Effectiveness trial of Mathematical Reasoning

Evaluation report

September 2026

Lillian Flemons, Chris Morton, Eleanor Bradley, Rachel Clutterbuck, and Helen Poet

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- to 19-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**. 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.

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

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

The project was independently evaluated by a team from the National Foundation for Educational Research. The trial director and principal investigator for this study was Helen Poet, Research Director. Lillian Flemons, Research Manager, led the evaluation team and the implementation and process evaluation (IPE). Andrew Smith, Senior Evaluation Analyst, and Chris Morton, Senior Statistician, led the impact evaluation. They were supported by Eleanor Bradley and Gustavo Lopes as IPE researchers, and Rachel Clutterbuck and Juan Manuel Pozo Segura as statisticians. Kathryn Hurd led the research operations with Elaine McCann and Katharine Stoodley as operations researchers.

Contact details:

Lillian Flemons
National Foundation for Educational Research,
The Mere,
Upton Park,
Slough, Berkshire
SL1 2DQ,
United Kingdom

Email: l.flemons@nfer.ac.uk

Acknowledgements

We are grateful to all of the primary schools that participated in this evaluation. We are especially thankful to the school staff members who completed the surveys, administered the baseline assessments, and co-ordinated the endpoint assessments with the National Foundation for Educational Research test administrators, and the staff and pupils who took part in the case study visits and interviews.

We would like to thank the delivery team, including Terezinha Nunes, Gabriel Stylianides, Louise Matthews, and Rossana Barros Baertl, for supporting the evaluation.

We would also like to thank the team at the Education Endowment Foundation, including Daniela Alvarado, Ben Sillitoe, and Amy Clark.

Office for National Statistics (ONS) Secure Research Service (SRS) publication status and statistical results

Executive summary

The project

Mathematical Reasoning is a programme designed by the University of Oxford to improve the maths attainment of Year 2 pupils (ages 6 to 7) by developing their understanding of the logical principles underlying maths. Through weekly, curriculum-aligned, structured lessons that replace one maths lesson, pupils learn and practice reasoning strategies for solving problem-based questions. Teachers receive training, expert support, and structured resources to strengthen their understanding of and ability to teach mathematical reasoning. The programme is delivered by the class teacher, supported by a teaching assistant (TA), over 12 to 15 weeks. Lessons involve both whole-class work and group work, combining teacher-led instruction with opportunities for pupils to reinforce mathematical reasoning with independent, collaborative, and digital learning activities. Teachers and TAs complete an online training course and attend three trainer-led webinars over the course of delivery. The programme was developed and refined by researchers at the University of Oxford through successive trial phases. Following smaller impacts in the first effectiveness trial, the training model was redesigned from in-person train-the-trainer delivery to an online model with webinar support, which was piloted in 2022/23.

The project was a two-armed randomised controlled effectiveness trial. It took place between December 2024 and June 2025 and involved 238 schools and 6,291 pupils. Schools were randomly allocated to either the programme or usual practice. Pupils were assessed at baseline by their teacher and at endpoint by an independent test administrator. The implementation and process evaluation (IPE) drew on teacher and TA surveys, programme records (including attendance logs and computer game data), lesson observations, pupil focus groups, and interviews with school staff and trainers, and observations of the online training and webinars.

Table 1: Key conclusions

Key conclusions	
1.	Pupils in Mathematical Reasoning schools made one additional month of progress in maths, on average, compared to pupils in other schools. This result has a high security rating.
2.	Among pupils eligible for Free School Meals (FSM), those in Mathematical Reasoning schools made two additional months of progress in maths, on average, compared to those in other schools. This result is less secure than the overall finding due to a smaller number of pupils.
3.	There was some cautious evidence that pupils with lower prior attainment benefited more from the programme than pupils with higher prior attainment. However, the IPE did not explain this finding, with teachers more likely to perceive the programme as better suited to higher attaining pupils.
4.	The IPE showed that the programme had a positive impact on self-reported teacher understanding of and confidence in teaching mathematical reasoning, and most teachers felt it had a positive impact on pupils' mathematical reasoning and arithmetic skills, as well as their confidence in and enjoyment of maths.
5.	Teachers reported that Teaching Assistant (TA) support was important for delivering group activities as intended. Although programme materials supported delivery, training and programme delivery required substantial time commitment, contributing to some issues with fidelity.

EEF security rating

These findings have high security rating. This was an effectiveness trial that tested whether the intervention worked under everyday conditions across a large number of schools. The trial was a well-designed, well-powered, two-arm, randomised controlled trial. The pupils in Mathematical Reasoning schools were similar to those in the comparison schools in terms of gender, Special Educational Needs and Disabilities (SEND) status and prior attainment. However, 14.6% of the pupils who started the trial could not be included in the final analysis, because they left the school or were absent on the final assessment day.

Additional findings

Pupils in Mathematical Reasoning schools made, on average, one additional month of progress compared to those in the control group. This is our best estimate of impact, which has a high security rating. As with any study, there is always some uncertainty around the result: the impact of this programme probably lies between zero and two months of additional progress. Our best estimate for pupils eligible for FSM is two months of additional progress, with the programme's impact probably lying between zero and three months of additional progress. Our analysis suggests that Mathematical Reasoning may have a larger effect of FSM-eligible pupils. However, this finding differs from the pattern observed in previous trials, which found similar effects for FSM-eligible pupils and their non-eligible peers.

Impacts appear greater for pupils with lower prior attainment and higher attendance. Exploratory analyses also suggest that higher engagement with a larger number and wider variety of computer games is associated with better outcomes, although the strength of evidence for these findings is more limited. Pupils in the programme also made greater progress, on average, in the Progress Test in Maths 7 (PTM7) mathematical reasoning subscale, compared to the other subscales.

Implementation varied across schools. Not all delivered the full content, and there was variation in key components such as pupil grouping, computer games, and use of manipulatives. Some schools reported staffing challenges, particularly around training cover and TA support.

This second effectiveness trial tested whether an online training model would improve fidelity and lead to impacts closer to those observed in the efficacy trial. In practice, it produced the same headline effect as the first effectiveness trial. IPE evidence suggests implementation challenges related to logistical and structural barriers (such as TA availability, time for preparation, and delivery), which remain consistent with previous findings.

Overall, there is some evidence that impacts can be achieved despite deviations in delivery. While greater fidelity may increase the size of impact, this cannot be evidenced at this stage. Furthermore, the time commitment required from teachers, alongside logistical and structural barriers (e.g. sustained availability of TAs), may make full implementation of the current model challenging for some schools.

Cost

The average cost of Mathematical Reasoning for one class is around £652 per year, or £19 per pupil per year when averaged over three years. The majority of this cost is for the training and programme materials. Prerequisite materials for programme delivery include a computer for staff use, a screen for use in the classroom, laptops or tablets for half of the class, and enough manipulatives for all pupils to use. The training and preparation time for the programme is ten hours for teachers' time and 7 hours for TAs, on top of the time required for session delivery, which occurs during usual lesson time.

Impact

Table 2: Summary of impact on primary outcome

Outcome / group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No. of pupils	P Value	EEF cost rating
Maths attainment (all pupils)	0.09 (0.00, 0.18)	1	🔒🔒🔒🔒 (4 of 5 locks)	5,373	0.052	£ £ £ £ £ (1 of 5)
Maths attainment (pupils eligible for FSM)	0.14 (0.01, 0.26)	2	🔒🔒🔒🔒 (4 of 5 locks)	1,424	0.025	£ £ £ £ £ (1 of 5)

Introduction

Background

Mathematical Reasoning (MR)) programme is designed for pupils in Year 2 that aims to improve mathematical attainment for all pupils by developing their understanding of the logical principles underlying mathematics. The programme was developed by Professor Terezinha Nunes and Professor Peter Bryant at the University of Oxford in response to their own research that found these two abilities at age eight and nine years predicted Key Stage 2 and 3 mathematical attainment (Nunes *et al.*, 2011). This project also built on research that found additive reasoning and logical abilities to be predictive of six-year-olds' mathematical attainment 12 to 16 months later (Nunes *et al.*, 2007; Ching and Nunes, 2017). Early number sense has likewise been found to predict later mathematics achievement (Jordan, Devlin, and Botello, 2022). The programme is based on the Key Stage 1 National Curriculum and introduces no new content. Mathematical Reasoning was not commercially available at the time of the trial.

The programme focuses on quantitative reasoning and number sense and replaces one lesson per week for 12 weeks with a programme session. Each session involves both whole-class teacher-led time and differentiated group time, alternating between tailored teacher support and time spent on programme-specific computer games to embed learning. Quantitative reasoning and number sense are distinct but related skills, and both have a role to play in learning arithmetic. The developer team defines quantitative reasoning as the ability to reason about quantities and relations between quantities, with or without numbers. Number sense is defined as the ability to reason about relations between numbers using the four operations. Within this latter domain, the programme focuses specifically on additive composition (i.e. any number can be seen as the sum of two other numbers) and the inverse relation between addition and subtraction (i.e. when a certain number is both added and taken away, the original amount remains the same) (Ching *et al.*, 2020). While the Key Stage 1 curriculum tends to teach arithmetic operations before applying these concepts to problem-solving, the MR programme seeks to provide pupils with the quantitative reasoning and number sense first as a foundation for problem-solving skills. The programme treats quantitative reasoning and number sense as complementary but separate strands of learning, reflecting previous research that has found that a child's understanding of quantities and their understanding of numbers are not always connected (Ching *et al.*, 2020).

Teachers and teaching assistants (TAs) are trained to deliver the approach through e-learning and professional development support. The programme aims to improve teacher and TA pedagogical knowledge around mathematical and numerical reasoning and to increase understanding of the importance of teaching these concepts and skills from a young age. While the core purpose of the Continuing Professional Development (CPD) is to train teachers and TAs to effectively deliver the Mathematical Reasoning programme, it is also intended to empower teachers and TAs to apply their learning from the programme to other areas of their work and to share it with their colleagues. In this way, pupils not directly involved in the programme may still benefit from the increased understanding and improved practice of teaching staff in these areas. The CPD element is delivered as an online training programme made up of five core modules and four webinars accompanied by tailored implementation support from trained professionals.

Previous trials

The Mathematical Reasoning programme has been the subject of two randomised controlled trials (RCTs) commissioned by the Education Endowment Foundation (EEF), both with a high security rating.

The efficacy trial (Worth *et al.*, 2015) was a three-arm RCT conducted by the National Foundation for Educational Research (NFER), with 17 out of 55 schools allocated to the MR group. In that efficacy trial, the Oxford University team trained teachers directly through a one-day in-person training session, with one follow-up visit to each participating school to observe delivery and provide personalised feedback. The trial found that the programme had a positive impact on pupils' numeracy abilities, equating to, on average, three months' progress (effect size of 0.20), compared to pupils who had not received the programme. A slightly smaller impact of two months' progress was found for pupils eligible for Free School Meals (FSM) (effect size of 0.14).

The subsequent effectiveness trial (Stokes *et al.*, 2018), carried out by the National Institute of Economic and Social Research (NIESR), was an RCT involving 160 schools. An in-person train-the-trainer model was employed for this trial. The National Centre for Excellence in the Teaching of Mathematics (NCETM) helped to develop the training model, which was delivered in person through the national network of 'Maths Hubs'. The trial found a smaller impact of one month's progress for all pupils and pupils eligible for FSM specifically (effect size of 0.08 and 0.09, respectively); however, these results were not statistically significant.

In response to the limited impact observed in the effectiveness compared to the efficacy trial, the Oxford University developer team led by Professor Gabriel Stylianides and Professor Terezinha Nunes, sought to improve the fidelity of delivery at a larger scale by developing a fully asynchronous online training course created by the programme developer to remove the intermediary effect of the train-the-trainer model. The new online training model, co-designed by Professor Terezinha Nunes and Professor Gabriel Stylianides, was piloted in 2022/2023 in 32 schools by the Institute of Education (IOE) at University College London (UCL), with the final report published in October 2024 (Sheldrake *et al.*, 2024). In the new online training model, additional tailored support over the course of the implementation period is provided to schools by Teacher Leaders (TLs) trained by the Oxford team via webinars, email, and an online forum. Other elements of the programme remained unchanged from previous iterations of the MR programme, including the structure and content of the sessions (including both teacher-led and differentiated group work) and the use of computer games. The pilot study sought to assess how effective the new training model was for preparing teachers to deliver the programme, as well as the feasibility of implementing this model and its scalability. Interviews, observations, and surveys were carried out to inform this assessment.

More information about the version of the MR programme being evaluated can be found in the 'Intervention' section below.

The current trial

Given the positive impact identified in the efficacy trial, the EEF was interested in determining how to implement the Mathematical Reasoning programme at a larger scale most effectively. The interim findings from a recent pilot study indicated positive results for the acceptability and feasibility of the new training model and its effectiveness in preparing teachers to deliver the programme. As a result, the EEF commissioned this second effectiveness trial to understand whether this new training model may better retain the scale of impact seen at the efficacy stage by enabling direct contact with the Oxford University team's teaching material. A summary of the differences in both the programme and evaluation for each of the trials can be found in Appendix C.

Intervention

The Mathematical Reasoning programme was implemented for the purpose of the effectiveness trial between December 2024 and June 2025.

A detailed description of the programme in the context of the Template for Intervention Description and Replication (TIDieR) checklist is presented below. A summary of the training and preparation carried out by TLs can be found in Appendix I.

Why: Rationale, theory, and/or goal of essential elements of the programme

The aim of the Mathematical Reasoning programme is to improve mathematical attainment by developing pupils' understanding of the logical principles underlying mathematics, primarily:

1. **Quantitative reasoning.** The ability to reason about quantities and relations between quantities (with or without numbers).
2. **Number sense.** The ability to reason about relations between numbers using the four operations, specifically focusing on additive composition and the inverse relation between addition and subtraction.

The programme does not introduce any new subject content outside of the national curriculum but instead seeks to improve reasoning and understanding of existing concepts through a teaching approach that emphasises quantitative reasoning and number sense as the foundation for problem-solving and arithmetic. The programme promotes discussion and the use of manipulatives (by both the teacher and the pupils) to support mathematical thinking.

The causal logic of the programme (see Figure 1) is based on previous research by the developers that found an association between pupils' quantitative reasoning and arithmetic abilities and their subsequent mathematical attainment, even from early primary school (Nunes *et al.*, 2007; Nunes *et al.,* 2011; Ching and Nunes, 2017). The programme also seeks to support the achievement of these outcomes by increasing pupil confidence and enjoyment around maths, as there is evidence to suggest that these factors may be predictors of subsequent mathematical attainment (Putwain *et al.*, 2018; Çiftçi and Yildiz, 2019).

Who: Recipients of the programme

The programme recipients are Year 2 pupils in state schools in England. The programme is delivered to the whole class, with reasonable adjustments made for pupils with Special Educational Needs and Disabilities (SEND) where necessary.

In most cases, trial schools registered for one class to participate in the trial, although a small number registered for two or three. In a small number of cases, schools that requested additional classes later in the process were permitted to deliver the programme to other Year 2 classes without having them participate in the trial, so as to manage evaluation costs.

What: Materials

Training for schools

All teachers and TAs are given access to an online training course comprising nine modules. Modules 1–5 prepare the teacher/TA pair to implement the programme, while modules 6–9 aim to show how mathematical reasoning can be used in teaching other mathematics topics in primary schools. Each module includes brief video lectures from a member of the Oxford University team explaining key ideas before the programme and implementation guidelines, as well as downloadable research briefs and presentation slides with further details about the programme and the theory and evidence behind it. The course is self-guided and includes interactive elements, opportunities for reflection, and videos demonstrating how different sessions should be delivered. Teachers and TAs can interact with their TL and other schools in their cohort (see below) via a chat function embedded in the course.

Programme delivery

Each school receives digital presentation slides for each of the 12 sessions, as well as a teaching handbook for the programme. This handbook includes a detailed unit plan for each session and instructions for each whole-class and group activity, as well as a glossary. A document with Frequency Asked Questions (FAQs) is also provided.

As part of the programme, each pupil receives a Pupil Workbook, which includes written activities and extension worksheets for the teacher-led components, as well as cut-out shapes to be used as manipulatives.

Schools are also provided with printable worksheets with supplementary games that pupils can complete before starting on the computer games if they are not yet at the ability level required for the first online game.

Schools are expected to provide the necessary IT equipment, including a screen for presenting the slides during the whole-class component of the sessions. In addition, schools are expected to provide additional manipulatives for pupil use, such as counters, blocks, and coins.

Computer games

In the second part of the Mathematical Reasoning session, around half of the class is allocated to play computer games. These games were created by the University of Oxford for this programme as a tool for pupils to practice their Mathematical Reasoning skills. Children access the games through a website. They are assigned an individual log-in so that progress can be tracked and achievements rewarded. Schools are given access to and instructions for the computer games website, which includes bonus games and certificates of achievement that can be downloaded and printed separately. Each pupil receives a Pupil Record Sheet to track the games they play, which is included in the Pupil Workbook. Schools must provide access to computers or tablets for pupils to play computer games during the group component of the sessions. There is no minimum number of computers or tablets required—this is up to the capacity and discretion of each school.

What: Procedures, activities, and/or processes

Training for schools

The teacher and a TA for each participating class are expected to complete the five core online modules before starting programme delivery. Modules 1–4 cover the theory and rationale and provide an overview of the programme. Module 5 provides practical guidance for implementing the programme as well as highlighting the importance of fidelity and the kinds of adaptations that would be acceptable. The TL provides each teacher/TA pair with access to module 5 once they judge them to have satisfactorily completed modules 1–4 based on their responses to the activities in each of the modules. These activities are not intended to be a knowledge assessment but instead require the respondents to share reflections on what they have learned.

The training course includes an additional four online modules (modules 6–9) that offer teachers and TAs the opportunity to learn more about how mathematical reasoning can be promoted when teaching other topics in the maths curriculum (such as fractions and diagrams), including for older age groups. Only the last of these modules (module 9) is compulsory. It covers how participants can use what they have learned in other areas of their teaching practice and share this knowledge with other teachers in the school. Teachers and TAs are expected to complete this module ahead of the final webinar (see below).

The TLs monitor the online course and provide support via the discussion forums and chat function in the online training course and by email. Each TL supports a separate cohort of around 20 schools. There are no criteria for allocating schools to particular cohorts or TLs. Online interaction is restricted to within each cohort.

Each teacher/TA pair receives one set of log-in details, and the teacher and TA are encouraged to complete the training course together to promote knowledge sharing and discussion.

In addition to the online training, teachers and TAs are expected to attend three live webinars delivered by their TL over the delivery period. Each TL delivers a webinar attended by their own cohort only. The materials for these webinars were designed by the Oxford University team with a focus on participant-led discussion. TLs are able to make minor adaptations to meet the needs of their cohort. The webinars seek to:

- provide support to teachers and TAs in tackling practical challenges that may arise at any phase of the programme;
- create a community of practice for reflective thinking and peer learning, with sharing between schools; and
- support integration of learning into practice beyond the programme.

