability to form positive relationships and support students emotionally**. In its simplest form, this factor manifested itself in basic behaviour management. Both teachers and students identified the importance in Maths-for-Life lessons of ensuring that more vocal students did not dominate. They also described regular low-level behavioural problems such as lateness, talking over the teacher, ignoring instructions, and using mobile phones (all of which were observed by the researcher), which needed to be managed by the teacher for the intervention to be effectively implemented. There were also more subtle issues in this category that required the teachers' attention. Encouraging the use of new representations, allowing cognitive conflict, and asking students to try to understand the maths required skilful questioning and emotional support from the teachers in the lessons observed. For example, in one group where confidence was very low, the teacher spent a lot of time and energy encouraging the students to get started by trying any approach with which they felt comfortable, initially ignoring the model of structure that was supposed to be applied. When a student had been working on their own with trial and error for a while, the teacher reapproached the student, gave them some enthusiastic positive reinforcement and then asked some questions to encourage them to try some algebra. This caused anxiety in the student, so the teacher stepped back again.

> *Don't worry, you're doing brilliantly! Step back from any method, trust your instinct and start again. I'll back off.* (Class teacher 2)

The teacher returned to the same table ten minutes later and tried again, this time asking the student to pair up and exchange methods with a peer who had made more progress, which resulted in a productive conversation where the student began to understand the algebraic approach. This process required a close attention to the student's feelings, knowledge of their level of understanding, a visibly positive attitude, and a strong commitment to the pedagogy. In interviews, students in this class emphasised the positive and supportive atmosphere as key to their learning.

> *[The teacher is] a very lively character and she just allows us to be us, and to ask questions and she goes, 'Don't hold back!'. And I think that is the biggest thing: we have to just be us and ask anything.* (Student 9)

Causal mechanisms

PD programme

Class teachers were asked by survey to report whether they believed that each of the PD mechanisms hypothesised in the logic model (Figure 1) were present for them. Table 39 below shows the findings. A majority felt that most of these mechanisms were present, with two notable exceptions. Half of respondents felt that their own mathematical understanding and reasoning had improved as a result of the intervention and a similar number felt that they were strategically planning for high-quality student dialogue. The first of these mechanisms was also absent in the class teacher interview data, but the second (strategic planning for dialogue) was mentioned by some teachers. The results of the qualitative analysis are given below, with descriptions of how the mechanisms may have worked, and why they sometimes may not have worked.

Table 39: Presence of causal mechanisms according to class teachers

Item	%
% class teachers who reported…	
being exposed to other teachers' practices	94
being better at reflecting on practice	73
improving their own mathematical understanding and reasoning	51
improving their understanding of student learning	67
focusing more on deepening student understanding	75
encouraging and guiding peer-to-peer reasoning	69
strategically planning for high-quality student dialogue	55

N (class teachers in treatment group) = 64. In total, there were 84 class teachers in the treatment group. Response options were: Strongly agree; Agree; Neither agree nor disagree; Disagree; Strongly disagree; Don't know; and Don't want to answer. For the latter set, we report the percentages of teachers who Agreed or Strongly agreed.

Exposure to other teachers' practices was seen as beneficial by interviewees for one or more of three reasons. First, seeing another teacher doing things differently gave some teachers' the confidence to change their approach. For example, after observations teachers described allowing cognitive conflict to develop more readily and giving more time for students to work on problems than they otherwise would have. For some teachers, the benefits of this mechanism went beyond classroom practice.

> *Because we are a niche provision here, it's so interesting to see how other teachers do it in a 16 to 18 environment—not the actual delivery of the lesson. That hasn't varied much in terms of how I do it to be honest. It's what resources they have, how they manage the planning…* (Class teacher 2)

The peer observation element of the programme was closely linked to teachers feeling **better at reflecting on practice** (when they did feel this). For some teachers, the most powerful ideas they had for improving their own practice came to them while observing other teachers. In particular, teachers described **improving their understanding of student learning**; thinking in more detail about how different students would react to different parts of lessons, predicting the challenges and confusions that may arise, and planning for these accordingly. Teachers with more experience and seniority—for example, those with roles supporting practice in their department—described less of an effect in terms of their personal reflection. Reflection for them was a regular part of their work, and they were also supporting colleagues to reflect and develop their practice.

Interviews with students suggested that they experienced a **greater focus on deepening their understanding** during Maths-for-Life lessons. This increased focus was, according to these interviews, supported by: i) an emphasis on students

solving problems (rather than following teacher demonstrations and copying from the board); ii) **encouraging and guiding peer-to-peer reasoning**; iii) an emphasis on rough working and making mistakes (encouraged by the use of mini whiteboards); and iv) a progression in each lesson from simpler to more complex problems, based on a single topic.

> *[The teacher] made us do it by ourselves to be honest. She restricted the guidance…[S]he wanted us to do it and work it out for ourselves so we would be able to understand it.* (Student 4)

> *I liked the fact that whenever you're [working on a problem], [the teacher] throws in little questions just to make sure you are understanding as well. He's not just there to be like 'okay, this is what you do.'* (Student 27)

The findings on quality above, suggest that teachers considered the time given to planning and design of the lessons and activities as key in **encouraging high-quality student dialogue**. Even teachers who described themselves as experienced and confident with the dialogic approach said that the Maths-for-Life lessons resources helped them to facilitate better quality conversations.

> *What we have done [at this college] for a while is trying to get [the students] to talk. But we haven't put enough structure in…[T]hese activities, [e.g. ask students to] put something down and explain actually why.* (Class teacher 3)

Maths-for-Life lessons

Class teachers were asked by survey to report whether they believed that each of the classroom mechanisms hypothesised in the logic model (Figure 1) were present at some point during their Maths-for-Life lessons. Table 40 below shows the findings. These numbers suggest that teachers believed that all key hypothesised mechanisms were present some of the time. The small/negative effects of the programme estimated in this study, do not therefore, seem to be down to a failure in this regard.