The first webinar provides the theoretical background to the programme and outlines each of its key components. The second webinar focuses on knowledge sharing between schools and introduces some of the activities that feature in later sessions. The third webinar looks at the sustainability and future use of the programme in the school and how it can apply to other year groups and/or areas of the curriculum.

Finally, TLs also provide schools with practical and administrative support.

Programme delivery

The programme consists of 12 units delivered by the teacher, with TA support, across 12–15 sessions. Each session comprises a whole-class component and a group component. For the group component, the teacher divides the class into two groups: L1 and L2. L1 consists of pupils for whom the teacher feels additional support could be beneficial, while L2 is for pupils perceived by the teacher to be ready for further learning. These groupings are intended to be flexible according to perceived pupil needs in response to each topic covered on a session-by-session basis. During the group work component of each session, one group completes activities with the teacher, while the other works on the computer games. The groups alternate by unit between the computer games and teacher-led work, as specified in the handbook.

The programme consists of two conceptual strands (number sense and quantitative reasoning). The activities that explore each of these are intended to be interleaved such that the pupils' skills in each develop in tandem. The number sense strand focuses on the additive composition of numbers, place value, and inverse relations between addition and subtraction. Pupils are encouraged to think about the changes caused by performing and undoing actions, not just by counting forward or backwards. The quantitative reasoning strand focuses on the different relations that can be established between quantities and asks pupils to visualise relational scenarios using manipulatives. The programme starts by working with smaller numbers before progressing to larger numbers.

The first part of each session involves the whole-class component led by the teacher and is expected to last around 40 minutes. In each session, the teacher introduces the (new) concept, often with manipulatives and an animated presentation, and presents the class with story problems to solve. The pupils each write an answer in their Pupil Workbook and discuss it in pairs, using manipulatives to inform their working out by enacting the story problems. The teacher then talks through the work, asking the pupils to explain the thinking and processes behind their answers. There are extension exercises at the back of the Pupil Workbook for pupils who finish an activity early to complete while waiting for the rest of the class. During this session, the TA makes sure pupils have the materials they need and are answering the questions in their workbooks. The TA also supports pupils to turn to the extension activities and/or to make reasonable adjustments to the activities according to pupil needs.

The remainder of each session is spent in group activities. One group works with the teacher, who provides extra support or pre-teaching for L1 and extension opportunities for L2. The handbook provides detailed instructions for the activities to be carried out in the teacher-led group component for each unit and whether the activities should be with L1 or L2 in each case.

The other group plays computer games, which provide pupils with the opportunity to practice the concepts taught in the whole-class session to consolidate their learning. The games are divided into units that reflect the structure and content of the lessons. However, pupils can work through the units at their own pace. Pupils record the games they have completed on their individual Child Record Sheets. The TA supervises the group playing the computer games, helping pupils to log on and complete their Child Record Sheets, showing them how different games work and encouraging them to play a variety of games. Pupils who achieve 100% on a game receive a certificate (shared with the pupil at the TA/teacher's discretion) and the opportunity to play a short bonus game. The computer games can also be accessed from home, and schools are provided with guidance on how to facilitate this.

There are supplementary activities for any pupils who are not yet at the ability level required to play the first set of additive composition computer games.

Who: Programme providers/implementers

The developer team at the University of Oxford provides all the material and trains the TLs.

The TLs support teachers in completing the training and delivering the programme. They are specialist teachers trained in the programme by the developers and have extensive experience in training other teachers to implement programmes and/or have delivered the programme as a teacher at least twice in the past.

Nominated Year 2 class teachers are responsible for delivering the teacher-led components of the programme (whole class and group).

Nominated TAs are responsible for supporting the teachers in delivering the programme, including the whole-class components, and monitoring the use of computer games during the group components of the sessions.

How: Mode of delivery

Teacher and TA training is online via a training course, webinars, and a forum.
The programme is delivered to pupils in person. Some of the pupils play games online for part of each session.

See the 'What: Materials' and 'What: Procedures, activities, and/or processes' sections above for further details.

Where: Location of the programme

The programme is intended to be delivered in the regular classrooms of participating schools. It is recommended that pupils sit on the carpet for the whole-class component, but this is up to the teacher's discretion, as long as the pupils are situated so as to be focused on the teacher and questions on the screen at the front and are able to engage in discussions with the whole class.

Schools can be located anywhere in England for the purpose of the trial.

When and how much: Duration and dosage

Training for schools

Modules 1–5 can be completed at times convenient to the teachers within a designated period to create a learning community of teachers and to ensure they are completed in good time. Headteachers are asked to release teachers and TAs for one and a half days to complete modules 1–5 plus module 9 and to set the pupils up on the computer games. Ideally this time is allocated in half-day blocks, rather than short sessions across a longer period.

The first webinar takes place once schools have completed modules 1–5 but before they start delivery, the second webinar takes place two to three weeks into the delivery period, and the third webinar takes place midway through the delivery period.

Programme delivery

The programme consists of 12 units delivered weekly across 12–15 weeks, allowing for flexibility to deliver some units across more than one session if needed. Each session should last approximately 60 minutes (or the length of a normal lesson) and take place during normal maths lesson time. Approximately 40 minutes of each session should be spent on the whole-class component and 20 minutes on the group activities. If the teacher is not able to complete the activities from a whole-class lesson within 40 minutes, they should still move on to the group component of the session and start the whole-class session in the next lesson at the point they stopped.

While in practice the programme can be delivered in classrooms at any point after completing the training, for the purpose of the trial, it was delivered across the Autumn Term and Spring Term (following randomisation, baseline testing, and training completion). Endpoint testing occurred in the Summer Term.

Tailoring: Adaptation of the programme

High fidelity to the course structure, material, and content is required. Teachers are told not to skip any of the content and to continue the next session where they left off if a whole unit cannot be covered in a single session.

Duration and timing of different aspects of the programme are more flexible, including when teachers and TAs complete the training, how long sessions last, how long is spent on each component of the session, and how many weeks the programme lasts. The proportion of the class assigned to each of the groups (L1/L2) is also flexible, depending on perceived pupil need.

Teachers are encouraged to make any necessary adjustments for pupils with SEND, as they would in a normal lesson. They should also tailor the pace and approach of each teacher-led group component to the needs of that particular group.

Teachers were also advised of some adjustments to the session schedule that they could make if required, for example, doing two sessions in a week if they had to miss a session, or repeating the previous session if two weeks had been missed.

However, more significant adaptations to the programme were considered deviations from intended practice, including:

- changing the order in which the programme content is delivered;
- starting from session 5 because the early sessions seem too easy;
- replacing inverse relation tasks with practice in number facts;

- skipping the whole-class discussion to catch-up when behind or to go faster during the session;
- skipping the group component of a session (which would mean skipping the computer games) or alternating between whole-class and group sessions; and
- using larger numbers in the activities right from the beginning because the pupils already know how to count to 100.

How well (planned): Strategies to maximise effective implementation

Teachers are provided with a list of 'keys to success' to optimise for effective implementation (see Appendix C). These include what to do should the teacher be unable to deliver all the content for each unit within the space of a single session —something which had emerged as an issue in the previous trials.

A section of the online training course explicitly emphasises the importance of fidelity and the kinds of adaptations to the programme that may or may not be acceptable.

TLs are provided with training (see Appendix D) to carry out their role, including supporting schools with any queries or challenges in delivering the programme. Schools are also provided with numerous different opportunities and forums for raising questions or concerns with their TL, including online forums, webinars, and email.

Both teachers and TAs are asked to complete the training together, with the aim of promoting discussion about the programme and consequently reducing the risk of misunderstandings about how it should be delivered. The developers strongly recommend that schools use TA support to facilitate the group work component of the sessions.

Teachers are provided with a self-evaluation form and encouraged to use it to monitor their implementation once they are midway through delivery.

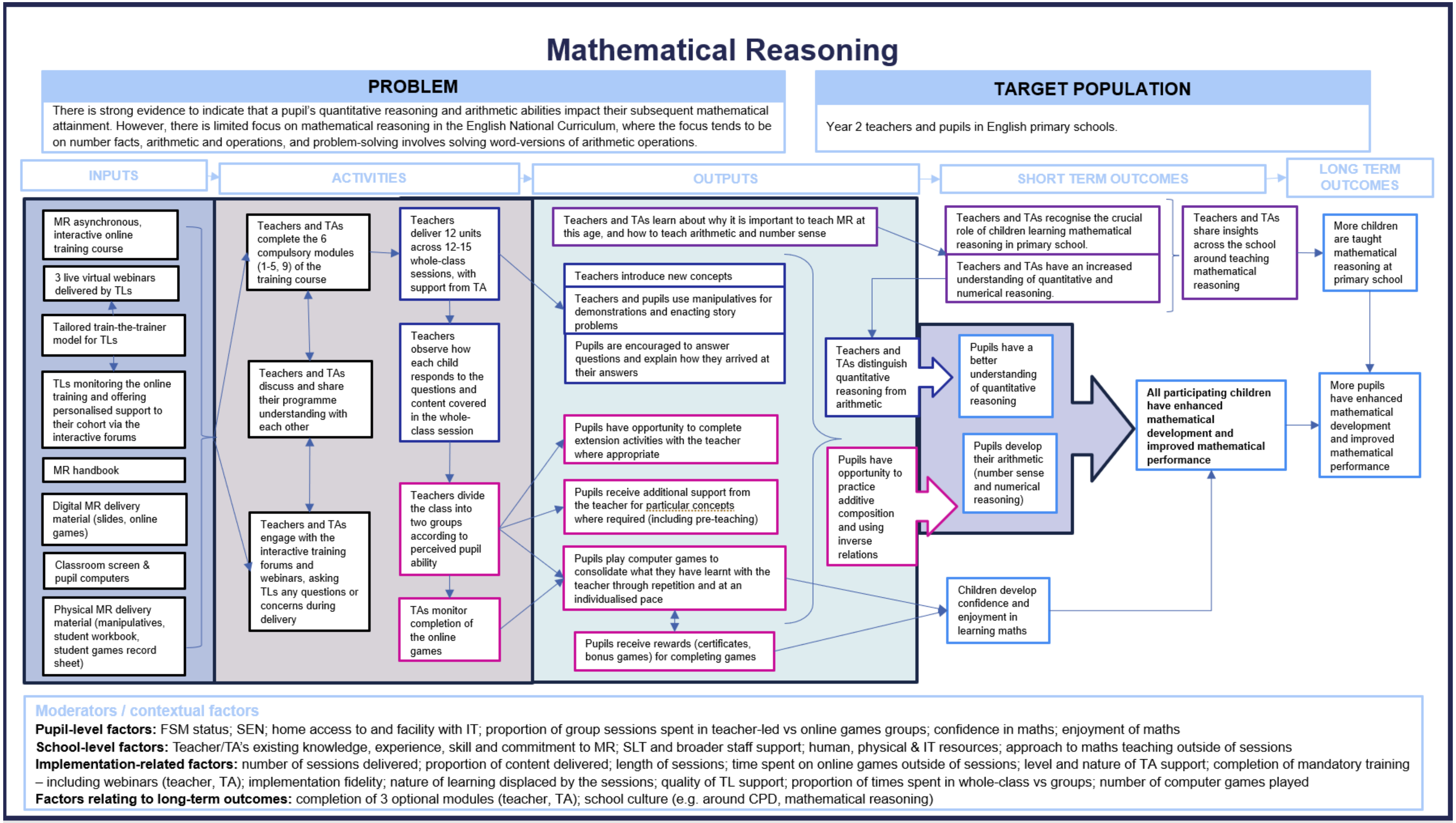
How well (actual, based on previous trials): Evidence of implementation variability

Significant variation in the proportion of the programme content delivered was observed in the efficacy trial (Worth *et al.*, 2015). For example, only 15 out of the 22 participating teachers in that trial reported that they had delivered all of the programme units. This issue appears to have been less prevalent in the first effectiveness trial, however, with 49 out of 52 teachers reporting having delivered all ten units (Stokes *et al.*, 2018). Significant variation in the use of computer games was, however, observed across both the efficacy and the first effectiveness trials, with a fifth of pupils in the latter found to have played no games at all (Stokes *et al.*, 2018). Some variation in grouping practices and use of extension activities was also observed in the first effectiveness trial (Stokes *et al.*, 2018).

Theory of Change

The protocol version of the Theory of Change for the Mathematical Reasoning programme is shown in Figure 1. This was developed at the start of the trial in consultation with the delivery team and informed the trial design. It outlines the target population (all Year 2 pupils) and the activities, outputs, and short-term outcomes that are intended to ultimately lead to more pupils having enhanced mathematical development and improved mathematical performance. An additional strand of outputs and outcomes reflects the intended trajectory for teacher and TA learning and broader impact.

Figure 1: Theory of Change for the Mathematical Reasoning programme



While an initial distinction has been drawn between the whole-class and group activities to clarify that they are distinct programme components, the model articulates the manner in which the whole-class session informs group allocation, as well as their distinct but complementary contribution to the same set of outcomes. Similarly, while quantitative reasoning and number sense are treated within the programme as separate strands and developed as distinct abilities, they are positioned as complementary skills feeding into broader mathematical outcomes.

While Mathematical Reasoning is a scripted programme, it encourages teachers to be proactive in tailoring questions and support to the children's specific needs. It also relies on teachers employing a range of teaching techniques, including scaffolding and the use of manipulatives. This means that the Theory of Change relies on teachers having the necessary existing skills to deliver the quality of teaching required following training completion. In addition to supporting these skills, the training must also be of sufficient quality to develop the teacher and TA's understanding of what quantitative reasoning (as compared to arithmetic) and what the use of inverse relations look like in practice, as the programme relies on supporting children to approach questions in a particular way, not simply to obtain the right answer.

Given that the programme sessions are intended to replace normal lessons, the Theory of Change relies on the assumption that the learning pupils receive through the programme will be more beneficial for their mathematical development than the learning they would otherwise have received through their usual lessons.

The role of differentiated teaching within the programme relies on teachers being able to correctly identify whether a pupil would most benefit from extension activities or additional support in each lesson. The effectiveness of this approach also depends on pupils in both groups drawing equal benefit from tailored teacher-led support.

Technology has a central role in the programme, which relies on schools having the necessary technology available. Moreover, the number of individual devices available would need to be sufficient so as not to influence the number of children allocated to each group. Internet access within the school would also have to be of sufficient quality to allow this number of children to access the internet at the same time.

Finally, to achieve the programme's intended long-term outcomes, teachers and TAs must have the opportunity and means to share their learning within the school community, and other staff members must be willing and able to take this on board and integrate it into their own practice.

Evaluation objectives

The research questions (RQs) for the impact evaluation and the implementation and process evaluation (IPE) are provided below. Further detail on the evaluation design can be found in the **study protocol** (Flemons *et al.*, 2024) and the **Statistical Analysis Plan** (SAP) (Morton and Smith, 2025).

Primary impact RQ

> **RQ1:** What is the impact of the Mathematical Reasoning programme on Year 2 pupils' attainment in mathematics (measured using GL Assessment Progress Test in Maths [PTM] 7)?

Secondary impact RQs

> **RQ2.1:** What is the impact of the Mathematical Reasoning programme on Year 2 FSM-eligible pupils' attainment in mathematics (measured using GL Assessment PTM7)?
> **RQ2.2:** What is the impact of the Mathematical Reasoning programme on each of the PTM7 'process' categories (subscales): i) fluency in facts and procedures; ii) fluency in conceptual understanding; iii) problem-solving; and iv) mathematical reasoning (measured using GL Assessment PTM7)?
> (a) For all pupils.
> (b) For FSM-eligible pupils.
> **RQ2.3:** How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by the number of sessions attended by the pupil?
> (a) For all pupils.
> (b) For FSM-eligible pupils.

RQ2.4: How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by: i) the number of computer games played by the pupil; and ii) the number of different computer games played by the pupil?

(a) For all pupils.
(b) For FSM-eligible pupils.

RQ2.5: How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by pupils' prior attainment?

(a) For all pupils;
(b) For FSM-eligible pupils.

RQ2.6: How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by whether their teacher completed the training?

(a) For all pupils.
(b) For FSM-eligible pupils.

IPE RQs

IPE RQ1: To what extent was the Mathematical Reasoning programme delivered as intended at scale? (Fidelity, adaptation, reach)

IPE RQ2: What are the key moderators and contextual factors that influenced how effectively the programme was delivered? (Context, moderators, dosage)

IPE RQ3: How did pupils' experiences of the programme vary depending on pupil characteristics? (Context, moderators, dosage)

IPE RQ4: What was the perceived impact of the Mathematical Reasoning programme for: i) staff members; ii) pupils; and iii) disadvantaged pupils specifically? (Perceived impact)

IPE RQ5: What was business as usual (BAU) in relation to maths teaching, and to what extent did it differ from the Mathematical Reasoning programme? (Programme differentiation)

IPE RQ6: How much does it cost a school to deliver the Mathematical Reasoning programme? (Cost)

Issues arising

While only two schools withdrew fully from the trial and did not complete the endpoint assessment, a further 11 schools formally withdrew from delivering the programme after having delivered none or very little of it. The reasons for withdrawal included staff absence, staffing changes, participation in other programmes or trials, the trial being deemed unsuitable for school requirements, and dissatisfaction with not having been selected as one of the schools permitted to have more than one class participate in the trial.

The other key issue that emerged during the trial was in relation to the tracking of pupil use of the computer games. In order to collect data for the trial, schools were provided with an allocated username for each pupil via their Session Delivery Log. However, this is not the procedure the programme uses outside of the trial and therefore, departed from the instructions on the website. Some teachers created their own usernames for pupils in their class (i.e. following the programme website instructions) instead of using the allocated ones. This was related to the timing of when these were shared (to align with the launch event) and that the Session Delivery Logs were shared with the key contact rather than directly with the teacher meant that teachers were not always aware of these allocated usernames when they came to creating the profiles for their pupils. This resulted in some confusion and ultimately compromised the amount and quality of the computer games data that could be retrieved for analysis. While significant efforts were made to collect the usernames that were used for each child in place of the allocated usernames, these were ultimately not provided by every school.

Ethics and trial registration

An ethical review was undertaken as part of the NFER start-up meeting in November 2023 where consideration was given to consent and the impact of the research on trial participants (pupils and practitioners). This evaluation was conducted in accordance with NFER's Code of Practice.

Each participating school's headteacher provided their agreement on behalf of the school to participate in the trial by signing the Memorandum of Understanding (MoU) (see Appendix J in Further Appendices), which outlines the responsibilities of all parties involved in the trial.

NFER shared a parent/carer letter and withdrawal form with schools who then sent it to parents/carers of all pupils in participating classes. Through the withdrawal form, parents/carers had the opportunity to withdraw their child from the evaluation and associated data processing at any stage of the trial.

A separate opt-in consent process was used for pupils selected to take part in the pupil focus groups. Given that pupils participating in this study are only six to seven years old, we could not assume that all pupils had the capacity to provide fully informed consent to participate. As a result, we provided parents/carers with a written information sheet about the focus groups containing full details about the focus group and what their child would be asked to do. Parents/carers completed the opt-in consent form for their child to be invited to participate in the focus group, who passed this information on to the research team.

Pupil participation in the focus groups is voluntary, therefore even if a parent/carer had given consent for their child to participant, their child could still choose not to take part. Age-appropriate information about the focus groups was provided to pupils at the same time as parents/carers, to allow them to discuss participation together. The researchers also read this information to pupils at the beginning of the focus group to ensure pupils understood it and had the chance to ask any questions. If, at this point, a pupil decided that they would prefer not to participate, then they could return to their class. Prior to beginning the focus group, the researchers agreed some ground rules for the group with the pupils and had a discussion with them about the types of scenarios in which we would need to break confidentiality, to ensure they fully understood what this means.

Interviewees (e.g. school staff) were provided with information about the research and how we use their data before our visit and informed consent will be obtained from interviewees at the start of the interview. If an individual staff member within a case study school did not wish to participate in the data collection, they could choose to decline.

This trial is designed, conducted, and reported to the Consolidated Standards of Reporting Trials (CONSORT) standards. It was registered in the ISRCTN (International Standard Randomised Controlled Trial Number) Registry after the protocol was finalised (reference ISRCTN18042440).

Data protection

Personal data was processed as part of this trial. All data gathered during the trial was held in accordance with the data protection framework created by the Data Protection Act 2018 (*Data Protection Act 2018, C. 12*, 2018) and the General Data Protection Regulation (GDPR) 2016/679 (*Regulation (EU) 2016/679 of the European Parliament and of the Council*, 2016) and treated in the strictest confidence by the NFER, the Oxford University team, and the EEF. No individual or school will be identified in any report.

NFER is the data controller for evaluation and made decisions about how and what personal data was used in accordance with the objectives of the study set by EEF. The University of Oxford is the data processor for the evaluation and data controller for the programme delivery.

The legal basis for processing personal data is covered by GDPR Article 6 (1) (f):

> Legitimate interests: The processing is necessary for your (or a third party's) legitimate interests unless there is a good reason to protect the individual's personal data, which overrides those legitimate interests (GDPR, 2016).

A legitimate interest assessment has been undertaken. The evaluation fulfils one of NFER's core business purposes (undertaking research, evaluation, and information activities). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence about the impact of teaching techniques used in the classroom—in this case, the teaching of mathematical reasoning in the primary school classroom, and the relative effectiveness of the Mathematical Reasoning programme specifically. The legal basis for processing pupils' special personal data (SEND status) is covered by GDPR Article 9 (2) (j) which states that:

> 'processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) (as supplemented by section 19 of the 2018 Act) based on domestic law which shall be proportionate to the aim pursued, respect the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject (GDPR, 2016)'.

We do not believe this processing will cause damage or distress to the pupils. The outcomes of the evaluation will not result in the creation of measures or decisions being made about individual pupils.

Prior to any data sharing, NFER and the Oxford University team signed a Data Sharing Agreement (DSA) to govern the collection and sharing of personal data during this trial. This DSA included a description of the nature of the data being collected and how it will be shared, stored, protected, and reported by each party. In addition, the Oxford University team provided an MoU to schools, explaining the nature of the data being requested of schools, teachers, and pupils, how it would be collected, and how it would be passed to and shared with NFER. Two separate Privacy Notices were available: one for schools, and another one for parents/carers (see Appendices L and N in Further Appendices). All personal data was stored securely and shared via secure, password-protected data sharing portals.

The full name and contact details of the headteacher were collected when a school signed up to participate in the project. As part of the MoU, the full name, contact details, and job role was collected for the key contact on the project, who facilitated communications between the school and the Oxford University team, and the school and NFER. The full name, role, and contact details for the class teacher and TA in each participating class, was collected for the evaluation to facilitate contact with participants. The names and contact details were collected to enable both the Oxford University and NFER teams to make 'thank you' payments to the schools. The Oxford University team shared pseudonymised teacher and TA responses to questions asked as part of the online training course with NFER so that NFER could analyse rates of training completion. The surveys and interviews asked about staff experiences of delivering the programme, challenges, and facilitators relevant to their school context and self-reported knowledge about and attitudes towards mathematical reasoning.

NFER also collected pupil data from schools including names, date of birth, Unique Pupil Number (UPN), FSM eligibility status, and baseline and endpoint assessment scores for all pupils in the participating classes. NFER shared pupil personal data with GL Assessment to enable the assessment marking process. For these pupils, background data including gender, FSM eligibility, English as an Additional Language (EAL) status, and SEND status was collected from the National Pupil Database (NPD). To obtain the information from the NPD, NFER securely provided the Data Sharing Team at the Department for Education (DfE) with the names of the pupils, their dates of birth, and UPNs, allowing a match to NPD. Pupil perspectives on the programme and maths in general were collected as part of pupil focus groups.

Within three months of the end of project, NFER will send school data and pupil data to EEF's data archive partner. At this point, EEF's data archive partner will keep a copy of the data and EEF will become the Data Controller. NFER will retain personal data for one year after report publication in case there are any queries about the report. One year after the report publication, all personal data will be securely deleted.

After the evaluation report has been published, NFER will share anonymised pupil data with the Oxford University team, who will then match this data with pseudonymised teacher data and school identifications (IDs) that they will collect and hold.
NFER will not store or transfer any data outside of the UK. When we use Questback to administer online surveys, data is stored in the European Union. GL Assessment may transfer personal data outside of the UK. However, this is safeguarded by the appropriate contractual safeguards.

Project team

Table 3: Project team and their roles and responsibilities

Name	Organisation	Role and Responsibilities
Evaluation team		
Helen Poet	NFER	Project Director: Responsible for overall delivery of the evaluation to agreed specifications and overseeing the integration of the impact and IPE. Responsible for strategic oversight and the quality of the outputs
Lillian Flemons	NFER	Trial Manager and IPE Lead: Responsible for day-to-day management of the trial, and design and delivery of the IPE
Andrew Smith	NFER	Impact Evaluation Design Lead: Responsible for the design of the trial and analytical considerations, responsible for the protocol, study plan, and quality assurance of the impact analysis
Chris Morton	NFER	Senior Statistician: Responsible for contributing to the analytical design and running the quantitative analyses for the impact and IPE strands. Reporting lead for the 'Impact evaluation results' section
Eleanor Bradley	NFER	IPE researcher: Responsible for IPE data collection, analysis, and reporting
Gustavo Lopes	NFER	IPE researcher: Responsible for IPE data collection
Kathryn Hurd	NFER	Research Operations Lead: Responsible for leadership and strategy around data collection and school communications
Katharine Stoodley	NFER	Operations Manager: Responsible for day-to-day operations, including preparation of recruitment documents, co-ordinating data collection and point of contact for schools participating in the trial
Delivery team		
Professor Gabriel Stylianides	University of Oxford	Principal Investigator and co-designer of the online professional development training for teachers: Responsible for strategic oversight of, and contributor to, all aspects of delivery and related outputs
Professor Terezinha Nunes	University of Oxford	Programme designer, co-designer of the online professional development training for teachers and the principal investigator: Responsible for contributing to all aspects of delivery, including academic, technical, and administrative, and to writing and reviewing documents for programme implementation and evaluation
Louise Matthews	University of Oxford	Research Project Manager. Responsible for the day-to-day management of the delivery, including the professional development of the TLs
Dr Rossana Barros	University of Oxford	Research Officer. Responsible for IT management and contributor to academic, technical, and administrative aspects of the trial

Methods

Trial design

This was an effectiveness randomised controlled trial, in which schools were randomised to either the intervention or control arm. The trial cohort were Year 2 pupils in participating classes. Schools with more than one Year 2 class could choose to nominate more than one class for the trial. However, to balance sample size and cost, only some schools were permitted to have more than one class. The teacher(s) of the nominated Year 2 class(es) would then receive the Mathematical Reasoning training if randomised to the intervention. All trial pupils in intervention schools were eligible to receive the Mathematical Reasoning intervention, with intervention sessions being delivered to whole classes.

Control schools continued with 'business as usual'. Control schools were not permitted to access the Mathematical Reasoning programme. However, as this is an effectiveness trial and hence concerned with 'real-world' practice, control schools were permitted to deliver other maths interventions and forms of support (as decided by their school), including in relation to mathematical reasoning, as they would have otherwise done. The control schools received a payment of £500 in recognition of the contribution to this research upon completion of the endpoint testing. The primary outcome for the trial was total score from the PTM7, a maths assessment developed by GL Assessment (see Table 4 and the 'Outcome measures' section below). Four subscales from the PTM7 were used as secondary outcomes.

Table 4: Trial design

Trial design, including number of arms		Two-arm cluster randomised controlled trial with random allocation at school level
Unit of randomisation		Schools
Stratification variable (s) (if applicable)		None
Primary outcome	Variable	Maths attainment
	Measure (instrument, scale, source)	Progress Test in Maths 7 (PTM7) (test administered by NFER test administrators, 0-43 raw score, GL Assessment)
Secondary outcome(s)	Variable(s)	Maths attainment (specific domains)
	Measure(s) (instrument, scale, source)	Progress Test in Maths 7 (PTM7) 'process' categories (subscales): i. fluency in facts and procedures (0-6 raw score) ii. fluency in conceptual understanding (0-13 raw score) iii. problem-solving (0-4 raw score) iv. mathematical reasoning (0-20 raw score).
Baseline for primary outcome	Variable	Maths attainment
	Measure (instrument, scale, source)	Progress Test in Maths 6 (PTM6) (test administered by teachers, 0-31 raw score, GL Assessment)
Baseline for secondary outcome(s)	Variable	Maths attainment (specific domains)

	Measure (instrument, scale, source)	Progress Test in Maths 6 (PTM6) 'process' categories (subscales): i. fluency in facts and procedures (0-2 raw score) ii. fluency in conceptual understanding (0-12 raw score) iii. problem-solving (0-3 raw score) iv. mathematical reasoning (0-14 raw score)

Participant selection

All state primary schools in England were eligible to participate in this evaluation, excluding those participating in the Maths-Whizz and Maths Mastery trials or who participated in the 2023 pilot study of the new Mathematical Reasoning training model (Mujtaba and Sheldrake, 2023). Recruitment was nationwide and did not prioritise specific regions. In total 238 schools were randomised and included in the trial (119 intervention, 119 control).

TLs were primarily responsible for recruitment. They approached schools via both promotional and personalised emails, events in the TLs' local areas, flyer distribution at other relevant events, and social media. The trial was also advertised on the websites of the EEF, NFER, and the University of Oxford, as well as the Reasoning First Website,[1] which is maintained by the Oxford University team. Once a school joined the project, teachers and TAs were nominated by the school in accordance with the procedures indicated in an MoU.

It was expected that most schools would choose to enter one class into the evaluation, but schools could request to enter more than one. In addition, while most classes were expected to comprise Year 2 pupils only (the target group for the programme), mixed year group classes covering the target year group (e.g. Years 1 and 2, Years 2 and 3) were also eligible to participate. This was to ensure that small schools (e.g. in rural areas) had the opportunity to participate. While pupils from year groups other than Year 2 could participate in the programme, only outcomes for Year 2 pupils were assessed.

All pupils in participating classes were included in the trial: there were no pupil-level eligibility criteria. Pupils completed a baseline assessment (the PTM6, see 'Outcome measures' section below) to control for prior attainment and increase the precision of the analysis, but baseline assessment scores were not used to determine pupil eligibility for the trial. At the point of randomisation, 6,291 pupils were included in the trial (3,281 intervention, 3,010 control).

Outcome measures

Primary outcome

The primary outcome for the trial was maths attainment, measured using overall raw scores from GL Assessment's PTM7 (Form A). The choice of 'maths attainment' for the primary outcome was consistent with the Theory of Change, which includes 'improved mathematical performance' as a short-term pupil outcome. GL Assessment's PTM offered a suitable measure for the primary outcome of mathematical performance, particularly as it aligns with national curriculum objectives in English schools. Furthermore, the use of PTM7 was consistent with the previous effectiveness trial, thus increasing comparability with previous evaluations of the programme.[2]

For this trial, we expected raw scores to be less vulnerable to floor and ceiling effects than age-standardised scores.[3] The choice of raw score rather than age-standardised score also made the primary and secondary outcomes more comparable (age-standardised scores were not available for the PTM7 subscales). There was no adjustment for pupil age, either through standardisation or by including age as a model covariate.

[1] The Mathematical Reasoning programme website, see: **https://reasoningfirst.org.uk/**
[2] The efficacy trial used Progress in Maths 7 (PiM7), the forerunner to PTM7. GL Assessment report the correlation between the PiM7 and PTM7 to be 0.8, based on a sample of 350 pupils.
[3] This was based on inspection of the distribution of baseline PTM6 raw and age-standardised scores for this trial, as well as comparing Figures 3 and G.1 in the previous effectiveness trial.

Multiple versions of the PTM assessment are available, representing different levels (i.e. year groups). This means it can be used to measure progress within a single year or across multiple years, thus making it suitable for both a baseline and outcome measure in this study. GL Assessment's tests are designed to measure progress using different test versions, rather than by repeated administration of the same questions. The PTM questions were developed by the Mathematics Assessment Resource Service (MARS) team, which is a collaboration between the University of California, Berkeley, and the Shell Centre, Nottingham. Standardisation was conducted with 34,762 pupils in the UK for all test versions (4,071 pupils for PTM7). Internal consistency is reported at 0.91 (Cronbach's Alpha), and the assessment has been found to have no significant difference in Standard Age Scores between male and female pupils. Given the robust development process of the test and its established psychometric properties, we used the instrument as developed by GL Assessment and without modification for our main analysis.

PTM7 assessments were administered at the endpoint by NFER test administrators, who were blinded to school allocation to intervention or control. Pupils completed the assessment within the classroom and under test conditions. The PTM7 is not time-limited, but GL Assessment has suggested that approximately 35 minutes would be needed for pupils to demonstrate their abilities. Test administrators used a secure courier to send the completed assessment papers to GL Assessment, who completed the marking and scoring, blinded to group allocation. GL Assessment then used the NFER secure portal to share with NFER a spreadsheet containing the pupil-level data, including overall raw scores and subtotal raw scores for the PTM7 subscales described below, alongside item-level scoring. Standard Age Scores, Stanine Scores, and National Percentile Ranks were also included in this data.

Secondary outcomes

All PTM assessments are based on categories of mathematical proficiency, which have been derived by GL Assessment from the Curriculum Aims in the Key Stages 1, 2, and 3 National Curriculum for England 2013. The assessment comprises of 34 questions aligned with these categories, as follows:

- fluency in facts and procedures;
- fluency in conceptual understanding;
- mathematical reasoning; and
- problem-solving.

The four categories were used as subscales for the secondary outcomes,[4] thus providing additional findings which helped understand how the programme's impact is associated with specific aspects of mathematical attainment. One of the subscales in particular (mathematical reasoning) was hypothesised to be directly associated with receipt of the programme and is identified by the Theory of Change as a short-term outcome ('Pupils have a better understanding of quantitative reasoning'). However, the developers of Mathematical Reasoning emphasise that the programme promotes reasoning about relations between quantities and about relations between numbers. This is a narrower focus than the 'mathematical reasoning' PTM7 subscale; not all questions on the subscale relate to programme content.

The PTM7 subscales—as defined by GL Assessment—were used in the secondary analysis, rather than developed using data-driven methods such as exploratory factor analysis. We analysed the trial test data for reliability as a sensitivity check (see 'Confirmatory Factor Analysis (CFA) on the PTM7 subscales' section).

Baseline measures

The baseline measure was GL Assessment's PTM6 Form A (for RQs 1, 2.1, 2.3, 2.4, 2.5, and 2.6), which is designed for administration at the end of Year 1 or the start of Year 2. This assessment is based on the same categories as PTM7, with a Pearson correlation between PTM6 and PTM7 of 0.67 (Bishenden, 2023). As a point of comparison, the pre-post correlation in EEF trials where a commercial test has been used at baseline and endpoint varies between 0.28 and 0.75, with a median

[4] For RQ2.2. the other secondary RQs use the same outcome as primary RQ1.

of 0.54 (Singh *et al.,* 2023).[5] As the PTM6 tests content covered in the previous academic year, it provides a more accurate assessment of pupil ability at baseline, compared to using the PTM7. For RQ2.2, the baseline measure was the PTM6 subscale corresponding to the PTM7 subscale used for the outcome.

The baseline assessment was administered by class teachers, who were asked to do so within the classroom and under exam conditions. The reason for this approach—rather than using NFER test administrators, as for the endpoint PTM7 test—was to reduce costs and because the PTM6 was completed before randomisation, so bias caused by unblinded test administration was not a consideration.

Sample size

All sample size calculations were performed using the 'PowerUpR' package (Bulus *et al.*, 2021) in the R statistical software, using the function 'mdes.cra2'.

Protocol stage

Given that the previous Mathematical Reasoning effectiveness trial (Stokes *et al.*, 2018) found an effect size of 0.08 among all pupils, we considered it appropriate to power this trial for a relatively low minimum detectable effect size (MDES) at the protocol stage. We therefore recommended that the Oxford University team recruit schools to the upper limit of their capacity to deliver the intervention (120 schools). We assumed one class per school (25.7 pupils per class, seven of whom are FSM-eligible[6]).

While GL Assessment reports that the correlation between PTM6 and PTM7 is 0.67, we assumed a slightly lower correlation (0.62). This was due to the use of these assessments in the context of a programme, where we would expect different correlations between the intervention and control groups, resulting in an overall lower correlation than what would be found outside of a programme context. The formula used by 'PowerUpR' does not use the pre- and post-correlation directly, instead requiring the proportion of variance explained at level 1 and level 2 (Bloom and colleagues provide an explanation of these parameters; Bloom, Richburg-Hayes, and Black, 2007). We assumed the variance explained at both level 1 and level 2 would be equal,[7] in which case both would be equal to the pre- and post-correlation (0.62) squared. We also assumed a school-level intracluster correlation coefficient (ICC) of 0.11, which aligned with analysis from the previous effectiveness trial (Stokes *et al.*, 2018).

Using the specified parameters, we estimated the MDES that could be detected with a power of 0.8 to be 0.108 for all pupils and 0.140 in the FSM-eligible subgroup, as shown in Table 9. However, this figure did not account for school and pupil attrition before the analysis stage. We anticipated 90% of schools would remain in the trial and within these schools 85% of pupils would be included in the primary analysis (i.e. 10% school-level attrition and 15% pupil-level attrition). The MDES we anticipated would in fact be achieved at the analysis stage was therefore 0.117 (0.154 in the FSM-eligible subgroup).