Table 40: Presence of causal mechanisms according to class teachers

Item	%
% class teachers who reported seeing…	
teacher-to-student meaningful dialogue	74
student-to-student meaningful dialogue	84
students using representations to understand mathematical structures	80
students building shared chains of reasoning	70
students experiencing cognitive conflict[a]	77
students increasing in confidence	80
students improving in their mathematical reasoning	75
students improving their verbal communication skills	69

[a]*The role of cognitive conflict is addressed in detail on the findings on quality so is not addressed in the qualitative findings here.*
N (class teachers in treatment group) = 64. In total, there were 84 class teachers in the treatment group. Response options were: Strongly agree; Agree; Neither agree nor disagree; Disagree; Strongly disagree; Don't know; and Don't want to answer. For the latter set, we report the percentages of teachers who Agreed or Strongly agreed.

Interviews with students supported the survey findings that **meaningful teacher-to-student dialogue** was taking place during the intervention lessons. Students described teachers taking more time over problem-solving conversations, being proactive about starting conversations with students (rather than waiting for people to ask for help), and showing more

enthusiasm for dialogue. One student thought that this increased enthusiasm from the teacher was a result of increased engagement from the students.

> *I think when we do the Maths-for-Life [lessons the teacher] is much more bubbly and going around and helping everyone. Whereas if we were just doing the exam papers, he would just go around asking if we are okay, and then go through the answers. I think [that's] because when we actually do [the Maths-for-Life activities and] we are interested…it makes him happy and feels like he can go around and help people. If we are just sitting there doing exams, the exam questions are hard and a lot of the time people don't do them anyway.* (Student 15)

Interviews with students suggested that **student-to-student dialogue** did take place during Maths-for-Life lessons, but that this dialogue was not always helpful. When it worked, students were able to develop **shared chains of reasoning** and described this dialogue as: i) helping them to reinforce their knowledge and understanding by explaining to others; ii) helping them to work more at their own pace (as compared to whole-class demonstrations); iii) adding another source of help when the teacher was busy; iv) providing explanations from peers that were sometimes easier to understand (as compared to those from the teacher); v) exposing them to different approaches to a problem such that their understanding of the underlying topic was deepened; and vi) making maths work more memorable (as compared to working through questions alone).

> *Like the expanding brackets one, I was looking at it and literally knew the answer in my head straightaway. But because he didn't and me explaining it to him basically reinforced what I already knew. It was quite good for me to go over it and it was good for him to learn the explanations.* (Student 14)

When it was less effective, students described peer-to-peer dialogue as: i) a source of distraction (with off-topic chat); ii) difficult when a discussion partner was reluctant to engage; and iii) difficult when discussion partners were too far apart in their level of understanding.

Students described liking the **representations** in Maths-for-Life lessons, and some described them as helping them to understand mathematical structures (though not in those words). Three characteristics of the representations were identified by students as particularly helpful: i) they helped students who struggled with reading to access problems; ii) they provided a starting point for students to tackle a problem (as opposed to having a blank sheet of paper); and iii) they reduced the level of abstraction (e.g. replacing algebraic notation with boxes).

> *I feel like my example would be the number line [question that asked us to convert] pounds to dollars. I feel like I already knew [the topic] and what to do, but [the diagram helped me to understand] what to do…[S]o I have broken it all down. Now it fits easily into my mind and if it was to come up on an exam paper, I would just think two number lines, blah, blah, blah.* (Student 9)

Both students and class teachers also described positive effects of the intervention on **student confidence**. This finding is supported somewhat by the impact analysis that found a small positive effect on mathematical self-efficacy.[33] This was seen to be supported in part by the pictorial representations, which gave some students with very low confidence (according to their class teachers) a way to begin problems that they would likely have not even started previously.

> *I think the stuff they were doing today, I could see one or two little lightbulbs going on where people are thinking. They'[d] look at those questions [in the past] and they wouldn't even try. Whereas today, when we got to question three or four they just got on with it.* (Class teacher 5)

Students also described their confidence being increased by working in small groups (when they felt unable to participate in whole-class discussions), and by being given problems that were pitched at the right level, with appropriate progression

[33] Notwithstanding the wide CI and risk of bias introduced by attrition.

in difficulty. This meant problems that were not so hard that they were inaccessible, but hard enough such that they provided challenge and satisfaction.

Teachers and students described positive effects on **students' mathematical reasoning.** In particular, students described applying problem-solving approaches that they had learned in Maths-for-Life lessons to new problems, having more than one way to approach a range of problems, and taking more time over reading and understanding problems before trying to solve them.

Finally, **verbal communication** did seem to improve for some students, but this seemed to be due to an increase in confidence and a more conducive classroom environment (factors described above), rather than a matter of 'skill'.

Cost

The financial cost of the intervention over three years in a school/college is summarised in Table 41 below. This shows that most of the financial costs of delivering the intervention during the evaluation period were incurred by the University of Nottingham and subsidised by the EEF. Were the programme continued without subsidy, this cost would have to be covered by settings in the form of a programme fee. The total cost of delivering the intervention per setting over three years is estimated to be £3,942.67. This is equivalent to £33.63 per student per year. Seven resources were identified as prerequisites for delivering the intervention: a computer; projector and screen (for the lesson slide show); a printer and photocopier (for lesson materials such as sorting cards); and small white boards and white board pens for each student. Some settings also used a laminator, Post-it™ notes, and paper clips, but these items were not seen as essential. The prerequisite and non-essential items that settings already had are not included as direct costs.

Table 41: Direct cost of delivering Maths-for-Life in a school/college

Item	Type of cost	Average cost per setting in year one	Total cost over three years	Total cost per student per year over three years[a]
The EEF subsidy	Start-up cost per setting	£3,566.32	£3,566.32	£28.08
Printing and photocopying	Running cost per setting	£15.58	£44.92	£0.55
Travel and subsistence for Continuing Professional Development (CPD) sessions	Start-up cost per setting	£316.10	£316.10	£4.90
Hosting CPD sessions (lunch and refreshments)	Start-up cost per setting	£15.33	£15.33	£0.10
Total		£3,913.33	£3,942.67	£33.63

[a]*Assuming 42 students per year—the mean average in our cost data sample.*

Table 42 below, shows the estimated cumulative costs for a setting running Maths-for-Life over a three-year period.