Randomisation stage

The number of schools randomised was slightly fewer than planned (238 vs 240). However, the number of pupils per class was slightly higher than expected, as was the proportion of FSM-eligible pupils.[8] The overall effect was that the MDES among all pupils (0.108) and FSM-eligible pupils (0.138) remained virtually unchanged between the protocol and randomisation stages. Adjusting for anticipated attrition again produced an MDES of 0.117 for all pupils and 0.153 for the FSM-eligible subgroup, as for the protocol stage. For simplicity, all the anticipated 10% school-level attrition was assumed to occur between randomisation and analysis.

[5] EEF trials will implement an educational intervention for some pupils, which may affect the correlation between pre- and post-test, somewhat limiting the comparability of these results with the correlation between the PTM6 and PTM7.
[6] Seven FSM-eligible pupils per class implies 26.9% of pupils indicated as eligible for FSM by the EVERFSM_6_P NPD variable. The 2022/2023 DfE figures indicated 24.6% of Year 2 pupils were FSM-eligible, a figure we expected to rise further by 2024/2025.
[7] The parameters in Appendix 5 of the Mathematical Reasoning efficacy trial (Worth *et al.*, 2015) support this assumption.
[8] Estimated using FSM status provided directly by schools, which was the only source of FSM data at that point in the trial. All other impact analyses, including MDES calculations at the analysis stage, used the EVERFSM_6_P variable from the NPD.

Analysis stage

At the analysis stage, three parameters differed from what we had anticipated at earlier stages. First, there was almost no school-level attrition: 236 (N=117 intervention, N=119 control) of the 238 schools randomised remained in the analysed sample. The proportion of variance explained by the PTM6 was also higher than anticipated (0.75 vs 0.62) at the pupil level and higher than anticipated (0.68 vs 0.62) at the school level. Both these factors would be expected to lower the MDES, relative to the 0.117 anticipated for the whole sample. However, the ICC observed in the analysed data was 0.21, much higher than the 0.11 anticipated, which increased the MDES. The net effect of these three factors was an increase in MDES to 0.131 for all pupils (0.161 in the FSM-eligible subgroup). The trial was therefore slightly underpowered for both the all pupils and FSM-eligible group, relative to the targeted MDES of 0.117 and 0.153.

Randomisation

Schools were randomised to the intervention and control groups in a 1:1 ratio. The randomisation was performed using the R statistical software by an NFER statistician. The statistician was not blinded to group allocation. Code was stored for transparency and replicability of the randomisation process and is available as an appendix in the **SAP** (Morton and Smith, 2025).

A simple randomisation was performed for this trial, without stratification. Simple randomisation is considered to be safer than restricted (i.e. stratified) randomisation, which may increase the chance of selection bias and errors during the randomisation process (Hewitt and Torgerson, 2006). While not stratifying, we included a sensitivity check that added socio-economic variables to the primary analysis model (see 'Additional analyses and robustness checks' section below).

Statistical analysis

An intention-to-treat (ITT) approach was followed for all models that included the intervention indicator (except the compliance analysis). This means pupils were analysed according to their intervention or control group assignment, regardless of their degree of participation in the intervention. Analysis was conducted on complete cases only; pupils with missing values for any variables used in an analysis model were excluded from modelling. The missing data analysis investigated the sensitivity of results to this choice. The analyst was not blinded to intervention assignment for any part of the analysis.

All analysis was conducted in accordance with the EEF statistical analysis guidance (Education Endowment Foundation, 2022) using the R software (R Core Team, 2026). Mixed effects models were analysed using the R package 'lme4' (Bates *et al.*, 2015).

Primary analysis (RQ1)

The purpose of the primary analysis is to answer the RQ:

> **RQ1:** What is the impact of the Mathematical Reasoning programme on Year 2 pupils' attainment in mathematics (measured using GL Assessment PTM7)?

A two-level (pupil and school) linear mixed effects model was used for this analysis:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

In this model:

$\mathrm{PTM7_{ij}}$ = total raw PTM7 score of pupil i in school j, measured at endpoint;

β_0 = intercept term;

$\mathrm{intervention_j}$ = indicator for whether school j was randomised to the intervention (1) or control (0);

$\mathrm{PTM6_{ij}}$ = total raw PTM6 score of pupil i in school j, measured at baseline;

β_2 = average impact of each additional point of the baseline PTM6 score on the PTM7 score;

b_j = school-level error term (random intercept); and

ϵ_{ij} = pupil-level residual error term.

The main estimate of interest was β_1, which represents the average impact of Mathematical Reasoning on PTM7 score, with a 95% confidence interval (CI) for this estimate calculated using the profile likelihood. The estimate and CI were converted to a standardised effect size, as described in 'Estimation of effect sizes' below.

FSM subgroup analysis (RQ2.1)

The effect of Mathematical Reasoning among FSM-eligible pupils[9] (RQ2.1) was determined by repeating the primary analysis model, restricted to pupils from the FSM subgroup. Additionally, we investigated the differential effect of the programme for FSM-eligible pupils relative to non-FSM pupils using the model:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6_{ij}} + \beta_3 \mathrm{FSM_{ij}} + \beta_4 \mathrm{FSM_{ij}} \times \mathrm{intervention_j} + \mathrm{b_j} + \epsilon_{ij}$$

where $\mathrm{FSM_{ij}}$ indicates whether a pupil is eligible for FSM (1) or not eligible (0). $\mathrm{B_3}$ is the difference in PTM7 score for FSM pupils compared to non-FSM pupils, among pupils in the control group. The interaction term coefficient β_4 measures the differential impact of the intervention for FSM-eligible pupils relative to non-FSM pupils, which is the key aspect of interest from this model. If β_4 has a low p-value and is positive, this suggests more benefit for FSM-eligible pupils than for non-FSM pupils when receiving the intervention.

Secondary analysis (RQ2.2)

Four subscales of the PTM7 were included as secondary outcomes (for RQ2.2 only—other secondary RQs used the primary outcome of PTM7 total score): i) fluency in facts and procedures; ii) fluency in conceptual understanding; iii) problem-solving; and iv) mathematical reasoning. These subscales were the dependent variables in four linear two-level (pupil, school) regressions. For example, in the case of the fluency in facts and procedures subscale, the regression model was:

$$\mathrm{PTM7\_FFP_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6\_FFP_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

where $\mathrm{PTM7\_FFP_{ij}}$ is the score of pupil *i* in school *j* for the fluency in facts and procedures subscale of the PTM7. Similarly, $\mathrm{PTM6\_FFP_{ij}}$ is the pupil's score for the fluency in facts and procedures subscale of the PTM6 at baseline. This regression was then repeated for each of the other three subscales, with $\mathrm{PTM7\_FFP_{ij}}$ and $\mathrm{PTM6\_FFP_{ij}}$ replaced by the appropriate PTM7 and PTM6 subscales.

Due to the smaller range of possible scores for the PTM7 subscales (the smallest range is 0–4 for problem-solving), we anticipated that they could be more vulnerable to floor and ceiling effects, as observed in the first evaluation of Mathematical Reasoning (Stokes *et al.*, 2018). Histograms were produced for all PTM6 and PTM7 subscale scores (see Appendix F of Further Appendices), which suggested floor effects (for fluency in facts and procedures and fluency in conceptual understanding) or ceiling effects (for problem-solving) were present for most scales. The implications for the reliability of the secondary analysis estimates are discussed in the 'impact evaluation results' section below.

As there are multiple outcomes in the secondary analysis, they have an inflated 'family-wise error rate' (chance of one or more false positives) when conducting significance tests. We did not implement a multiple testing correction (e.g. Bonferroni) to address this, as we expect the resulting analysis would be severely underpowered to detect an intervention

[9] 'EVERFSM_6_P' from the 2024/2025 spring census was used to measure FSM eligibility throughout the impact analysis. The variable indicates whether a pupil has been eligible for FSM at any point in the last six years. For this evaluation of Year 2 pupils' FSM records do not go back the full six years, since FSM status is first recorded in the Reception year.

effect. Because of this we report isolated low p-values among the secondary analysis results with caution and avoid a dichotomous 'significant' versus 'non-significant' interpretation of p-values.

Analysis in the presence of non-compliance

Teachers at intervention schools completed a Session Delivery Log of Mathematical Reasoning sessions. For each session, the unit completed in that session was recorded (or 'unit not complete' if no unit was completed), together with which pupils attended the session. By combining this information, we calculated the number of units attended by each pupil and across how many sessions. Where a unit spanned multiple sessions, a pupil's attendance at all those sessions was required for the unit to be counted. There were three compliance measures for this evaluation, all of which were based on a pupil's attendance of up to 12 units delivered by their teacher.

1. Number of units attended by a pupil (continuous measure).
2. Whether a pupil attended ten or more units (dichotomised variable derived from continuous measure).
3. Whether a pupil attended one or more units (dichotomised variable derived from continuous measure).

For this evaluation, the main compliance measure (1) is continuous. Compliance measure (2) investigates the impact of attending most units. Compliance measure (3) exists as a lower bound for measure (2), as described below.

For each of these measures an instrumental variable analysis was performed, using two-stage least squares methods (Angrist and Imbens, 1995) to estimate the effect of compliance with the intervention on PTM7 scores. For instrumental variable analysis to produce unbiased results the 'exclusion restriction' must hold: intervention assignment can only affect PTM7 scores via the chosen compliance measure. One implication of the exclusion restriction is that when dichotomising a continuous measure (here number of units), intervention pupils receive no benefit below the chosen compliance threshold. This is unlikely to hold for compliance measure (2): we expected some benefit for pupils attending fewer than ten units, which is likely to upwardly bias this result. For this reason, we included compliance measure (3) to act as a 'lower bound' for measure (2). Because measure (3) is relatively robust to the exclusion restriction, the unbiased estimate for measure (2) must be higher than the estimate for measure (3) (Gerber and Green, 2012).

None of the compliance measures (1–3) involve a teacher's completion of training sessions. We considered including this information, either as a composite measure with pupils' units attended or as a stand-alone measure. However, it would be difficult to obtain an accurate estimate of the impact of teacher training using instrumental variable methods, as a pupil may benefit from the intervention just by receiving Mathematical Reasoning units, whether or not their teacher completed the training (i.e. another form of exclusion restriction violation). Instead, we investigated the impact using regression modelling outside of the instrumental variable framework (see RQ 2.6).

The impact of each compliance measure described above was modelled for all pupils and for the FSM-eligible subgroup separately, resulting in a total of six models. For the first stage, the compliance variable was regressed on the intervention indicator and baseline PTM6 score. This first stage linear regression was therefore:

$$\mathrm{compliance}_{\mathrm{ij}} = \beta_0 + \beta_1 \mathrm{intervention}_{\mathrm{j}} + \beta_2 \mathrm{PTM6}_{\mathrm{ij}} + \epsilon_{\mathrm{ij}}$$

In each model, the continuous or dichotomised compliance variable, $\mathrm{compliance}_{\mathrm{ij}}$, was defined based on one of the three measures above and took the value zero for control pupils (the trial design ensured control pupils could not receive Mathematical Reasoning). For the second stage linear regression PTM7 scores were regressed on each pupil's predicted compliance value, $\widehat{\mathrm{compliance}}_{\mathrm{ij}}$, obtained from the first stage:

$$\mathrm{PTM7}_{\mathrm{ij}} = \beta_0 + \beta_1 \widehat{\mathrm{compliance}}_{\mathrm{ij}} + \beta_2 \mathrm{PTM6}_{\mathrm{ij}} + \varepsilon_{\mathrm{ij}}$$

The coefficient for predicted compliance β_1 in this second stage is the Complier Average Causal Effect (CACE) estimate for the effect of compliance on PTM7 scores. For the linear CACE measure, this can be interpreted as the average change in PTM7 score per additional unit attended.

All instrumental variable analyses was performed using the R package 'ivreg' (Fox *et al.*, 2021). These models do not include school-level random effects, so instead cluster-robust standard errors were calculated using the R package 'sandwich' (Zeileis, 2006; Zeileis, Köll, and Graham, 2020).

Missing data analysis

The number and proportion pupils with any missing primary analysis variables (PTM6 or PTM7 score) are reported in the 'Impact evaluation results' section below. PTM6 score was missing for individual pupils included in the trial, but not for entire schools (as completing PTM6 assessments was a condition of a school being randomised). To identify patterns of missingness, a mixed effects logistic regression with two levels (pupil and school) was run, where the outcome was the logit probability of the PTM7 outcome being missing. The PTM6 and intervention indicator variables from the primary analysis model were included as covariates, together with auxiliary variables that may be associated with missingness:

- FSM eligibility in the Spring Term of 2024/2025;
- SEND status in the Spring Term of 2024/2025;
- Proportion of half-days pupil was absent from school (authorised or unauthorised) in the 2024/2025 Autumn Term and Spring Term;[10]
- Number of pupils at the school per full-time teacher in 2024/2025; and
- Proportion of pupils eligible for FSM at the pupil's school in 2024/2025.

A second logistic regression was run in which the outcome was missingness of the PTM6 baseline. PTM7 score replaced PTM6 score as a predictor in this second model, but it was otherwise the same model as that described above.

Any of the auxiliary variables which demonstrated an association with missingness in the PTM7 scores (indicated by a p-value below 0.05) were included as covariates to rerun the primary analysis model as a sensitivity check. This did not replace the primary analysis result; it was an additional result provided for context. Additional variables associated with missingness in either PTM6 or PTM7 score were included as predictors in the multiple imputation procedure described later in this section.

In general, it is not possible to determine whether data is missing at random (MAR) or missing not at random (MNAR) using only the observed data: it is impossible to be sure that the set of available variables are sufficient for the data to be MNAR. We therefore also performed further sensitivity analyses that incorporated MNAR assumptions, using multiple imputation to explore three simple missing data patterns for pupils with a missing PTM7 outcome:

a) Outcomes are MAR, conditional on available covariates.
b) Outcomes are MNAR and on average missing outcome scores are lower than non-missing scores by 1 standard deviation (SD).[11] Missing scores are 'equally worse' between the intervention and control groups.

[10] This is slightly different to the variable specified in the **SAP** (Morton and Smith, 2025), which was the number (not proportion) of absences. The proportion was calculated as $\frac{Number\ of\ half-day\ absences}{Number\ of\ half-days\ possible}$ using the NPD variables 'OverallAbsence_2Term' and 'SessionsPossible_2Term'. As 'SessionsPossible_2Term' varied between pupils it was decided that the proportion was a better summary than the number of absences for how much of the available learning in the Autumn Term and Spring Term was lost.
[11] That is,1 SD of the PTM7 outcome among all pupils with a recorded outcome. This was a somewhat arbitrary choice: it was intended to be large but within the range of the observed data.

c) Outcomes are MNAR for intervention pupils only: intervention pupils with missing outcomes have their score reduced to the average among control pupils.[12] This models a situation where, among pupils with missing outcomes, intervention pupils do not receive any benefit from the intervention.

The plausibility of these missing data patterns depends on the reason why a pupil did not have a PTM7 score available. We were able to distinguish between the reasons based on the endpoint test absence codes we collected from schools (there was no missing data in the absence codes themselves). Each reason (absence code) is given in the first column of Table 5, along with the missing data pattern we considered could apply in a reasonable 'worst case' scenario (worst in terms of biasing the primary analysis result). A small number[13] of pupils had missing PTM7 scores due to pupils being withdrawn from the trial by their parents/carers, including withdrawal from data processing (i.e. their data could not be used in the trial). Due to the data processing withdrawal, most information for these pupils was deleted from NFER records. This caused errors in the imputation procedure, as most variables used to impute the missing values (e.g. PTM6 score, SEND status) were themselves missing, so the pupils could not be included in modelling. This is contrary to the **SAP** (Morton and Smith, 2025), where we stated these pupils would be included, with their outcomes assumed to follow pattern (a) (outcomes are MAR), above.

We used the full dataset to impute outcomes for pupils corresponding to pattern (b) and (c) according to the assumptions in Table 5; each imputation of PTM7 scores was edited to reflect the relevant MNAR assumption (e.g. by having one standard deviation subtracted for pattern [b]). Missing data was imputed using a two-level normal model with the '2l.lmer' function, using chained equations in the R package 'mice' (van Buuren and Groothuis-Oudshoorn, 2011). The imputation model included all variables from the primary analysis, together with auxiliary variables associated with missingness of PTM6 or PTM7 scores. Ten datasets were generated, each using 20 iterations. The mean and standard deviation of imputed variables were plotted against imputation number, to check for convergence. If means or standard deviations were systematically increasing/decreasing with iteration number, or the lines from different imputations were not 'mixing' (crossing each other), this could have indicated non-convergence. In practice, there was no evidence of non-convergence, so the analysis proceeded without any further checks or changes needed. The primary analysis model was then rerun on each of the ten datasets generated and results from each model were pooled into a single set of estimates and standard errors. This pooled estimate was compared to the primary analysis result.

Table 5: Reason why a pupils' PTM7 outcome is not available and the corresponding missing data pattern in a reasonable worst case scenario[a]

Reason for missing PTM7 outcome	Missing data pattern in reasonable worst case scenario	Explanation
Pupil absent on the day of test	Pattern (a) (MAR)	No reason to believe that absence on the day of the test is correlated with worse outcomes, except possibly via persistent absence (we potentially include absence rates in the imputation model, as described above)
Pupil did not want to take the test	Pattern (b) (MNAR)	Reluctance to take the test could indicate a lack of confidence and/or knowledge of test material
Pupil present but excluded from the test	Pattern (b) (MNAR)	Poor behaviour may be correlated with lower attainment. It may also be an expression of reluctance to take the test, as in the cell above This code could also be used for pupils that schools do not think are suitable for testing (e.g. due to SEND)
School withdrawal	Pattern (c) (MNAR)	School withdrawal may indicate lack of resources (e.g. staff time) to implement the intervention properly or lack of engagement with the intervention

[12] For this scenario to be applicable a positive intervention effect of course needed to be observed in the primary analysis, which was the case in practice.
[13] Less than ten in both the intervention and controls groups. Exact numbers are not shared to prevent disclosure of individual pupil data.

Pupil left school	Pattern (c) (MNAR)	Intervention pupils who leave the school during the trial period will miss some or all intervention sessions (effectively non-compliance)

[a] *Patterns (a–c) are not an exhaustive list of missing data patterns. In particular, we do not consider a pattern where intervention pupils with missing outcomes perform worse than control pupils with missing outcomes (but the primary analysis estimate is positive). This pattern might be considered in, for example, a medical trial where side effects of the intervention are linked to both attrition and poor outcomes, but we do not consider it likely for an educational intervention.*

Additional analyses and robustness checks

Impact of the number of sessions attended (RQ2.3)

The compliance analysis investigated the impact of the number of Mathematical Reasoning units attended; we also separately explored the impact of the number of sessions attended (RQ2.3) outside of an instrumental variable framework, as described here. There are 12 Mathematical Reasoning units in total, each of which may be delivered to pupils in a single session, or more than one session may be required to complete a unit. To investigate the impact of the number of sessions attended by a pupil, a two-level (pupil, school) linear regression was run, restricted to intervention pupils only.

$$\mathrm{PTM7}_{ij} = \beta_0 + \beta_1 \mathrm{N\_sessions}_{ij} + \beta_2 \mathrm{PTM6}_{ij} + \beta_3 \mathrm{prop\_absent}_{ij} + b_j + \epsilon_{ij}$$

In this equation $\mathrm{N\_}sessions_{ij}$ is the number of Mathematical Reasoning sessions attended by of pupil i in school j and β_1 is the average change in PTM7 score per session attended. $\mathrm{prop\_absent}_{ij}$ is the proportion of half-days absent from school (not just Mathematical Reasoning sessions) in the Autumn Term and Spring Term of 2024/2025[14] and β_3 is the average change in PTM7 score per additional 1% increase in the proportion of half-days absent. The rationale for including absences as a covariate in the model was that persistent absence was likely to impact both number of sessions attended and attainment (i.e. it was a potential confounder). The model above was run for both all pupils and for the subgroup of FSM-eligible pupils.

Impact of computer game participation (RQ2.4)

As part of the Mathematical Reasoning intervention pupils played educational computer games: there were 60 different games that were categorised into six subject topics. The Mathematical Reasoning programme did not instruct teachers on the details of how the games should be played, including the number of games played and how long pupils spent on each game. In principle pupils could repeat the same game any number of times, even if they scored 100%.

Teachers at intervention schools were given a spreadsheet which specified the username each trial pupil should be assigned when setting up their computer game account. This username was the only identifier provided within the computer game data shared with NFER by the Oxford University team. This meant that if teachers did not assign usernames as instructed, it was not possible to correctly link a pupil's computer game data to their other records held by NFER (e.g. PTM7 score). We categorised the quality of the match between a pupil's computer game data and their other data as follows:

- **'Green' match (N=730 pupils).** Expected usernames were used and school confirmed that usernames were assigned to pupils in the correct order. We can be reasonably sure that the computer game data was correctly matched for these pupils.

- **'Amber' match (N=1,269 pupils).** Expected usernames were used but the school did not confirm that they assigned usernames to pupils in the order instructed. There remains a possibility that the correct usernames were assigned in the wrong order, which would mean the computer game records would be matched to the wrong pupils in some cases.

[14] This is the same variable used in the missing data analysis and is described further in that section.

- **'Red' match (N=1,307 pupils[15]).** Expected usernames were not used, so it was not possible to match computer game data with other data at all for these pupils. They therefore could not be included in the computer game impact analysis.

Pupils with both 'green' and 'amber' matches were included in the modelling described in this section.

Two models were analysed, investigating the moderating effect of the number of Mathematical Reasoning computer games played (Model 1) and the number of subject topics[16] (Model 2). All models were restricted to intervention pupils only.

$$\mathrm{PTM7}_{ij} = \beta_0 + \beta_1 \mathrm{N\_games}_{ij} + \beta_2 \mathrm{PTM6}_{ij} + b_j + \epsilon_{ij} \qquad \text{(Model 1)}$$

$$\mathrm{PTM7}_{ij} = \beta_0 + \beta_1 \mathrm{N\_subject\_topics}_{ij} + \beta_2 \mathrm{PTM6}_{ij} + b_j + \epsilon_{ij} \qquad \text{(Model 2)}$$

Here $\mathrm{N\_games}_{ij}$ counts the number of games played by pupil i in school j. If a pupil played the same game multiple times, all these repeated games were counted towards the total for calculating $\mathrm{N\_games}_{ij}$. $\mathrm{N\_subject\_topics}_{ij}$ counts the number of subject topics from which the pupil played at least one game. As the benefit of additional games may depend on the number of subject topics and vice versa, a further model (Model 3) was analysed:

$$\mathrm{PTM7}_{ij} = \beta_0 + \beta_1 \mathrm{N\_games}_{ij} + \beta_2 \mathrm{N\_subject\_topics}_{ij} + \beta_3 \mathrm{N\_games}_{ij} \times \mathrm{N\_subject\_topics}_{ij} + \beta_4 \mathrm{PTM6}_{ij} + b_j + \epsilon_{ij} \qquad \text{(Model 3)}$$

In this model β_3 is the coefficient for the interaction between $\mathrm{N\_games}_{ij}$ and $\mathrm{N\_subject\_topics}_{ij}$, representing how the impact of each additional game played changed with the number of subject topics (and vice versa). The three models described above were repeated in the FSM subgroup (labelled Models 4–6 in the 'Impact evaluation results' section below).

In the **SAP** (Morton and Smith, 2025), we stated that we would perform further exploratory analyses on the computer game data but did not specify the details (only the models above were pre-specified in detail). This was because there were several properties of the computer game data we were unsure of, including how reliably it could be linked to other pupil data such as PTM7 score. Described below are the remaining computer game analyses, none of which were pre-specified in the **SAP** (Morton and Smith, 2025).

When performing the analysis, we noticed that a small number of pupils played far more games than their peers: 11 pupils played between 150 and 767 games (the median was 21 games). There was evidence of non-effortful game playing (i.e. just 'clicking through' without thinking) for most of these pupils, as indicated by strings of several games with scores close to 0 completed within the same minute.[17] Aside from this fact, it also seemed unlikely that a linear relationship between number of games played and PTM7 score would hold in this high range. By default, these 11 pupils were therefore excluded from modelling. However, a sensitivity analysis was performed in which the 11 pupils were added back into the analysis and Models 1–3 were rerun. Additionally, a sensitivity analysis was performed in which Models 1–3 were restricted to only pupils with a 'green' quality match for their computer game data (as described above).

We also investigated the relationship between PTM7 score and how many games a pupil scored 100% (Models 7–8), as well as between PTM7 score and how many games they scored at least 50% (Models 9–10). In Model 8, the variable $\mathrm{N\_games}_{ij}$ (described above) was added to Model 7, to explore whether there was any benefit to games were a pupil did not score 100% (and similarly for Model 9–10).

[15] This is 1,307 records in the computer game data, but there was not an exact one-on-one correspondence between these records and the remaining intervention pupils in the trial (e.g. due to account sharing or pupils joining the class after randomisation).
[16] 'Subject topics' is used in the place of the term 'different games' used in the **SAP** (Morton and Smith, 2025), as the former was decided to be less prone to misinterpretation. The methods remain the same, only the terminology has changed.
[17] We also observed some evidence of non-effortful game playing among the remaining pupils, albeit to a lesser extent. This was based on inspection on a small subset of the data: we did not attempt to systematically quantify non-effortful game playing among all pupils.

$$PTM7_{ij} = \beta_0 + \beta_1 N\_games\_100\_percent_{ij} + \beta_2 PTM6_{ij} + b_j + \epsilon_{ij} \quad \text{(Model 7)}$$

$$PTM7_{ij} = \beta_0 + \beta_1 N\_games\_100\_percent_{ij} + \beta_2 N\_games_{ij} + \beta_3 PTM6_{ij} + b_j + \epsilon_{ij} \quad \text{(Model 8)}$$

$$PTM7_{ij} = \beta_0 + \beta_1 N\_games\_at\_least\_50\_percent_{ij} + \beta_2 PTM6_{ij} + b_j + \epsilon_{ij} \quad \text{(Model 9)}$$

$$PTM7_{ij} = \beta_0 + \beta_1 N\_games\_at\_least\_50\_percent_{ij} + \beta_2 N\_games_{ij} + \beta_3 PTM6_{ij} + b_j + \epsilon_{ij} \quad \text{(Model 10)}$$

In these models $N\_games\_100\_percent_{ij}$ was the number of games in which pupil i in school j scored 100% and $N\_games\_at\_least\_50\_percent_{ij}$ was the number of games in which they scored 50% or more. These variables did not make a distinction in terms of whether a pupil was scoring highly across different games or repeatedly in the same game. For example, playing the same game three times and scoring 100% each time was treated as equivalent to scoring 100% once in three different games.

We considered modelling whether computer games played outside of school hours have a different relationship with PTM7 score, compared with those played within school hours. However, when we analysed the computer game data we found that only 100 pupils played any computer games outside of school hours, almost certainly too few to accurately estimate a relationship with PTM7 score. We therefore decided not to run these models.

None of the computer game variables were randomised, nor was it possible to use quasi-experimental methods such as instrumental variables analysis. The analysis was therefore vulnerable to confounding from unobserved variables. For example, pupils that played more computer games may have been more engaged to learn maths generally and so scored higher in the PTM7 because of this (rather than benefiting from the games per se).

Prior attainment moderator analysis (RQ2.5)

Further analysis was performed investigating how the impact of the intervention varies with prior attainment (RQ2.5a). Prior attainment (PTM6) is already included as a continuous variable in the primary analysis model; this investigation required the addition of an interaction term to that model:

$$PTM7_{ij} = \beta_0 + \beta_1 intervention_j + \beta_2 PTM6_{ij} + \beta_3 PTM6_{ij} \times intervention_j + b_j + \epsilon_{ij}$$

Here β_3 estimated the additional (or lesser) impact of the intervention per additional point of prior attainment on the PTM6. We did not attempt to dichotomise PTM6 scores to form a 'lower prior attainment' subgroup, instead restricting our investigation to the model above that treated PTM6 score as continuous.

This same analysis was also repeated for only FSM-eligible pupils, investigating the differential impact of the intervention as prior attainment increased for this subgroup (RQ2.5b). This second analysis, restricted to FSM pupils, was exploratory as it was likely to be underpowered.

Impact of teacher training attendance (RQ2.6)

Teacher completion of the Mathematical Reasoning training was not investigated in the compliance analysis, but was included as a separate RQ (RQ2.6) as described in this section. Training completion for the purpose of this trial was defined as the first five training modules and two out of three training webinars completed. We understand that completion of training modules 1–4 is a prerequisite for completing module 5, so it was not possible to investigate the impact of completing fewer than five modules. A two-level (pupil, school) linear regression was run, restricted to intervention pupils only.

$$PTM7_{ij} = \beta_0 + \beta_1 N_{units_{ij}} + \beta_2 PTM6_{ij} + \beta_3 Training_{completed_{ij}} + b_j + \epsilon_{ij}$$

In this equation $\mathrm{Training}_{\mathrm{attendance}_{ij}}$ is a binary indicator for whether a pupil's teacher fully completed the Mathematical Reasoning training (as defined above). B_3 was therefore the average impact of receiving Mathematical Reasoning from a teacher with complete training, compared to one with incomplete training, on pupil PTM7 scores. This was adjusted for number of units a pupil attended, $\mathrm{N}_{\mathrm{units}_{ij}}$, so that β_3 better reflected the benefit of the training itself, rather than any correlation with the number of units delivered. The model above was run for both all pupils and for the subgroup of FSM-eligible pupils.

Confirmatory Factor Analysis (CFA) on the PTM7 subscales

The PTM7 subscales, which are the outcomes for RQ2.2, were identified by GL Assessment for general usage by mapping assessment questions to national curriculum areas, rather than being developed using a data-driven approach such as exploratory factor analysis and so their psychometric properties are unknown. To investigate the reliability and validity of the PTM7 subscales, CFA was performed on the PTM7 items to confirm the four-subscale structure proposed by GL Assessment.

Missing data on individual items were imputed with zeros for the CFA and other reliability analysis—this was deemed the appropriate method given that this is how missing data is handled when calculating total scores for the PTM7. Also, given that there was no time limit for the assessment, it can be assumed that missed items could not be completed by the pupil, as opposed to the items not having been reached. This approach to missing data meant that individual item scores were only missing when the total PTM7 score was also missing. Where there were missing data for baseline PTM6 or endpoint PTM7 scores, those pupils were excluded, so that the analysed sample was the same as the primary analysis.

As item data were categorical (27 binary items, five items with three categories, two items with four categories), a diagonally weighted least squares means and variance (WLSMV) estimator was used (Li, 2016). The CFA assumption of multivariate normality was assessed using Mardia's (1970) tests. The results of the Mardia's tests provided clear evidence against multivariate normality, with each subscale and all items together deviating from normality on the skewness and/or kurtosis components of the Mardia's test. Multivariate non-normality was adjusted for in the CFA analyses by using the WLSMV estimator. The current sample size was also deemed to be sufficient for CFA (e.g. Myers, Ahn, and Jin, 2011).

To test statistically whether a four-factor model of the PTM7 was a better fit to the data than the original one-factor model, two CFAs were conducted, testing a four-factor model and a one-factor model. The fit statistics of both models were evaluated using the commonly referenced criteria by Hu and Bentler (1999); Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) above 0.95, Root Mean Square Error of Approximation (RMSEA) below 0.06, and Standardised Root Mean Square Residual (SRMR) below 0.08. Standardised factor loadings were also evaluated alongside fit indices to understand individual item fit; factor loadings greater than 0.3 were considered sufficient (e.g. Hair *et al.*, 1998). Owing to the ordinal estimation method used in this analysis, fit indices were interpreted jointly rather than by single cut-offs(Marsh, Hau, and Wen, 2004). Due to the sensitivity of the chi-squared statistic to large samples (Barrett, 2007) fit indices and factor loadings were used to determine model fit. The two CFA models were also compared statistically using a likelihood ratio test.

In addition to the CFAs, correlations between items were examined to understand the directionality and strength of relationships. Reliability estimates were also calculated for each subscale and the PTM7 overall score. Alpha-if-item-removed estimates were calculated for subscales with lower reliability estimates to determine whether any single item could be lowering the internal consistency of the scale.

It was specified in the **SAP** (Morton and Smith, 2025) that if certain items performed poorly—as identified by low factor loadings and the removal of the item improving overall model fit—the poorly performing items would be removed from their secondary outcome subscale. Any secondary analysis models that were affected by this change would then be rerun, with the relevant items removed from their PTM7 subscale, as a sensitivity check. However, in practice, dropping individual items did not improve any of the subscales, so this sensitivity check did not occur.

Addition of further covariates to the primary analysis

As described previously, the randomisation for this trial was not stratified and the primary analysis model did not include any covariates other than baseline PTM6 score. To assess the sensitivity of results to any chance imbalances in pupil-level characteristics,[18] we will rerun the primary analysis with additional covariates:

- FSM eligibility in the Spring Term of 2024/2025;
- SEND status in the Spring Term of 2024/2025;
- whether the pupil speaks EAL; and
- pupil gender.

The model was therefore:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6_{ij}} + \beta_3 \mathrm{FSM_{ij}} + \beta_4 \mathrm{SEN_{ij}} + \beta_5 \mathrm{EAL_{ij}} + \beta_6 \mathrm{male_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

where $\mathrm{FSM_{ij}}$, $\mathrm{SEN_{ij}}$, $\mathrm{EAL_{ij}}$ and $\mathrm{male_{ij}}$ were indicators for whether a pupil is FSM-eligible, has SEND, speaks EAL, and is male, respectively (yes [1] or no [0]). All four variables were added to the model, regardless of the degree of imbalance between the control and intervention groups observed at baseline.

Estimation of ICC

The ICC was calculated as the proportion of outcome variance attributable to level 2 (between-school) variation:

$$\mathrm{ICC} = \frac{\Sigma_B^2}{\Sigma_B^2 + \Sigma_W^2}$$

Here Σ_B^2 and Σ_W^2 are the between-school and within-school outcome variation, which were extracted directly from a mixed effects regression fitted by the 'lme4' package. ICCs were calculated for the PTM7 total score from the primary analysis and the four PTM7 subscales from the secondary (RQ2.2) analysis models. This was repeated for both all pupils and the FSM subgroup. The ICC was calculated twice in each case, once for the analysed model (with covariates) and once for an empty model (with no covariates). There were therefore $5 \times 2 \times 2 = 20$ ICCs calculated in total.

Estimation of effect sizes

Impact estimates from the models described above are presented as an effect size, as described by Hedges (2007):

$$\mathrm{ES} = \frac{\hat{\mathrm{B}}}{\sqrt{\Sigma_B^2 + \Sigma_W^2}}$$

$\hat{\mathrm{B}}$ is the coefficient for the binary predictor of interest, typically the intervention indicator, which was extracted from a conditional model (including any covariates). Σ_B^2 and Σ_W^2 are the between-school and within-school variance from the corresponding empty model (the same outcome but no covariates). To obtain a 95% CI for the effect size, a CI for B was first calculated. The end points of this CI were then divided by the denominator in the above effect size formula. The coefficients for continuous predictors were not converted into an effect size, they were simply presented on their original scale.

[18] School-level variables were also considered but we decided against their inclusion, as typically these have less potential to explain outcome variance, as reflected by the small ICCs seen in many educational studies.

Implementation and process evaluation (IPE)

The IPE was designed to complement the impact evaluation and to enable a stronger understanding of the mechanisms within the Theory of Change. The mode of training in this trial is different to that of the previous trial, with the aim of increasing fidelity by creating a more direct link with the developers. As a result, the IPE particularly focused on understanding variation and adaptation in implementation. Particular attention was given to the use of the online games and adherence to the programme structure, as the implementation of both of these aspects varied between schools in the previous effectiveness trial. The role of differentiated teaching within the programme was also of interest, including understanding how well teachers were able to adopt the flexible and adaptive grouping approaches intended by the programme. Moderators and contextual factors were interrogated, including access to IT facilities and the level and nature of the support from the TA and Senior Leadership Team. These findings help to contextualise the impact evaluation findings by providing insight into what does and does not work in relation to the programme and external factors that may influence this. Perceived impacts and possible mediating pathways were likewise explored to develop the Theory of Change further. Finally, the intention of the programme to both replace and complement standard maths lessons meant careful consideration was given to both programme differentiation and the potential risk of learning displacement.

Research methods

The evaluation team drew on a range of data collection methods to answer the IPE RQs (set out in the 'Evaluation objectives' subsection in the 'Introduction' section above). An overview of the IPE methods is presented in Table 6.