Table 42: Cumulative costs of Maths-for-Life (assuming delivery over three years)

Programme	Year one	Year two	Year three
Maths-for-Life	£3,913.33	£3,932.25	£3,942.67

Table 43, below, shows the estimated time spent per school, by school staff on the intervention, broken down by activity. The intervention is delivered at the class level—i.e. it is a series of five lessons to be delivered to a whole class—but can be delivered to more than one class. Among the settings interviewed for the cost evaluation, the number of classes receiving the intervention ranged between 1 and 14. To take this into account, the 'Year one time' and 'Total time over three years' figures in Table 43 are based on mean averages across interviewed settings. The 'Total time per class per year' is also given in the final column, based on the average time per class estimated for each item in the table.

Table 43: Time spent by setting staff on Maths-for-Life

Activity	Type of time	Year one time (hours)	Total time over three years (hours)	Total time per class per year (hours)
Class teacher delivery in school/college day (teaching Maths-for-Life lessons)	Running cost per setting	34	102	6.4
Teaching assistant delivery in school/college day	Running cost per setting	3	15	0.4
Class teacher delivery out of school day	Running cost per setting	0	0	0
Teaching assistant delivery out of school/college day	Running cost per setting	0	3	0.9
Class teacher CPD (lesson studies)	Start-up cost per setting	30	30	3.3
Class teacher lesson preparation and follow-up (e.g. marking)	Running cost per setting	10	23	3.8
Teaching assistant lesson preparation and follow-up (e.g. marking)	Running cost per setting	1	8	0.2
Supply cover	Running cost per setting	5	7	0.6
Other staff time	Running cost per setting	2	3	0.1
Total		85	190	16

Notes on the calculations

Of the 100 settings participating in the trial, six were interviewed for the cost evaluation, one for each setting type. For each of the items in the tables above, a mean average was calculated across all respondents.

Conclusion

Table 44: Key conclusions

Key conclusions	
1.	Learners in Maths-for-Life settings made two months' less progress in GCSE Maths scores, on average, compared to learners in other settings. This result has a low security rating.
2.	Among learners previously eligible for free school meals (FSM), those in Maths-for-Life schools made one month's less progress in GCSE Maths scores, on average, compared to those in other settings. These results may have lower security than the overall findings because of the smaller number of learners.
3.	There is no evidence that Maths-for-Life had an impact, either positive or negative on GCSE Maths pass rate. The result is uncertain due to high attrition.
4.	Learners in the intervention group were more likely to attend their GCSE Maths exam—a key improvement, given that attendance was low across both groups. Given the importance of exam attendance for achieving qualifications, this finding is noteworthy.
5.	The intervention was delivered broadly as planned, though some teachers made changes. Some classroom activities were adapted by teachers, which may have affected how closely the approach matched the original design.

Impact evaluation and IPE integration

Evidence to support the logic model

The evaluation findings provide partial support for the original logic model of the Maths-for-Life intervention. Some mechanisms functioned as expected, while others showed variation in implementation or lacked strong supporting evidence.

Activities

Fidelity survey data indicated that lead teachers generally delivered the PD component as intended. However, attendance at PD planning sessions and observations was lower than expected. Survey, observational, and interview data suggested that class teachers followed the Maths-for-Life lesson plans, though many made small adaptations. While teachers applied the five pedagogies promoted by the programme, interviews and observations revealed variation in the quality of dialogic teaching.

Mechanisms

The evaluation provided strong evidence for several key mechanisms of the PD programme. Through PD sessions and observations, teachers were exposed to new teaching practices, engaged in self-reflection, improved their understanding of student learning, and placed greater emphasis on deepening student understanding and encouraging peer-to-peer reasoning. There was less evidence to support the mechanisms that teachers improved their own mathematical understanding and reasoning, and their ability to successfully plan for high-quality student dialogue.

In terms of the Maths-for-Life lessons, the majority of teachers surveyed for the IPE believed that all hypothesised mechanisms were present for some students. Interviews and observations supported this conclusion. However, variation in the quality of dialogic teaching led to inconsistent experiences of cognitive conflict for students. Observations and interviews also indicated that peer-to-peer dialogue quality varied widely.

Moderating factors

The moderating factors hypothesised in the original logic model were extensive and fairly broad. Interviews and observations identified the following more specific moderating factors in four categories.

- **Lead teachers.** Preparation and attention to detail, facilitation skill and approach.
- **Class teachers.** Preparation, attitudes towards dialogic teaching and the intervention, ability to manage a tension between collaborative learning and closure, and ability to form positive relationships and support students emotionally.
- **Students.** Attitudes towards dialogic learning and level of mathematical understanding.
- **Wider context.** The composition of the group of class teachers in PD sessions and student class size.

Given the complexity of the Maths-for-Life materials and approach—both for the PD programme and lessons—the preparation, attitudes, and capabilities of the lead teachers and class teachers were absolutely key to the success or otherwise of implementation.

Interpretation

The evaluation found no evidence that the intervention improved GCSE Maths performance, the primary outcome. This inconclusive result is likely due to high attrition in the outcome data, leading to a smaller-than-anticipated sample and the possible introduction of bias. Evidence for the secondary outcome, mathematical self-efficacy, was also inconclusive due to high levels of missing data. However, interview findings suggested that Maths-for-Life lessons did increase some students' confidence. An exploratory analysis of exam sitting rates found that Maths-for-Life students were 6 percentage points more likely to sit the GCSE resit exam compared to control students (21.6% vs 27.8%), also suggesting a positive effect on student confidence and participation.

If the programme did have a negative or null effect (our best guess), this could have been down to one or more of two main factors. First, some lead teachers adapted the PD programme in ways that may have made it less effective; not always following the suggested topics for discussion and sometimes struggling to facilitate effective learning conversations. This lack of adherence to the plan and approach seemed to be driven by one or more of: the complexity and presentation of the PD resources; the preparation and attention to detail of the lead teachers; and the class teachers' attitude towards and understanding of the PD process.