Table 6: IPE methods overview

Research methods	Data collection methods	Participants/ data sources	Data analysis methods	Research questions addressed	Implementation / logic model relevance
Qualitative	Semi-structured observation	TL training session	Thematic analysis	1	Fidelity
		Webinars		1, 2, 3	Fidelity, Context and Moderators
		Case study programme sessions		1, 2, 3, 5	Fidelity, Context and Moderators, Programme differentiation
	Semi-structured review	Online training course		1	Fidelity
	Semi-structured interview	TLs		1, 2, 3, 5	Fidelity, Context and Moderators, Programme differentiation
		Case study teachers, TAs and Maths Leads		1, 2, 3, 4, 5	Fidelity, Context and Moderators, Perceived impact, Programme differentiation
	Focus group	Case study pupils		1, 2, 3, 5	Fidelity, Context and Moderators, Programme differentiation
Quantitative	Online surveys	All control teachers	Descriptive statistics and pre- and post-comparison between intervention and control	1, 4, 5	Fidelity, Perceived impact, Programme differentiation
		All intervention teachers and TAs		1, 2, 3, 4, 5	Fidelity, Context and Moderators, Perceived impact, Programme differentiation
	Training course completion data	All intervention schools	Descriptive statistics	1, 2	Fidelity, Context and Moderators

	Webinar attendance data				
	Session delivery and attendance logs	All intervention pupils		1, 2, 3, 5	Fidelity, Context and Moderators, Programme differentiation
	Computer games platform analytics				
	Case study cost pro formas	Case study schools		2, 4	Context and Moderators, Perceived impact

The intended and actual sample for each of these research methods is presented in Table 7. More details about the data collection process, including reasons for the actual sample not aligning with the intended sample, are provided in the 'Data collection' section below.

Table 7: Sample size for each data collection activity

Data collection activity	Intended sample	Actual sample
Webinar observations	Two for each round of webinars, six in total (out of 27)	As intended
TL interviews	Six in total (out of nine), aligned with the observed webinars, range of different levels of experience with teacher training and the programme itself	As intended
Case studies	Eight schools of different sizes, with different TLs, and a relatively high proportion of FSM pupils in the trial • Interviews with eight teachers, TAs, and maths leads • Eight focus groups with four pupils in each • Eight session observations • Eight school cost pro formas Staff interviews took place at two time points	11 schools (see the 'Case studies' subsection below) of different sizes, mostly different TLs and mostly with a relatively high proportion of FSM pupils • Interviews with 11 teachers, nine maths leads, and six TAs • Seven focus groups with average of four pupils in each • Seven session observations • Eight school cost pro formas An interview was completed at two time points for at least one staff member at seven schools There was some overlap in interviewees (e.g. individual was both the teacher and the maths lead), and some interviewees took part in an interview together
Baseline BAU survey	262 teachers across intervention and control	244 teachers (93%) with 122 each of intervention and control
Endpoint BAU survey	131 control teachers	108 control teachers (82%)
Teacher and TA survey	131 intervention teachers and TAs	102 intervention teachers (78%) 63 intervention TAs (49%)
Training completion data	118 intervention schools	118 intervention schools (100%)
Webinar attendance data	118 intervention schools	118 intervention schools (100%)
Session Delivery Log and attendance log	129 intervention classes	104 intervention classes (81%)
Computer games platform analytics	3,282 pupils	3,167 pupils[a]

Note: The total sample included 12 classes (from 11 schools) that formally withdrew from programme delivery but did not withdraw from the trial, as well as two classes (from two schools) that did not complete endpoint testing.
[a] This includes some non-trial pupils at intervention schools, as this number exceeds the total number of possible trial pupils in the computer game data (3,018). The figure includes 11 pupils with over 150 games played, who were included in IPE descriptives, despite being excluded from the computer games impact analysis. No computer games data was received for 12 intervention schools.

Data collection

Semi-structured observation of TL training

In July 2024, an NFER researcher observed the second in-person TL training day, which focused on preparing the TLs to support programme implementation. The researcher used a semi-structured observation tool developed from the programme material to record points of interest in relation to the training structure and content.

Review of the teacher and TA online training course resources

In November 2024, an NFER researcher completed a structured review of the online training course material. This review was completed in parallel to training completion by the schools and informed the development of the instruments for subsequent data collection activities—particularly for the webinar and session observations. The researcher used a semi-structured review tool informed by other programme materials and the Theory of Change to record points of interest in relation to the training structure and content.

Semi-structured observations of webinars

An NFER researcher observed one of each of the three of the webinars (delivered by two different TLs), covering a range of TL experience levels in relation to teacher training and the programme itself. The webinars for observation were purposively sampled accordingly. The researcher used a semi-structured observation tool informed by other programme materials and the Theory of Change to record the level of participant engagement, fidelity, adaptations, variation between TLs and the types of questions and feedback from the participants, as well as a high-level summary of the content covered.

TL interviews

Following each webinar observation, the NFER researcher carried out an interview with the TL who had presented it. Each interview lasted between 30 and 45 minutes and was conducted remotely. The semi-structured topic guide was designed to address aspects of specific IPE dimensions of interest and covered both broader questions and specifics relating to the webinar just observed. TLs were asked about the training they received, how they understood their role and the kinds of support they provided to schools, how they felt the webinar went, how they felt schools were responding to the programme and any feedback or concerns they had or of which they were aware.

BAU surveys

In September 2024, prior to randomisation and group allocation, teachers from all trial schools were asked to complete an online survey about their "business as usual". The survey asked teachers about their existing maths schemes, interventions, and continuous professional development, as well as the strategies they employed for teaching maths with Year 2. This was to inform our understanding of relevant usual practice in schools prior to the trial.

In July 2025, teachers from the control schools were asked to complete a very similar survey to provide information about their usual practice over the trial period, for comparison with the intervention group.

Both surveys included questions on teachers' understanding of and confidence teaching various maths topics, including components of mathematical reasoning, as well as the year group in which they thought those topics should be first taught. These questions were included to gauge impact of the programme on these areas by comparing intervention and control at baseline and endpoint.

Response rates to the BAU surveys were high (92% at baseline and 82% at endpoint). Completion of the endpoint survey was not a requirement for control schools to receive their incentive payment.

Case studies

A total of 45 teachers were contacted with the request to act as a case study class. Small numbers of teachers were contacted at a time, with new teachers contacted in cases where the teacher declined or where no response was received after three emails. Teachers were selected pseudo-randomly to ensure a sufficient spread across TLs, as well as different school sizes and a disproportionately high number of FSM pupils in the case study sample overall. Of the eight teachers who initially agreed, one withdrew from case study data collection late in the trial. This led us to contact additional teachers with the offer of remote interviews rather than an in-person visit. This resulted in interviews with three more schools than had been initially anticipated, but one fewer session observation and pupil focus group than intended. As a result, the final number of case study schools was 11.

The sample of schools covered a range of different sizes, as it was hypothesised that school size could be a moderator for delivery, with smaller schools more likely to struggle with the staffing requirements. There was an average of eight pupils eligible for FSM per case study class, ranging from 1 to 14.

The intention was to carry out interviews with the teacher, TA, and maths lead during the visit and then again (remotely) at the end of the trial. This was successful for teachers with the seven schools that hosted a case study visit. However, only a small number of TAs and maths leads participated in a second interview. We judged this adaptation to be acceptable because the data collected was showing evidence of saturation, particularly in terms of the information that each school individually was able to offer.

The case study instruments were developed on the basis of the training review, webinar observations, other programme information, and the Theory of Change with the aim of informing all of the IPE dimensions of focus for the trial.

The semi-structured observation tool supported the researcher to collect data on context, the nature of TA support, fidelity to the intended content and structure, adaptations, levels of pupil participation, use of materials (including manipulatives) and the ways in which the teachers scaffolded for and supported the children to learn. Part of the tool was adapted from Oxford University's fidelity tool for the programme (for more information please see the **study protocol**; Flemons *et al.*, 2024), which was also shared with schools by the TLs to support self-evaluation.

Semi-structured topic guides were used to ask staff about existing maths practice in the school, their views on the training and delivery, adaptations, any challenges or facilitators they had encountered, perceived impact for themselves and their pupils, and any intentions to continue with the programme. While the interviews were intended to be in person, some were carried out remotely via Microsoft Teams for the convenience of the school staff members involved. The interviews lasted between 15 and 60 minutes.

The pupil focus groups were in person and lasted around 15 minutes. The pupils were asked about how they felt doing each of the components of a Mathematical Reasoning session and why, as well as how different these sessions were from their usual maths lessons, and which they preferred. Pupils were given coloured smiley face charts which they could use to indicate their response to a question, and which were collected anonymously at the end. A member of school staff was present and relevant safeguarding procedures followed at all times.

The case study data collection was carried out between February and July 2025 by three trained and Disclosure and Barring Service (DBS)-checked IPE researchers. Schools received £100 for participating in a case study visit. Schools that only participated in remote interviews received £20 per interviewee.

Endpoint teacher and TA survey

In July 2025, all teachers and TAs from intervention schools were asked to complete a survey about their experiences of the programme. The survey asked respondents about their perceptions of the programme training, support, and materials, TA availability, resources, fidelity and adaptations, challenges and facilitators, pupil participation, perceived impact for themselves and their pupils, and any intentions to continue with the programme in future. Respondents were also asked about time spent on various components of the programme and when this occurred (e.g. during usual working hours with another teacher covering their classroom responsibilities). Finally, the survey also covered the same questions as used in the BAU surveys to allow for comparison between intervention and control. These questions included teachers' understanding of and confidence teaching various maths topics (including components of mathematical reasoning), and the year group in which they thought those topics should be first taught.

Teacher response rates to the survey were within the expected range for an endpoint survey (76%), however TA response rates were lower (49%). Given the high rates of movement of TAs within and between schools we did not have accurate TA contact details on file for all schools, so teachers were asked to pass the survey on to the TA for completion. No TA support for delivery was reported by 11 teachers.

Training course completion data

The Oxford University team provided NFER with a summary of which modules had been completed by each school account, as well as which webinars each school had attended. A module was counted as ‘complete’ if the account had responded to at least one of the interactive activities in the module.

Session delivery and attendance data

All intervention teachers were asked to complete a Session Delivery Log, where they recorded when each session was delivered, which unit it covered, whether the TA was involved, and which pupils were present. The Session Delivery Log was shared as an Excel spreadsheet to be completed online, with one for Weeks 1–6 of delivery and another for Weeks 7+ to aid completion. Return rates for the Session Delivery Logs were high (79% for both). Attendance data was complete for 2,524 of the 2,690 pupils named in the Session Delivery Logs (94%). Those pupils with missing attendance data for some or all of the sessions have not been included in the analysis (neither in the ‘IPE results’ nor ‘Impact evaluation results’ sections).

Computer games data

Back-end data from the computer games website was extracted by the developer and shared with NFER in pseudonymised form. The data covers the number of games played, when and which games for each pupil on the system.

Analysis

Qualitative data: Observations, interviews, and reviews

A detailed qualitative analysis plan was developed by the project lead in advance. Interview and observation notes were written up as intelligent verbatim transcripts[19] and uploaded to the qualitative data analysis software MAXQDA. The data was then analysed using thematic analysis to identify and interpret patterns or themes. High-level deductive coding (approaching data with a pre-established framework for interpreting it) was used to sort the data into relevant themes. Detailed inductive coding (identifying patterns of meaning present in the data) was then used to draw out the key findings under each of these themes. Two of the IPE researchers who conducted the data collection also completed the analysis. Each researcher started by independently coding one interview and then the two met to discuss their approach and develop greater inter-coder reliability. The researchers then met at regular interviews to sense-check their interpretations of the data. All deductive coding was carried out based on a coding frame that was developed in advance based on the programme Theory of Change and IPE RQs to ensure relevance and minimise bias.

Quantitative analysis: Surveys and administrative data

A detailed IPE quantitative analysis plan was developed by the project lead in advance, in consultation with the project statistician and project director. This outlined both the descriptive and inferential analysis required for each dataset to answer the IPE RQs. Cross-tabulations in relation to group allocation, role, and/or school-level proportion of FSM were requested where this was deemed likely to provide further insight into schools’ experiences. Descriptive analyses on the computer game and Session Delivery Logs data were conducted using R, while analysis of the cost and training attendance data were conducted in an Excel spreadsheet. The remaining IPE quantitative analyses were conducted in Stata (StataCorp LLC, College Station, Texas, USA). All code, the Excel spreadsheet as well as the outputs were reviewed and checked by another experienced analyst at NFER.

Survey data on self-reported teacher outcomes was statistically tested for differences between a small number of single items at baseline and endpoint (selected based on their relevance to the Theory of Change), comparing intervention and control groups using Mann-Whitney tests (treating the Likert-type scales as ordinal data). This allowed us to determine

[19] Intelligent verbatim transcription excludes fillers and redundancies that do not add meaning to the content to make the text more ‘readable’ (McMullin, 2023).

whether the change from baseline to endpoint for teachers in programme schools differs significantly from the same change for teachers in control schools. Some items included at baseline were removed from the endpoint survey due to high levels of reported understanding of the topics that significantly restricted the scope for change between baseline and endpoint. These items related to using part–whole relationships to solve problems, and the inverse relationship between addition and subtraction.

Triangulation of qualitative and quantitative data

The Project Lead developed an integrated analysis framework mapping out how each of the data sources fed into analysis and reporting for each of the IPE RQs. Findings from each of these data sources were then examined in tandem to ensure their integration when responding to each RQ in the final report. The IPE researchers collated and triangulated all data sources through an analysis workshop to ensure they were providing a comprehensive assessment of the implementation effectiveness and perceived outcomes of the Mathematical Reasoning programme. An emerging findings workshop with the full team then allowed for the IPE results and impact evaluation results to be brought together in a context where the team could reflect on how the IPE findings might inform our understanding and interpretation of the impact evaluation results.

Costs

Data collection

Data collection for the cost evaluation was embedded within the IPE activities to minimise the burden on schools while enabling triangulation from various sources. The surveys and Session Delivery Logs were used to collect information on the amount of teacher and TA time spent completing the training, attending the webinars, and delivering the sessions, as well as any paid cover for these activities. The surveys also asked about the prerequisite resources and whether schools had had to purchase any for the purpose of the programme.

All of the 11 case study schools were asked to complete a detailed pro forma outlining time spent on lesson preparation compared to usual practice, time spent liaising with the TL, and extra additional costs incurred as part of programme delivery—for example, printing or purchasing extra manipulatives. The pro forma was completed either online or as part of a remote interview.

Programme costs were provided by the Oxford University team. As the focus is on cost to schools, we have not considered the cost of TL training separately but as accounted for within the cost to schools.

Analysis

We conducted the cost evaluation analysis in line with the EEF’s latest cost evaluation guidance (EEF, 2023). When calculating the staff time spent on the programme, we used a range of sources and assumptions, as outlined in Appendix P (see Appendix P Table 1 in Further Appendices).

The cost of programme delivery to each class per year was calculated over a projected three-year period. Having established a cost per-class-per year, this figure was divided by the average number of pupils per class (according to baseline data from the trial) to estimate the cost per pupil per year. Only costs that were reported by at least half of the sample schools were included in the primary cost analysis. However, sensitivity analysis was carried out for any other costs reported by schools, including prerequisites such as manipulatives and devices for the computer games. Sources, methods, and assumptions for these calculations are outlined in Appendix P (see Appendix P Table 2 in Further Appendices).

Timeline

Table 8: Evaluation timeline

Dates	Activity	Staff responsible / leading
December 2023 – January 2024	Intervention Delivery and Evaluation Analysis (IDEA) workshop and set-up meetings Development of recruitment documents	NFER Oxford University
February 2024 – June 2024	School recruitment Teacher Leader Training	Oxford University
July 2024	Publication of protocol, trial registration, and NPD application Pupil data collection from schools	NFER
September 2024	Baseline assessment Baseline BAU survey	NFER
October 2024	Randomisation Share pre-populated Session Delivery Logs with schools	NFER
November 2024	Launch event, training, and first webinars Training data collection	Oxford University NFER
December 2024 – June 2025	Programme delivery Second and third webinars IPE case study data collection (*t1*) Further training data collection	Oxford University NFER
June 2025	Publication of the Statistical Analysis Plan (SAP) Endpoint assessment Endpoint surveys IPE case study data collection (*t2)*	NFER
July 2025	IPE incentive payments Control school payments	NFER Oxford University
September 2025	Endpoint assessment data shared with schools	NFER
January 2026	Submission of draft report	NFER
June 2026	Publication of final report	NFER

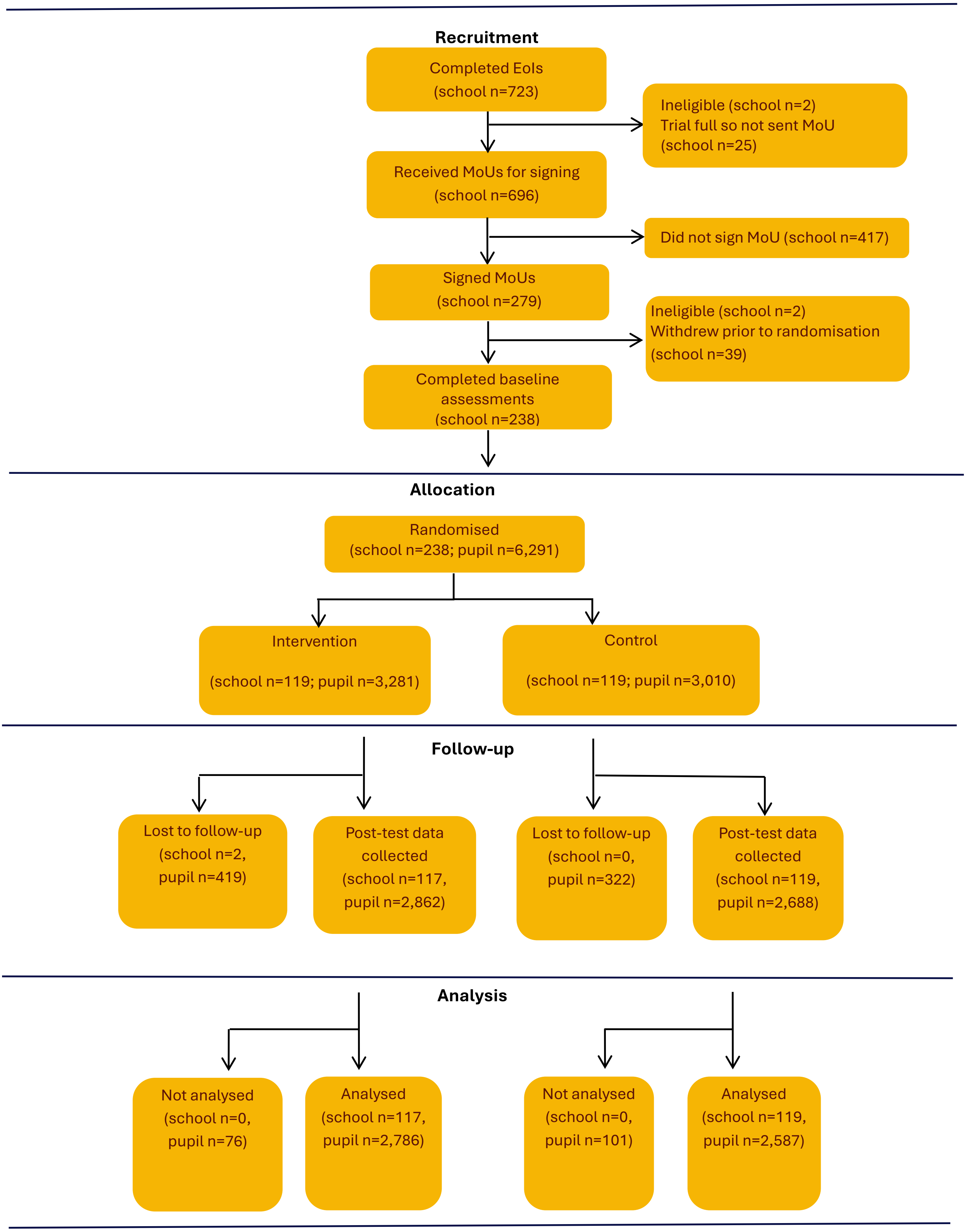
Impact evaluation results

Participant flow including losses and exclusions

The participant flow diagram for this evaluation is shown below in Figure 2. In the initial recruitment phase, 723 schools completed Expressions of Interest (EoIs), of which 696 were sent Memorandums of Understanding (MoUs) to sign. A total of 279 schools signed an MoU. Of these schools, 238 completed baseline data collection and went on to be randomised. This means 41 schools were lost between signing an MoU and randomisation. Two of these 41 schools were found to be ineligible for the trial and 39 withdrew: 11 due to work commitments or lack of time, one school withdrew due to concerns there would be high levels of parent/carer withdrawals, and two schools withdrew due to concerns around scheduling and disruption, while 25 schools gave no reason for withdrawal. The 238 remaining schools were randomised evenly (intervention N=119, control N=119), and the 6,291 pupils in participating Year 2 classes were randomised along with their school (intervention N=3,281, control N=3,010). There were 253 Year 2 classes participating in the trial (intervention N=129, control N=124).

Between randomisation and analysis, 741 pupils did not sit the endpoint PTM7 assessment (intervention N=419, control N=322) and so were lost to follow-up. This was largely due to pupils being absent on the day of the endpoint PTM7 assessment and pupils leaving the school (see 'Missing data analysis' section for details). Two intervention schools withdrew from the trial due to staffing and challenges related to the Office for Standards in Education, Children's Services and Skills (Ofsted) and did not agree to complete endpoint testing, which accounts for 44 of the 741 pupils. A further 177 pupils (intervention N=76, control N=101) were followed up but were not included in the complete case primary analysis due to a missing baseline PTM6 score. The final primary analysis sample, therefore, consisted of 5,373 pupils (intervention N=2,786, control N=2,587), 85.4% of those randomised.

Figure 2: Participant flow diagram



The MDES at the protocol, randomisation, and analysis stages is shown below in Table 9. The MDES calculations at the protocol and randomisation stages in Table 9, do not include anticipated attrition. If the 10% school-level and 15% pupil-level attrition we anticipated are included, the MDES at these stages increases to 0.117. At the analysis stage, some parameters in the MDES calculation were more favourable than anticipated in the protocol: there was a lower rate of school-level attrition and higher pre-post correlations. However, the ICC was much higher than anticipated, and as a result, overall, the MDES among all pupils increased to 0.131 at the analysis stage (after all attrition). It is not clear why the observed ICC of 0.21 was so much higher than 0.11, a figure obtained from the first effectiveness trial of Mathematical Reasoning. One possible explanation is that the recruitment process was different in the first effectiveness trial: schools were recruited by the NCETM through eight 'Maths Hubs', each of which covered one or more counties (e.g. Sussex). This may have led to schools being more similar to each other on average compared to the current trial, which recruited nationally. Further details on the parameters used in MDES calculations are provided in the 'Sample size' subsection within the 'Methods' section above.

Table 9: MDES at different stages

		Protocol		Randomisation		Analysis	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.108	0.140	0.108	0.138	0.131	0.161
Pre-test / post-test correlations[a]	Level 1 (pupil)	0.62	0.62	0.62	0.62	0.75	0.73
	Level 2 (school)	0.62	0.62	0.62	0.62	0.68	0.59
ICCs	Level 2 (school)	0.11	0.11	0.11	0.11	0.21	0.20
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		25.7	6.8	26.4	7.2	22.8	6.0
No. of schools	Intervention	120	120	119	119	117	117
	Control	120	120	119	119	119	119
	Total:	240	240	238	238	236	236
No. of pupils	Intervention	3,084	816	3,281[b]	857	2,786	714
	Control	3,084	816	3,010	867	2,587	710
	Total:	6,168	1,632	6,291	1,724	5,373	1,424

[a] More accurately, these figures are the square root of the proportion of variance explained at level 1 and level 2, as explained in the 'Sample size' subsection within the 'Methods' section above. We use 'pre-test/post-test correlations' in the table for consistency with the study protocol and SAP (Flemons et al., 2024; Morton and Smith, 2025).
[b] This is one fewer than the number given in the SAP (Morton and Smith, 2025), due to a duplicate pupil entry in the intervention group that was identified after the SAP was published.

Attrition

The pupil-level attrition rate between randomisation and the primary analysis was moderate at 14.6% (see Table 10). This level of attrition was close to what was expected: a 15% rate had been anticipated in sample size calculations. Attrition was generally caused by failure to complete the endpoint PTM7 assessment, either due to pupils leaving the school or being absent on the assessment day (see 'Missing data analysis' section for details). The attrition rate was slightly higher for pupils randomised to the intervention group (15.1%) compared to those in the control group (14.1%). Further discussion on attrition is provided in the 'Missing data analysis' section.

Table 10: Pupil-level attrition from the trial (primary outcome)

		Intervention	Control	Total
No. of pupils	Randomised	3,281	3,010	6,291
	Analysed	2,786	2,587	5,373
Pupil attrition (from randomisation to analysis)	Number	495	423	918
	Percentage	15.1%	14.1%	14.6%

Pupil and school characteristics

Baseline pupil- and school-level characteristics in each group, as well as national figures, are shown below in Table 11. The national population of schools was defined as primary schools (including infant, middle deemed primary, and all-through schools) in England, including only academies, free schools, and LA-maintained schools. School-level statistics (both national and for trial schools) were obtained from NFER's 'Record of Schools' dataset, which is derived from publicly available sources such as **Get Information About Schools** (Department for Education, 2026a)The national population of pupils was defined as Year 2 pupils attending state-funded primary schools in England. National pupil-level statistics were obtained from 2024/2025 DfE statistical releases (FSM and gender: (Department for Education, 2025); SEND: (Department for Education, 2026b)).[20]

As might be expected, due to randomisation both school- and pupil-level characteristics were generally well balanced between the intervention and control groups. Some of the larger imbalances were seen in the proportion of academies (5% higher in the intervention group) and the proportion of pupils that speak EAL (3.6% higher in the intervention group).

School- and pupil-level characteristics were also similar between the trial arms and the national population in most cases, although FSM pupils were overrepresented in the trial compared to the national population (28.6% vs 24.6%), as were urban schools (76.9% vs 71.0%) and LA-maintained schools (58.0% vs 51.8%). These differences are discussed in the 'Limitations and lessons learned' subsection within the 'Conclusion' section below.

Baseline PTM6 scores can also be seen to be similar between groups in Table 11, in terms of both the mean and spread (SD) of the data. Average PTM6 scores are 0.2 higher in the control group, a difference that can also be expressed as an effect size of -0.051 (95% CI: -0.175 to 0.072). This figure was estimated using a two-level linear model, in which the outcome was PTM6 score and the only predictor was the intervention indicator. The overall distribution of PTM6 scores was also similar between the two groups, as seen in Figure 3 below.

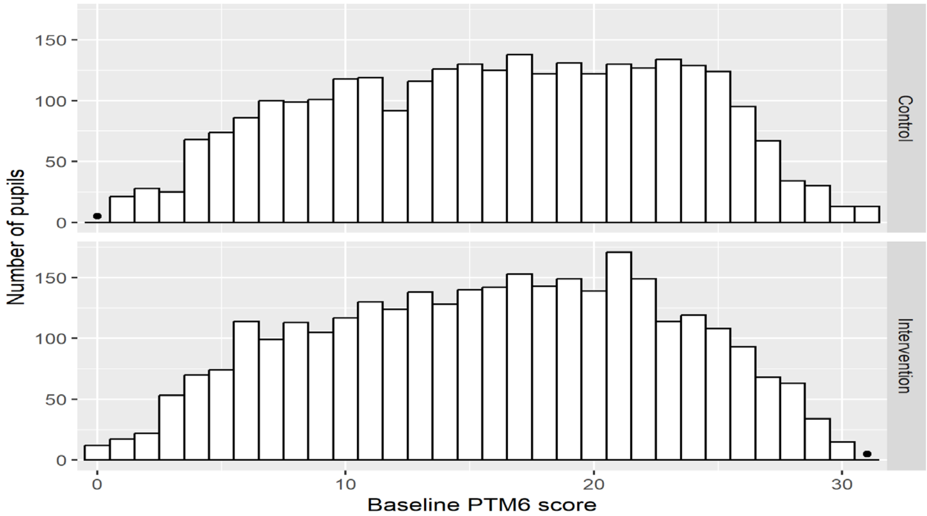
[20] As far as we are aware the national proportion of pupils that speak EAL in Year 2 is not published. The proportion of pupils that speak EAL in state-funded primary schools across all year groups is 22.8%.

Table 11: Baseline characteristics of groups as randomised compared to national figures

School level (categorical)	Intervention group		Control group		National	
	n/N	%	n/N	%	n/N	%
School phase Primary Infant All-through Middle deemed primary Missing	 112/119 6/119 0/119 0/119 1/119	 94.1 5.1 0 0 0.8	 112/119 7/119 0/119 0/119 0/119	 94.1 5.9 0 0 0	 14,624/15,921 1,132/15,921 159/15,921 6/15,921 0/15,921	 91.9 7.1 1.0 <0.1 0.0
School type Academies Free schools LA-maintained schools Missing	 52/119 0/119 66/119 1/119	 43.7 0.0 55.5 0.8	 46/119 1/119 72/119 0/119	 38.7 0.8 60.5 0.0	 7,345/15,921 331/15,921 8,245/15,921 0/15,921	 46.1 2.1 51.8 0.0
Most recent Ofsted Inadequate Requires improvement Good Outstanding Missing	 0/119 3/119 86/119 11/119 18/119	 0 2.5 72.3 9.2 15.1	 0/119 6/119 93/119 5/119 15/119	 0 5 78.2 4.2 12.6	 120/15,921 951/15,921 11,776/15,921 1,210/15,921 1,864/15,921	 0.8 6.0 74.0 7.6 11.7
Urban or rural Urban Rural Missing	 93/119 25/119 1/119	 78.2 21.0 0.8	 90/119 29/119 0/119	 75.6 24.4 0.0	 11,302/15,921 4,619/15,921 0/15,921	 71.0 29.0 0.0
School level (continuous)	N (missing)	Mean (SD)	N (missing)	Mean (SD)	N (missing)	Mean (SD)
% of pupils eligible for FSM	118 (1)	29.1 (17.7)	119 (0)	31.1 (17.4)	15,863 (58)	25.0 (16.0)
% of pupils with SEND	118 (1)	15.0 (6.7)	119 (0)	15.1 (5.4)	15,864 (57)	14.4 (6.2)
% of pupils meeting expected standard in Key Stage 2 maths	109 (10)	74.9 (14.2)	104 (15)	74.6 (13.1)	13,637 (2,284)	75.2 (14.0)
Pupil level (categorical)	n/N	%	n/N	%	n/N	%
Eligible for FSM Yes No Missing	 916/3,281 2,318/3,281 47/3,281	 27.9 70.6 1.4	 886/3,010 2108/3,010 16/3,010	 29.4 70.0 0.5	 150,104 (609,641) 459,537 (609,641) -	 24.6 75.4 -
Gender Male Female Missing	 1,658/3,281 1,576/3,281 47/3,281	 50.5 48.0 1.4	 1487/3,010 1507/3,010 16/3,010	 49.4 50.1 0.5	 310,315 (609,641) 299,326 (609,641) -	 50.9 49.1 -
Has SEND Yes No Missing	 591/3,281 2,643/3,281 47/3,281	 18.0 80.6 1.4	 547/3,010 2447/3,010 16/3,010	 18.2 81.3 0.5	 111,613 (609,641) 498,028 (609,641) -	 18.3 81.7 -
Speaks EAL Yes No Missing	 672/3,281 2,560/3,281 49/3,281	 20.5 78.0 1.5	 510/3,010 2480/3,010 20/3,010	 16.9 82.4 0.7		
Pupil level (continuous)	N (missing)	Mean (SD)	N (missing)	Mean (SD)	N (missing)	Mean (SD)
PTM6 total raw score	3,121 (160)	15.9 (7.0)	2,846 (164)	16.1 (7.1)		

Note: Time-variant characteristics were measured in the 2024/2025 academic year. FSM was defined using FSM eligibility at any point in the last six years, both at the school and pupil level.

Figure 3: Distribution of PTM6 scores in the control and intervention groups



Note: Black dots on the histogram represent pupil counts below ten. These have been suppressed to ensure individual pupils cannot be identified from the figures in this report.

Outcomes and analysis

Primary analysis (RQ1)

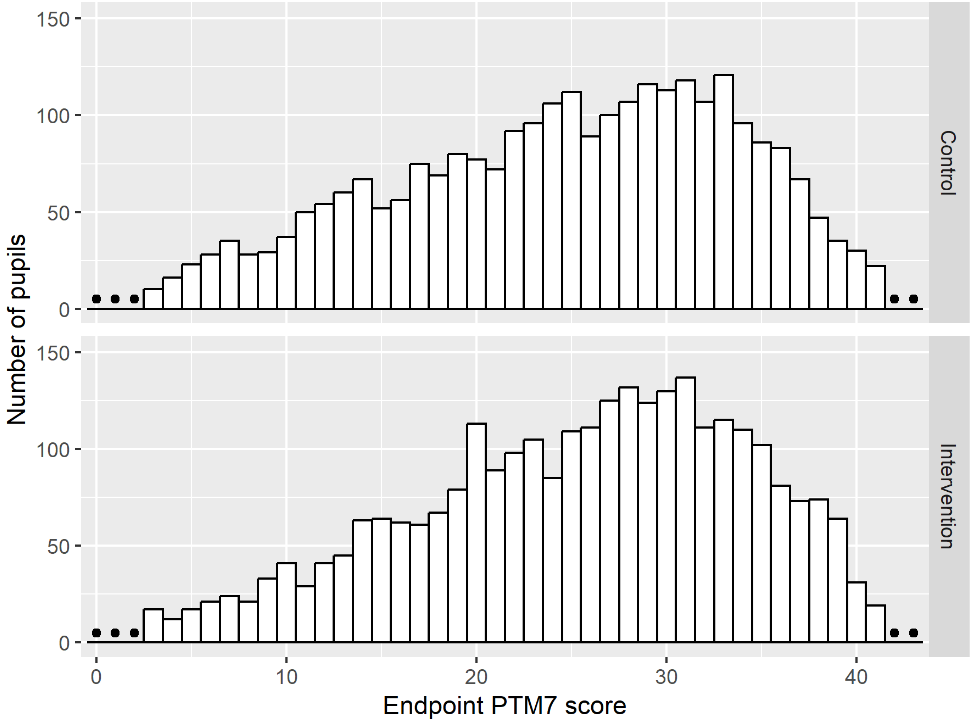
The primary analysis estimated the impact of the Mathematical Reasoning programme on Year 2 pupils' attainment in mathematics, as measured by the PTM7 (RQ1). Based on the primary analysis model, pupils who received Mathematical Reasoning scored an average of 0.84 points higher on the PTM7 than control pupils (95% CI: 0.00 to 1.68), which translates into an effect size of 0.093 (95% CI: 0.000 to 0.185, see Table 12). This is the best estimate of the intervention's impact but a range of small to moderate positive effect sizes are also supported by the data to a lesser extent. Zero impact is just within the 95% CI and equivalently the p-value is slightly over 0.05, so while a hypothesis test at the 5% significance level would fail to conclude that there is evidence of impact, our preferred approach is to consider the full range of values contained in the CI and to treat the p-value as a continuous measure of evidence against the hypothesis of zero impact. Considered in this context, the p-value and range of values supported by the CI would be unlikely to be observed if Mathematical Reasoning truly had zero impact. We therefore, consider it likely that Mathematical Reasoning does have a beneficial effect on maths attainment, as measured by the PTM7.

Figure 4 shows the distribution of PTM7 scores in the control and intervention groups. Intervention scores were slightly higher overall, in line with the primary analysis result, but there were no other visible differences (e.g. no difference in the spread of the distributions). The distribution of PTM7 scores in both groups was platykurtic (more pupils in the 'tails' of the distribution than for a normal distribution) and negatively skewed. This non-normality can sometimes be an indication that the residuals from the corresponding model (here the primary analysis model) will also not be normally distributed. However, inspection of a plot of the residual quantiles against those of a normal distribution suggested that the residuals from the primary and they were close to normally distributed. The distribution of PTM7 scores among FSM pupils follows a similar pattern (see Appendix F in Further Appendices).

Table 12: Primary 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)	Effect size (95% CI)	P-value
PTM7 Total raw score	2,862 (419)	25.32 (24.99, 25.64)	2,688 (322)	24.59 (24.24, 24.94)	5,373 (2,786, 2,587)	0.093 (0.000, 0.185)	0.052

Figure 4: Distribution of PTM7 scores in the control and intervention groups



Note: Black dots on the histogram represent pupil counts below ten. These have been suppressed to ensure individual pupils cannot be identified from the figures in this report.

FSM-eligible subgroup analysis (RQ2.1)

The impact of Mathematical Reasoning in the subgroup of FSM-eligible pupils (RQ2.1) is shown below in Table 13. The estimated impact was somewhat larger in the FSM-eligible subgroup than for all pupils, with an effect size of 0.140 (95% CI: 0.018 to 0.261) observed. It is therefore likely that Mathematical Reasoning has an impact in the FSM-eligible subgroup, although due to the smaller sample size in this subgroup a wide range of positive effect sizes—from very small to large—are possible. A further model was run that included an interaction between the FSM status and the intervention indicator, and the estimate for the interaction coefficient was 0.81 (95% CI: 0.10 to 1.52), which translates to an effect size of 0.089 (95% CI: 0.011 to 0.168). This suggests that there may be a larger effect of Mathematical Reasoning for FSM-eligible pupils compared to non-eligible pupils. This magnitude of the difference seems substantial in the context of the trial, given that it is similar in size to the overall primary analysis effect size (0.093).