Second, some class teachers adapted the lessons in ways that may also have limited their effectiveness; in particular, removing activities that were planned to support dialogue and/or adding content on completely different maths topics into what were supposed to be very focused lessons. Where adherence to the lesson plans and approach was low, students' attitudes seemed to play a big role. Teachers found it difficult where students were very resistant to group work, and during exercises that required students to assess their understanding, persist through the discomfort of problem-solving, and reflect on their mistakes. It was also especially hard when students had experienced a very different style of teaching and learning in secondary school. These challenges were compounded where the teacher was not fully bought into the Maths-for-Life approach. The pilot evaluation of Maths-for-Life (unpublished) that preceded this efficacy trial, identified risks related to low student confidence, the challenge of mastering the pedagogy, and the likelihood of quality dilution as the programme scaled. These risks do seem to have materialised.

Conversely, where the programme seemed to work well, five things seemed to be present. First, lead teachers were well prepared for the PD sessions; having taken the time to work through the complex resources, and to get to grips with the core concepts being taught. This preparation allowed them to facilitate much more effective discussions, and to explain core concepts more clearly. Second, class teachers did similar preparation for the lessons. Third, class teachers bought into the Maths-for-Life approach; either because they already believed in it, or because they were willing to try something new. Fourth, class teachers were adept at building strong relationships with, and providing emotional support to, students who were very low on confidence. Fifth, the composition of groups was conducive to the approach. For the PD component, this meant a group of teachers that was willing to reflect deeply on their practice and discuss these ideas with their peers. For the lessons, this meant classes that were small enough to enable teachers to support the dialogic approach.

There is very little causal evidence in the wider literature on the impact of dialogic teaching—or any other teaching approach—on post-16 maths attainment (Crisp *et al.,* 2023, p. 72). There is fairly strong evidence to say that interventions that facilitate purposeful, curriculum-focused dialogue can have large positive effects (up to seven months of additional progress), though these effects are greater in the early phases of education, and most studies have looked at the effects on reading (EEF, 2021a). Effects on maths appear to be much smaller (one month of progress) (EEF, 2021a). Collaborative learning (another dimension of the Maths-for-Life approach) might be able to have a large positive impact on maths attainment—estimated to be +5 months on average—though this estimate has relatively low security due to the age, quality, and lack of independence of the literature (EEF, 2021b).

The main impact findings from this study—on maths attainment and mathematical self-efficacy—do not add much to the existing literature on this teaching approach due to the insecurity of the findings. However, the relatively large effect that we have found on exam attendance (significant at the 5% level) is perhaps an interesting addition. Levels of exam attendance were extremely low in both the control and the intervention group (21.6% and 27.8%, respectively). It is obviously impossible for students to pass the GCSE if they do not sit the exam so, while the intervention may not have had a positive effect on attainment, it is promising in this regard. No other studies have been identified that have found an effect like this from the dialogic teaching approach, and we know that student motivation and engagement is a major barrier in the context of post-16 GCSE resits (Crisp *et al.,* 2023, p. 68).

There is a clear and pressing need for better provision in this field. The pass rate of students resitting their GCSE Maths is very low; 17% in 2024, 4 percentage points lower than when this trial ended in 2019 (Camden, 2024). Post-16 settings are stretched for resources, and the maths teaching workforce needs development. The supply of experienced maths teachers with a reasonable level of maths qualification in post-16 education is insufficient. This, along with a resit student intake that is low on motivation and confidence, makes the implementation of programmes like Maths-for-Life extremely challenging. There is, however, clearly a demand for high-quality PD opportunities for this cohort of teachers and ‘a lack of clear, developed and relevant programmes’ (Crisp *et al.,* 2023, pp. 69–70).

Limitations and lessons learned

The three main limitations of this evaluation were attrition, low power, and missing covariate data. We also discuss the limitations of the CACE analysis.

Attrition

High attrition was the most significant limitation. For the primary outcome analysis, 48% of participants were lost from randomisation to analysis, either due to missing outcome data, missing covariate data, or meeting the exclusion criteria.

Data collection from post-16 settings was very difficult, and not enough resources were allocated to the task. When it became apparent that so many post-16 settings were unwilling/unable to provide GCSE raw scores for students enrolled in the trial, we explored the option of obtaining these scores from the exam boards (the only other holders of the data). These boards were unwilling to support the request on this occasion, but we believe that this is one of the best options for improving data quality in trials like this in the future. A better, but harder to achieve, option would be for the DfE to collect and share this data in its national datasets.

The high level of attrition poses a substantial threat to internal validity. It has introduced potential bias into the estimated effect and has reduced the analytic sample (and therefore, the statistical power) substantially.

Low power

The MDES was estimated to be 0.20 SD at the point of analysis for the primary outcome. This puts it below/on the boundary of the EEF’s threshold for a threat to validity (EEF, 2019, p. 6). However, this is considerably larger than the average effect of

interventions evaluated by the EEF (+0.04 SD; Demack *et al.,* 2021, p. 12).[34] The study was therefore, likely underpowered for the primary outcome. Had we not experienced such high attrition, it may well have been a well-powered study however, as the initial power calculations seem to have overestimated the MDES.

Missing covariate data

Key Stage 2 Maths attainment, which was obtained through the NPD, was missing for 18.17% of the primary outcome sample, and Key Stage 4 Maths grade, for 8.4% of the sample. Both missing data rates were above the 5% threshold that was assumed at the design stage for risk of bias, and missingness was not completely random. Older students, those with lower Key Stage 4 attainment, and non-FSM-eligible students were more likely to have missing Key Stage 2 data. Older and higher Key Stage 2-attaining students were more likely to have missing Key Stage 4 data. These patterns suggest a mix of potential overestimation and underestimation of the treatment effect, making the overall direction of any bias unclear.

Sensitivity analyses showed that excluding covariates or using alternative imputation methods did not meaningfully change the results. However, the missing data rates on baseline covariates limited their usefulness in increasing the precision of the estimated treatment effects. Instead, their inclusion worsened the power problem caused by attrition.

While Key Stage 2 attainment is often used as a baseline measure in post-16 research, our results indicate that it may not be reliable for older students, as Key Stage 2 exams were introduced gradually between 1991 and 1998. Future studies should consider alternative measures, such as the most recent GCSE grade or mark prior to the intervention.

Additionally, missing dosage data led to limitations in the estimation of CACE. Specifically, we did not have data on actual student attendance in Maths-for-Life lessons, only whether each lesson had been delivered. As a result, the findings from the CACE analysis should be interpreted with caution, and the lack of accurate dosage data limits our ability to contextualise the findings effectively. Given the low attendance in post-16 settings, future field studies should prioritise obtaining reliable attendance data that can be used to more precisely assess the implementation of the programme.

Future research and publications

There remains a large gap in the literature on what works to support GCSE resit students to succeed in post-16 education. Future research could address this by re-running this trial, or trials on alternative promising approaches, but a radically different approach to data collection needs to be taken to make such research worthwhile.

[34] This analysis pooled a lot of different interventions, age groups, and subjects so, is not directly comparable, but still provides a helpful reminder that it is difficult for an intervention to have a large effect in the English education system.

References

Camden, B. (2024) *'GCSE Resits 2024: Maths Pass Rate Up But English Falls Again'*. FE Week. Available at: https://feweek.co.uk/gcse-resits-2024-maths-pass-rate-up-but-english-falls-again/ (accessed 01 March 2025).

Crisp, B., Hallgarten, J., Joshua, V., Morris, R., Perry, T. and Wardle, L. (2023) *'Post-16 GCSE Resit Practice Review'*. London: CfEY, University of Warwick, and the Education Endowment Foundation. Available at: d2tic4wvo1iusb.cloudfront.net/production/documents/Post-16-GCSE-Resit-Practice-Review.pdf (accessed 07 April 2025).

Demack, S., Maxwell, B., Coldwell, M., Stevens, A., Wolstenholme, C., Reaney-Wood, S., Stiell, B. and Lortie-Forgues, H., 2021. Review of EEF Projects. Evaluation Report. *Education Endowment Foundation*.

Department for Education (DfE). (2018) *'Schools, Learners and Their Characteristics: January 2018'*. Sheffield: Department for Education. Available at: https://assets.publishing.service.gov.uk/media/5b31eb0840f0b67f7f306124/Schools_Pupils_and_their_Characteristics_2018_Main_Text.pdf (accessed 03 December 2024).

Department for Education (DfE). (2019) *'Revised A level and Other 16-18 Results in England, 2017/2018'*. Coventry: Department for Education. Available at: https://assets.publishing.service.gov.uk/media/5c48878640f0b61704aec530/2018_revised_A_level_and_other_16-18_results_in_England.pdf (accessed 04 December 2024).

Department for Education (DfE). (2024) *'GCSE English and Maths Results'*. GOV.UK. Available at: www.ethnicity-facts-figures.service.gov.uk/education-skills-and-training/11-to-16-years-old/a-to-c-in-english-and-maths-gcse-attainment-for-children-aged-14-to-16-key-stage-4/latest/ (accessed 20 June 2025).

Department for Education (DfE). (2025) *'Key Stage 4 Performance'*. GOV.UK. Available at: Key stage 4 performance, Academic year 2024/25 - Explore education statistics - GOV.UK (accessed 01 November 2025).

Education Endowment Foundation (EEF). (2021a) *'Teaching and Learning Toolkit: Oral Language Interventions'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/oral-language-interventions (accessed 07 April 2025).

Education Endowment Foundation (EEF). (2021b) *'Teaching and Learning Toolkit: Collaborative Learning'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/collaborative-learning-approaches (accessed 07 April 2025).

Education Endowment Foundation (EEF). (2022) *Statistical analysis guidance for EEF evaluations*, Version 2022.14.11. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1768915411 (Accessed: 01 March 2022).

Hayward, H., Hunt, E. and Lord, A. (2014) *'The Economic Value of Key Intermediate Qualifications: Estimating the Returns and Lifetime Productivity Gains to GCSEs, A Levels and Apprenticeships'*. GOV.UK. Available at: www.gov.uk/government/publications/gcses-a-levels-and-apprenticeships-their-economic-value (accessed 01 March 2025).

Herman, J.L., Matrundola, D.L.T., Epstein, S., Leon, S., Dai, Y., Reber, S. and Choi, K. (2015) *'The Implementation and Effects of the Mathematics Design Collaborative (MDC): Early Findings from Kentucky Ninth-Grade Algebra 1 Courses'*. CRESST Report 845. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Available at: https://cresst.org/publication/the-implementation-and-effects-of-the-mathematics-design-collaborative-mdc-early-findings-from-kentucky-ninth-grade-algebra-1-courses/ (accessed 01 March 2025).

Hume, S., O'Reilly, F., Groot, B., Kozman, E., Barnes, J. Soon, X.-Z., Chande, R. and Sanders, M. (2018) *'Retention and Success in Maths and English: A Practitioner Guide to Applying Behavioural Insights'*. London: The Behavioural Insights Team and Department for Education. Available at: www.bi.team/publications/retention-and-success-in-maths-and-english-a-practitioner-guide-to-applying-behavioural-insights/ (accessed 01 March 2025).

Hupkau, C. and Ventura, G. (2017) *'Further Education in England: Learners and Institutions. Briefing Note 001'*. London: Centre for Vocational Education Research. Available at: **https://cver.lse.ac.uk/textonly/cver/pubs/cverbrf001.pdf** (accessed 03 December 2024).

Jadhav, C. (2018) *'The Ofqual Blog: GCSE 9 to 1 Grades: A Brief Guide For Parents'*. GOV.UK. Available at: https://ofqual.blog.gov.uk/2018/03/02/gcse-9-to-1-grades-a-brief-guide-for-parents/ (accessed 07 April 2025).

Johnson, E. (2020) 'Action Research'. *Oxford Research Encyclopedia of Education*. Available at: https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-696 (accessed 15 May 2025).

Johnston-Wilder, S., Lee, C., Brindley, J. and Garton, E. (2015) *'Developing Mathematical Resilience in School-Students Who Have Experienced Repeated Failure'*. Paper in Eighth Annual International Conference of Education, Research and Innovation (ICERI2015) , Seville, Spain, 16–18 November 2015.