Two coefficients from the interaction model (for the intervention and the interaction between FSM status and the intervention) were summed, which is an alternative method for estimating the impact within the FSM-eligible subgroup (as required by the EEF statistical analysis guidance: (Education Endowment Foundation, 2022)). This produced a similar but slightly higher effect size of 0.155 (95% CI: 0.047 to 0.263) compared to the effect size of 0.140 given above. It is common for the two approaches to produce a similar effect size, but the similarity in results slightly strengthens the conclusion that the effect size in the FSM-eligible subgroup lies close to 0.14.

Table 13: Impact of the intervention in the subgroup of FSM-eligible pupils

Group	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention, control)	Effect size (95% CI)	P-value
FSM-eligible subgroup	741 (175)	21.97 (21.32, 22.63)	751 (125)	20.33 (19.67, 20.98)	1,424 (714, 710)	0.140 (0.018, 0.261)	0.025

Secondary analysis (RQ2.2)

The estimated impact of Mathematical Reasoning on the four 'process category' subscales of the PTM7 (RQ2.2), among all pupils and in the FSM-eligible subgroup, is shown below in Table 14. For the 'Mathematical reasoning' subscale there was an effect size 0.105 (95% CI: 0.009 to 0.200). This shows some evidence of a positive impact but is not decisive and should be considered in the context of the caveats described below. There was little or no evidence of any benefit from Mathematical Reasoning for the other three subscales. It is tempting to conclude that only 'Mathematical reasoning' showed a positive result because it is the most aligned with the intervention.[21] However, these results should be interpreted in the context of the CFA (see 'CFA and associated analyses of the PTM7 subscales' in Appendix E of Further Appendices) and the histograms of PTM6 and PTM7 subscale scores (see Appendix F of Further Appendices) as described below.

The histograms suggest floor effects for the 'fluency in facts and procedures' and 'problem-solving' subscales and ceiling effects for the 'fluency in conceptual understanding' subscale, on both the PTM6 and PTM7. The 'fluency in facts and procedures' and 'problem-solving' PTM7 subscales also had relatively poor psychometric properties, with Cronbach's Alpha values of 0.51 and 0.39, respectively (see Appendix E of Further Appendices). It is also worth recalling that the 'fluency in facts and procedures' and 'problem-solving' PTM7 subscales have a small range of possible scores (0–6 and 0–4, respectively), which may limit their ability to fully capture the underlying constructs they purport to represent. In contrast, the 'mathematical reasoning' subscale has a relatively large number of items (20), showed little evidence of floor or ceiling effects, and had a high internal reliability, as measured by a Cronbach's Alpha of 0.84. Considering these facts together, it is notable that the only subscale without any issues ('mathematical reasoning') was also the only subscale that showed evidence of positive impact from the Mathematical Reasoning intervention.

The CFA and accompanying psychometric analyses also suggest that there may be a more fundamental conceptual problem with all four subscales (at least, when applied to the sample for this trial). The four-factor CFA performed well statistically overall, but correlations between the latent factors were very close to 1. Inspection of the correlations between all PTM7 items (see Appendix E Table 2 in Further Appendices) also revealed that while items within the same subscale were positively correlated, generally these correlations were not much higher than for pairs of items from different subscales (see Appendix E Table 1 in Further Appendices). The mean correlation between pairs of items within the 'mathematical reasoning' subscale was 0.25, compared to a mean correlation of 0.22 between mathematical reasoning items and items from other subscales. The equivalent results for the other subscales, reported in the format 'subscale (mean internal correlation, mean correlation with items from other subscales)' were: fluency in facts and procedures (0.17, 0.19); fluency in conceptual understanding (0.25, 0.22); and problem-solving (0.14, 0.17). The CFA latent factor correlations and the item correlations raise the same problem of poor discriminant validity: it is not clear that the four subscales measure distinct constructs (as opposed to just measuring mathematical ability generally).

Overall, we advise caution when interpreting the isolated positive result seen for the 'mathematical reasoning' subscale. We might tentatively conclude that scores on the 'mathematical reasoning' PTM7 subscale seem to be improved by the Mathematical Reasoning intervention. In principle we would wish to draw two more substantive conclusions: i) the 'mathematical reasoning' subscale represents a distinct and meaningful underlying construct; and ii) that this 'mathematical reasoning' construct benefits more from the intervention than the other three constructs. However, there is too much uncertainty to draw these further conclusions: i) is hindered by an apparent lack of discriminant validity for the scale; and ii) is limited by the statistical limitations of the other three scales described above.

[21] It is worth noting again that it is only somewhat aligned: the subscale includes content not covered by the Mathematical Reasoning intervention.

Table 14: Impact of the intervention on the four subscales of the PTM7

Group	Outcome	Unadjusted means				Effect size		
		Intervention group		Control group				
		n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention, control)	Effect size (95% CI)	P-value
All pupils	Fluency in facts and procedures score	2,862 (419)	4.41 (4.35, 4.46)	2,688 (322)	4.33 (4.28, 4.39)	5,373 (2,786, 2,587)	0.061 (-0.046, 0.168)	0.264
	Fluency in conceptual understanding score	2,862 (419)	8.45 (8.34, 8.57)	2,688 (322)	8.21 (8.09, 8.33)	5,373 (2,786, 2,587)	0.066 (-0.031, 0.162)	0.182
	Problem-solving score	2,862 (419)	1.21 (1.17, 1.25)	2,688 (322)	1.24 (1.20, 1.28)	5,373 (2,786, 2,587)	-0.022 (-0.121, 0.077)	0.665
	Mathematical reasoning score	2,862 (419)	11.25 (11.08, 11.42)	2,688 (322)	10.81 (10.63, 10.99)	5,373 (2,786, 2,587)	0.105 (0.009, 0.200)	0.032
FSM-eligible subgroup	Fluency in facts and procedures score	741 (175)	3.97 (3.86, 4.08)	751 (125)	3.81 (3.7, 3.92)	1,424 (714, 710)	0.097 (-0.042, 0.235)	0.172
	Fluency in conceptual understanding score	741 (175)	7.41 (7.17, 7.64)	751 (125)	6.86 (6.62, 7.1)	1,424 (714, 710)	0.112 (-0.019, 0.241)	0.094
	Problem-solving score	741 (175)	0.99 (0.92, 1.05)	751 (125)	0.96 (0.9, 1.03)	1,424 (714, 710)	-0.029 (-0.173, 0.113)	0.689
	Mathematical reasoning score	741 (175)	9.61 (9.26, 9.95)	751 (125)	8.69 (8.36, 9.03)	1,424 (714, 710)	0.185 (0.06, 0.309)	0.004

Analysis in the presence of non-compliance

The distribution of Mathematical Reasoning units attended among intervention pupils included in compliance modelling is shown below in Figure 5. Attendance was generally high, with over half of pupils (N=1,299) attending all units. Most pupils met the conditions for the binary compliance measures used in modelling below: 91% (N=2,022) attended ten or more units, and over 98%[22] attended one or more units. Similarly, among FSM-eligible pupils included in the compliance analysis 87% (N=496) attended ten or more units and over 98% attended one or more units. However, these high attendance rates are among analysed intervention pupils: 27% of intervention pupils (N=757) had missing compliance data and could not be included. In the FSM-eligible subgroup 29% of intervention pupils (N=207) had missing data. Missing compliance data could have introduced selection bias into the results described below, if the relationship between compliance and PTM7 score was different for pupils with missing compliance data, compared to those with complete data.

The impact of the three compliance measures was estimated for all pupils (N=4,802 analysed) and the three models were then repeated for the FSM-eligible subgroup (N=1,282 analysed), with the results shown below in Table 15. For the continuous compliance measure the estimate was 0.075 (95% CI: 0.045 to 0.106), meaning an average increase in PTM7 score of 0.075 per additional Mathematical Reasoning unit attended. A pupil attending all 12 units could therefore expect an average improvement of 0.9 points on the PTM7, compared to if they had not attended any units.

The effect size for pupils attending ten or more units was 0.102 (95% CI: 0.060 to 0.144), only slightly higher than the primary analysis effect size of 0.093 (95% CI: 0.055 to 0.131). As discussed in the 'Methods' section above, we expected this result to be upwardly biased due to the exclusion restriction. However, the degree of this bias depends on the proportion of pupils

[22] The exact number of pupils is not given, to prevent disclosure of individual pupil data.

that meet the compliance condition: the higher the proportion, the less severe the bias. As 91% of intervention pupils attended ten or more units, bias becomes a relatively minor consideration in this case. The unbiased estimate should be higher than the 'lower bound' created by the third compliance measure of one or more units, which had an effect size of 0.093.[23]

In the FSM-eligible subgroup all compliance measure estimates were over double the size seen among all pupils. To some extent this agrees with the ITT FSM-eligible subgroup result, which found a larger effect size (0.140) than the primary analysis (0.093). However, even given these results, it is notable how much higher the compliance estimates are in the FSM-eligible subgroup. It seems likely that this difference in compliance estimates between all pupils and the FSM-eligible subgroup is at least partially real but could also be partially artificial (e.g. due to bias caused by FSM pupils with missing compliance data).

Figure 5: Distribution of the number of Mathematical Reasoning units attended by pupils, among the N=2,215 intervention pupils included in the compliance modelling

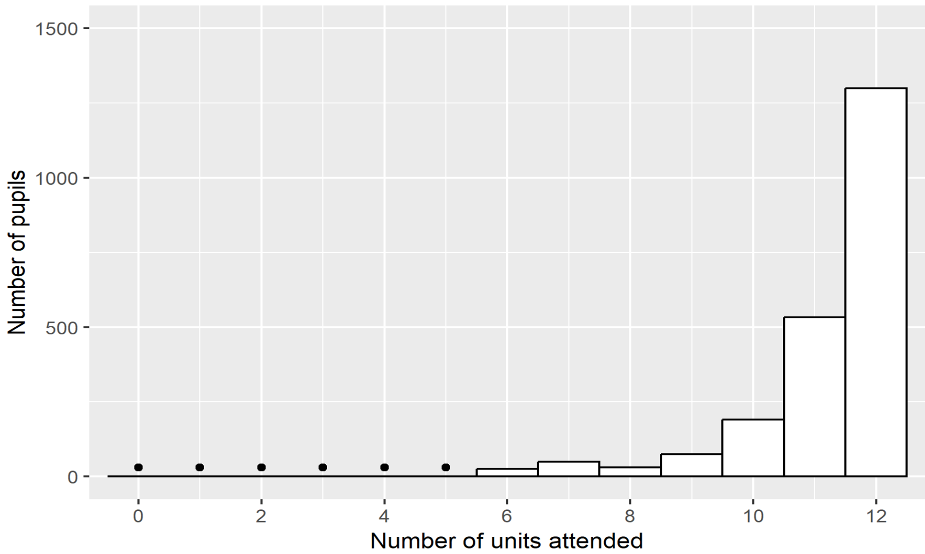


Table 15: Estimated impact of compliance from stage two of the instrumental variable two-stage least squares regressions

Group	Compliance definition	Estimate (95% CI)	Effect size (95% CI)	P-value
All pupils	Change per unit attended (continuous)	0.075 (0.045, 0.106)	Not applicable	<0.001
	Ten or more units attended (binary)	0.925 (0.546, 1.304)	0.102 (0.060, 0.144)	<0.001
	One or more units attended (binary)	0.845 (0.499, 1.192)	0.093 (0.055, 0.131)	<0.001
FSM-eligible subgroup	Change per unit attended (continuous)	0.170 (0.105, 0.234)	Not applicable	<0.001
	Ten or more units attended (binary)	2.139 (1.326, 2.953)	0.232 (0.144, 0.321)	<0.001
	One or more units attended (binary)	1.856 (1.149, 2.564)	0.202 (0.125, 0.278)	<0.001

Results from the first stage of the instrumental variable two-stage least squares regressions are shown in Table 16. The purpose of these results is to assess the association between the intervention indicator (the instrument) and the compliance variable: weak association leads to wide CIs and can also cause bias. However, there was clearly no such problem for this trial, as indicated by very large F-test statistics across all models. This finding is to be expected, given that control pupils could not access the intervention and most intervention pupils attended the majority of Mathematical Reasoning units.

[23] This means that in practice the third compliance measure was not useful as a lower bound, as we already know it is very unlikely that the unbiased compliance estimate will be lower than the primary analysis result.

Table 16: Instrument strength from stage one of the instrumental variable two-stage least squares regressions

Group	Compliance definition	Intervention indicator in stage 1 model	
		Estimate(95% CI)	F-test statistic (p-value)
All pupils	Change per unit attended (continuous)	10.891 (10.815, 10.967)	78325 (<0.001)
	Ten or more units attended (binary)	0.875 (0.863, 0.887)	20081 (<0.001)
	One or more units attended (binary)	0.986 (0.982, 0.991)	206537 (<0.001)
FSM-eligible subgroup	Change per unit attended (continuous)	10.453 (10.284, 10.622)	14711 (<0.001)
	Ten or more units attended (binary)	0.803 (0.775, 0.83)	3324 (<0.001)
	One or more units attended (binary)	0.985 (0.977, 0.993)	53428 (<0.001)

Missing data analysis

N=920 randomised pupils (14.6%) were not included in the primary analysis due to missing data from all sources (intervention N=497, 15.1%; control N=423, 14.1%). This was mainly due to missing outcome data: N=741 pupils (11.8%) had a missing PTM7 score (intervention N=419, 12.8%; control N=322, 10.7%). This small but noticeable difference in missing PTM7 scores between the two groups was due to more school withdrawals and excluded pupils in the intervention group (see Table 17). N=324 pupils (5.2%) had a missing PTM6 score (intervention N=160, 4.9%; control N=164, 5.4%). The multiple imputation analysis below accounts for this observed imbalance in attrition rates, as it reduces the imputed PTM7 outcome of intervention pupils with missing data due to school withdrawals and excluded pupils.

Multilevel logistic regressions were run in which the outcome was whether a pupil's baseline PTM6 score was missing (1) or not missing (0). Similarly, regressions were run in which the outcome was missingness of endpoint PTM7 scores (see Appendix G in Further Appendices for outputs from both regressions). Among the auxiliary variables (those not in the primary analysis), having SEND and the proportion of half-days absent were positively associated with missingness of both PTM6 and PTM7 scores, using p-values of less than 0.05 as a cut-off rule. Re-running the primary analysis with these variables included as covariates (as indicated by the EEF statistical analysis guidance: (Education Endowment Foundation, 2022)) produced a very similar effect size of 0.094 (95% CI: 0.003 to 0.186).

We conducted further sensitivity analysis allowing the possibility that outcome data was MNAR. The complete approach used is described in the 'Methods' section above but, to recap, three missing data patterns were considered possible:

a) Outcomes are MAR, conditional on available covariates.

b) Outcomes are MNAR and on average missing outcome scores are lower than non-missing scores by 1 SD. Missing scores are 'equally worse' between the intervention and control groups.

c) Outcomes are MNAR for intervention pupils only: intervention pupils with missing outcomes have their score reduced to the average among control pupils. This models a situation where, among pupils with missing outcomes, intervention pupils do not receive any benefit from the intervention.

The missing data pattern assumed for a pupil with a missing outcome depended on which of five absence codes they had on their record for the PTM7 (see the first two columns of Table 17). Multiple imputation that incorporated patterns (a–c) was performed for pupils with missing data to investigate whether the primary analysis result was robust to what we deemed a 'worst case' scenario (worst in terms of reducing the primary analysis result). The SEND and proportion of absences variables were included in the imputation procedure.

The estimated impact of Mathematical Reasoning derived from the multiple imputation was 0.72 (95% CI: -0.14 to 1.59), a reduction of 14% compared to the primary analysis estimate of 0.84 (95% CI: 0.00 to 1.68). This estimate from the multiple imputation converts to an effect size of approximately[24] 0.79, compared to the primary analysis effect size of 0.93. The size of the intervention impact from the multiple imputation analysis remains similar to the primary analysis: a small but positive effect that equates to one month of progress in EEF's **Teaching and Learning Toolkit** (Education Endowment Foundation, 2026). As this was also a 'worst case' scenario, we consider the size of the primary analysis result to be fairly robust to any bias that could have been introduced by missing outcome data. The CI from the multiple imputation includes a range of small negative values, somewhat decreasing the certainty with which we can conclude there is an intervention effect, although on balance it still seems likely that an effect does exist.

Table 17: Reason recorded for why PTM7 outcomes were missing

Reason for missing outcome	Missing data assumption	Intervention N	Control N
School withdrew	Pattern (c) MNAR	44	0
Pupil left school	Pattern (c) MNAR	115	120
Pupil did not want to complete test	Pattern (b) MNAR	15	12
Pupil present but excluded	Pattern (b) MNAR	56	33
Pupil absent	Pattern (a) MAR	160	153

Additional analyses and robustness checks

Impact of the number of sessions attended (RQ2.3)

The impact of the number of Mathematical Reasoning sessions attended is estimated below in Table 18. The increase in PTM7 score per session attended was estimated to be 0.368 (95%: 0.151 to 0.586). This was deemed remarkably high considering that the benefit per unit attended was only estimated to be 0.075 in the compliance analysis (Table 15; the number of sessions can be higher than the number of units, but the two numbers were similar for most pupils). Exploratory checks were performed to provide context for the result.

First, the number of sessions attended was plotted against the PTM7 score and a smoothed line[25] was drawn through the raw data points (that is, not dependent on model covariates), as seen in Figure 6. The shaded region indicates a 95% CI for the mean PTM7 score at each number of sessions attended. Underlying data points are not shown to ensure individual pupils cannot be identified from the figures in this report. The plot highlighted the fact that most pupils (96% of those included in regression modelling) attended between 10 and 15 sessions, as evidenced by the much narrower CIs within this range. It also highlighted that within the 10–15 sessions range the increase in PTM7 score per session was smaller than outside it. There is even some evidence of a slight decreasing trend around 13–15 sessions.

As a sensitivity check, a further model was run, restricted to pupils who attended between 10 and 15 sessions. In this range the increase in PTM7 score per session was much more modest, at 0.076 (95% CI: -0.259 to 0.412). It therefore, appears that the analysis was highly influenced by outlying pupils who attended fewer than ten or more than 15 sessions. Whether these outliers should be included in modelling is tied up with the question of whether bias is more likely at a very low or high number of sessions. It seems possible that pupils attending very few sessions were missing sessions because they had alternative teaching arrangements or Mathematical Reasoning was not considered appropriate for them (e.g. due to SEND), which could be associated with the PTM7 score. This was reported and/or observed among some of the case study schools as part of the IPE. Fewer than three schools ran more than 15 sessions. Overall, it seems safer to restrict our conclusions to pupils attending between 10 and 15 sessions, where the increase in PTM7 score per session is estimated at 0.076 (95% CI: -0.259 to 0.412) in this range. There is a small amount of data and increased potential for bias outside this range.

---