Joint Council for Qualifications. (2021) *'GCSE (Full Course) Outcomes for Key Grades for UK, England, Northern Ireland & Wales, Including UK Age Breakdowns Results Summer 2021'*. London: Joint Council for Qualifications. Available at: www.jcq.org.uk/wp-content/uploads/2021/08/GCSE-Full-Course-Results-Summer-2021.pdf (accessed 01 March 2025).

Kuczera, M., Field, S. and Windisch, H.C. (2016) *'Building Skills for All: A Review of England. Policy Insights from the Survey of Adult Skills'*. Paris: OECD.

Lauer, S.A., Kleinman, K.P. and Reich, N.G. (2015) 'The Effect of Cluster Size Variability on Statistical Power in Cluster-Randomized Trials'. *PLoS ONE,* 10: 4, e0119074. https://doi.org/10.1371/journal.pone.0119074

Menzies, L., Ramaiah, B. and Boulton, C. (2021) 'A Space for Maths: Exploring the Need for Maths Tutoring and the Potential Role of Third Space Learning'. London: The Centre fir Education & Youth. Available at: https://cfey.org/reports/2021/09/a-space-for-maths-exploring-the-need-for-maths-tutoring-and-the-potential-role-of-third-space-learning/ (accessed 01 March 2025).

Murray, C. (2017) *'English and Maths GCSE Resit Results 2017'*. FE Week. Available at: https://feweek.co.uk/2017/08/24/english-and-maths-gcse-resit-results-2017/ (accessed 30 April 2018).

Nolan, D. and Taylor, P. (2020) *'Statistical Analysis Plan'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/projects/Maths-for-Life_SAP_20200610_v1.1.pdf?v=1743418249 (accessed 01 March 2025).

Nolan, D., Taylor, P., Solomon, P. and Heal, J. (2020) *'Trial Evaluation Protocol'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/projects/M4L_trial_protocol_20200610_v1.2.pdf?v=1743418249 (accessed 01 March 2025).

Office of Qualifications and Examinations Regulation (Ofqual). (2013) *'Marking and Grading in GCSE and A Level Exams. Ofqual'*. GOV.UK. Available at: www.gov.uk/government/publications/gcse-and-a-level-exams-how-marking-and-grading-works/marking-and-grading-in-gcse-and-a-level-exams (accessed 07 April 2025).

Office of Qualifications and Examinations Regulation (Ofqual) Analytics. (2017) *'Grade Distributions of Reformed (9-1) GCSEs'*. Available at: https://analytics.ofqual.gov.uk/apps/2017/GCSE/9to1/ (accessed 30 April 2018).

Pampaka, M., Kleanthous, I., Hutcheson, G.D. and Wake, G. (2011) 'Measuring Mathematics Self-Efficacy as a Learning Outcome'. *Research in Mathematics Education,* 13: 2, 169–190. https://doi.org/10.1080/14794802.2011.585828

Pearson Edexcel. (2015) *'Grade Boundaries: Edexcel GCSE'*. London: Pearson. Available at: https://qualifications.pearson.com/content/dam/pdf/Support/Grade-boundaries/GCSE/1506-GCSE-Grade-Boundaries.pdf (accessed 30 April 2018).

Ritchie, J., Lewis, J., McNaughton Nicholls, C. and Ormston, R. (Eds.). (2013). Qualitative research practice: A guide for social science students and researchers. London, Thousand Oaks, CA: Sage Publications Ltd.

Schulz, K.F. and Grimes, D.A., 2002. Sample size slippages in randomised trials: exclusions and the lost and wayward. *The Lancet*, *359*(9308), pp.781-785.

Swan, M. and Green, M. (2002) *'Learning Mathematics Through Discussion and Reflection'*. London: Learning and Skills Development Agency.

Swan, M. (2006) 'Learning GCSE Mathematics Through Discussion: What Are the Effects on Students?' *Journal of Further and Higher Education,* 30: 3, 229–241. https://doi.org/10.1080/03098770600802263

Swan, M. (2007) 'The Impact of Task-Based Professional Development on Teachers' Practices and Beliefs: A Design Research Study'. *Journal of Mathematics Teacher Education,* 10: 4–6, 217–237. https://doi.org/10.1007/s10857-007-9038-8

Swan, M. and Swain, J. (2010) 'The Impact of a Professional Development Programme on the Practices and Beliefs of Numeracy Teachers'. *Journal of Further and Higher Education,* 34: 2, 165–177. **https://doi.org/10.1080/03098771003695445**

Takhashi A. and Wake G. (2023) *The Mathematics Practitioner's Guidebook for Collaborative Lesson Research*. Routledge.

Teddlie, C. and Tashakkori, A., 2008. *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage publications.

Thomson, D. (2017) *'GCSE Results Day 2017: Good News About Resits in English'*. FFT Education Datalab. Available at: https://ffteducationdatalab.org.uk/2017/08/gcse-results-day-2017-good-news-about-resits/ (accessed 30 April 2018).

Thomson, D. (2025) *'How Many Pupils Resit English and Maths?'* FFT Education Datalab. Available at: https://ffteducationdatalab.org.uk/2025/09/how-many-pupils-resit-english-and-maths/#:~:text=How%20many%20times%20do%20pupils,many%20do%20so%20only%20once. (accessed 22 October 2025).

Wake, G., Foster, C. and Nishimura, K. (2020) Lesson Study: A Case of Expansive Learning in Borko, H. and Potari, D. (Eds.) *Teachers of Mathematics Working and Learning in Collaborative Groups, ICMI Study 25*.

Wake, G., Swan, M. and Foster, C. (2016). Professional learning through the collaborative design of problem-solving lessons, *Journal of Mathematics Teacher Education*. 19 (2) 243-260.

Appendix A: The EEF cost rating

Appendix A Table 1: Cost rating

Cost rating	Description
£ £ £ £ £	*Very low:* less than £80 per student per year.
£ £ £ £ £	*Low:* up to about £200 per student per year.
£ £ £ £ £	*Moderate:* up to about £700 per student per year.
£ £ £ £ £	*High:* up to £1,200 per student per year.
£ £ £ £ £	*Very high:* over £1,200 per student per year.

Appendix B: Security classification of trial findings

PRIMARY OUTCOME: Key Stage 5 GCSE Maths resit performance for the 2018/2019 academic year, as measured by the UMS (Uniform Mark Scale) score.

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	**MDES**	**Attrition**					
5	Randomised design	<= 0.2	0-10%					
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%			Adjustment for threats to internal validity **[0]**		
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%					
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%	**1**				**1**
0	No comparator	>=0.6	>50%					

Threats to validity	Threat to internal validity?	Comments
Threat 1: Confounding	Low	The FSM eligibility and gender imbalances at baseline did not translate into significant differences in outcome attainment, suggesting that the gender split did not affect the balance in baseline attainment.
Threat 2: Concurrent interventions	No information available	There is no evidence of concurrent interventions influencing the results. The similar experiences reported in the survey by the control group refer to exposure to similar PD training content in the past five years, which helps to assert the programme differentiation. Therefore, this occurred before the intervention was delivered.
Threat 3: Experimental effects	Low	Unlikely to occur given the implementation model.
Threat 4: Implementation fidelity	Moderate	The intervention was well defined, aligned with the logic model, and generally delivered as intended. However, attendance was lower than expected, and interviews and observations revealed variation in the quality of dialogic teaching.

Threat 5: Missing data	Moderate	The missing data was high, which might introduce bias. However, the difference between groups was minimal (1.89 percentage points). The evaluator did not use multiple imputation as it would only increase the sample from 52% to 65%. The sensitivity analyses exploring missing data yield similar results to the primary analysis in all cases. The effects are negative and very small, and all CIs are extremely wide, encompassing zero, which reflects a high level of uncertainty. Attrition on dosage also affected the CACE analysis, which constrains the comparison with a model adjusting for missing data.
Threat 6: Measurement of outcomes	Low	Well justified, and no reasons for concerns related to its reliability, validity, utility, or acceptability. Ceiling effects were explored and deemed not concerning.
Threat 7: Selective reporting	Low	Study registered, trial protocol and Statistical Analysis Plan published prior analysis, and deviation reported. Data submission is planned.

- **Initial padlock score.** 1 Padlock – The randomised controlled trial was initially powered to detect an effect size of 0.15–0.22 for FSM students. However, due to fewer participating settings and smaller cluster sizes at randomisation, the actual MDES was higher, around 0.30–0.34. At the analysis stage, the MDES improved to approximately 0.20, as the observed ICC was substantially lower than anticipated (0.08 vs the assumed 0.20). Therefore, it can be argued that 5 Padlocks can be assigned to the Design and MDES criteria. Nonetheless, despite assuming 20% attrition in the protocol, 47.78% of attrition was found to be due to missing covariates, resulting from both student-level and setting-level attrition.

- **Reason for adjustment for threats to validity.** No adjustment was made as only two moderate threats were identified, and the bias direction is unknown.

- **Final padlock score.** Initial score adjusted for threats to validity = 1 Padlock – Very well conducted trial. Factors beyond the evaluator's control limited the validity and robustness of the evaluation.

Appendix C: Changes since the previous evaluation

Appendix C Table 1: Intervention and evaluation changes

	Feature	Pilot to efficacy stage
Intervention	Intervention content	Between the pilot phase and the efficacy trial, the Maths-for-Life programme moved to a more manualised and operational delivery model. While the pilot focused on testing and refining materials with 20 teachers in small clusters and preparing them to become lead teachers, the efficacy version specifies a clear, structured six-day model of professional development and lesson study cycle for clusters of teachers, including peer observation and reflection. The content focus shifted to five key challenging areas of the GCSE mathematics curriculum, with explicit use of student-centred, problem-solving, and dialogic teaching approaches. The trial implemented a cascade delivery model, where trained lead teachers support new teachers (expanding from 20 to ~100 teachers), and the intervention is applied in regular classrooms across multiple types of institutions. Overall, the efficacy version emphasises rigorous pedagogical frameworks, student-centred approaches, specific curriculum focus, structured lesson cycles, and broader scalability.
	Delivery model	In the pilot year, the PD programme was delivered by staff from the University of Nottingham and the lessons were delivered to students by maths teachers. In the trial year, a train-the-trainer programme was implemented, where staff from the University of Nottingham provided initial training, and some ongoing support, to a group of 'lead teachers'. These lead teachers then delivered the PD programme with a cohort of teachers in the local area. The teachers who received the PD then delivered the lessons with their students.
	Intervention duration	Duration of the intervention remained the same.
Evaluation	Eligibility criteria	In the pilot year, only colleges were eligible. In the trial year, colleges, schools and private training providers were all eligible, and the sample included representation from all these groups.
	Level of randomisation	Not applicable.
	Outcomes and baseline	Not applicable.
	Control condition	Not applicable.

Appendix D: Effect size estimation

Appendix D Table 1: Effect size estimation

			Intervention group		Control group		Pooled variance
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	
GCSE Maths UMS score	-0.888	-0.809	1,631 (1,440)	81.468	1,401 (1,334)	63.004	72.937
GCSE Maths GCSE pass rate	0.012	0.005	2,357 (714)	0.125	2,015 (720)	0.116	0.121
GCSE Maths pass rate (robustness check)	0.000	0.002	1,631 (1,440)	0.145	1,401 (1,334)	0.145	0.145
Maths self-efficacy score	0.021	0.015	564 (2,507)	0.170	533 (2,202)	0.157	0.164
GCSE Maths UMS score (CACE analysis)	-0.600	-1.124	1,631 (1,440)	81.468	1,401 (1,334)	63.004	72.937
GCSE Maths UMS score (HLM)	-0.881	-0.676	1,620 (1,451)	81.143	1,401 (1,334)	63.004	72.731
GCSE Maths pass rate (HLM)	0.013	-0.039	2,023 (1,048)	0.136	1,757 (978)	0.126	0.131
Maths self-efficacy score (HLM)	0.024	0.018	560 (2,511)	0.168	532 (2,203)	0.157	0.162

Appendix D Table 2: Effect size estimation, by FSM subgroup analysis

				Intervention group		Control group		
Outcome	Model	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
GCSE Maths UMS score	Interaction effect[a]	0.070	0.149	591 (405)	80.300	426 (374)	66.030	74.325
GCSE Maths UMS score	Subgroup (non-FSM)	-0.717	-0.828	1,040 (959)	81.838	975 (874)	61.622	72.056
GCSE Maths UMS score	Subgroup (FSM)	-1.086	-0.637	591 (405)	80.300	426 (374)	66.030	74.106
GCSE Maths pass rate	Interaction effect (exploratory analysis)	-0.001	-0.002	845 (151)	0.111	670 (130)	0.106	0.109
GCSE Maths pass rate	Subgroup (non-FSM) (exploratory analysis)	0.016	0.007	1,512 (487)	0.132	1,345 (504)	0.121	0.127
GCSE Maths pass rate	Subgroup (FSM) (exploratory analysis)	0.006	0.000	845 (151)	0.111	670 (130)	0.106	0.109

[a]*The Hedges' g for the interaction term has been computed using the unconditional SD of the FSM-eligible subgroup.*

Appendix D Table 3: Effect size estimation, by setting type subgroup analysis

				Intervention group		Control group		
Outcome	Setting type	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome[a]	n (missing)	Variance of outcome	Pooled variance
GCSE Maths UMS score	Further education college	-0.851	-0.852	1,406 (1,249)	80.786	1,113 (998)	63.601	73.194
GCSE Maths UMS score	School	0.048	0.323	73 (41)	64.676	124 (85)	59.287	61.277
GCSE Maths UMS score	Sixth-form college	-0.918	-1.367	148 (117)	90.935	136 (115)	64.368	78.217
GCSE Maths UMS score	Training provider	-9.540	3.206	** (**)[b]	312.004	28 (136)	48.947	75.253

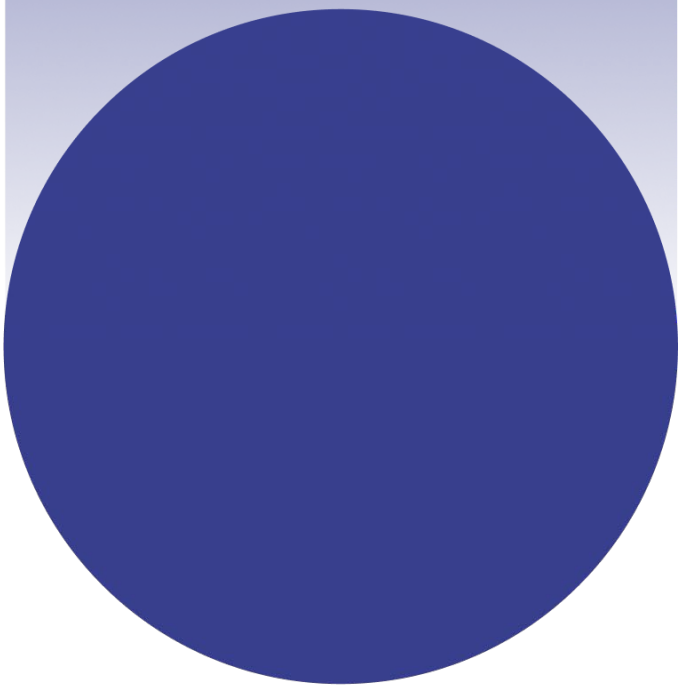
[a] *For the calculation of the Hedges' g in the interaction model by setting type we have used the variance in the treatment and control groups, which corresponds to the pooled SD in the primary outcome model.*

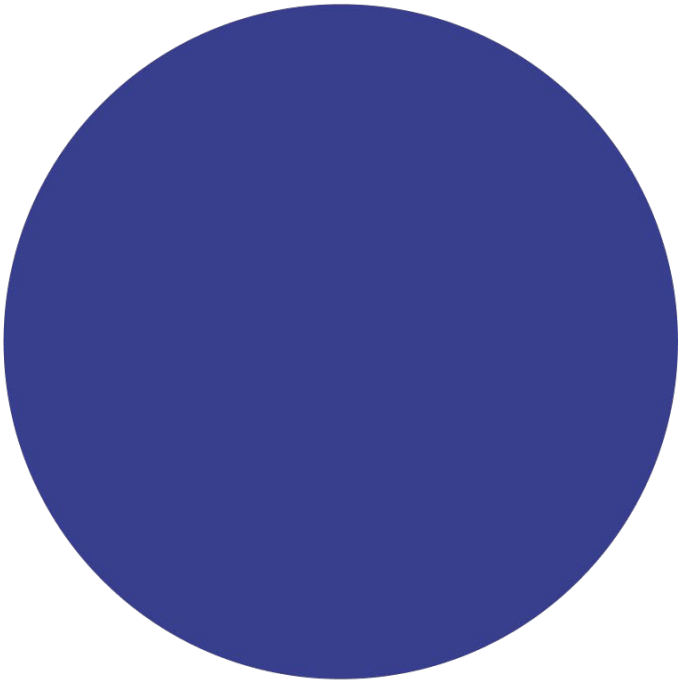
[b] *According to ONS disclosure rules, all counts under ten have not been reported.*

Further appendices:

You can find the further documents published in the accompanying document *'Further Appendices'*.

Appendix E: Intervention description
Appendix F: Programme resources
Appendix G: Memorandum of Understanding
Appendix H: Teacher information sheet
Appendix I: Student information sheets and withdrawal form
Appendix J: Logistic regression analysis for GCSE Maths pass/fail outcome
Appendix K: Lesson observation guide
Appendix L: PD observation guide
Appendix M: Lead teacher interview guide
Appendix N: Class teacher interview guide
Appendix O: Student interview guide
Appendix P: Histograms of pre-test results
Appendix Q: Technical note on how the UMS scores were obtained
Appendix R: Part E of the Teleprism survey
Appendix S: Deviations from the protocol
Appendix T: Analysis code (Stata)





Education Endowment Foundation

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

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn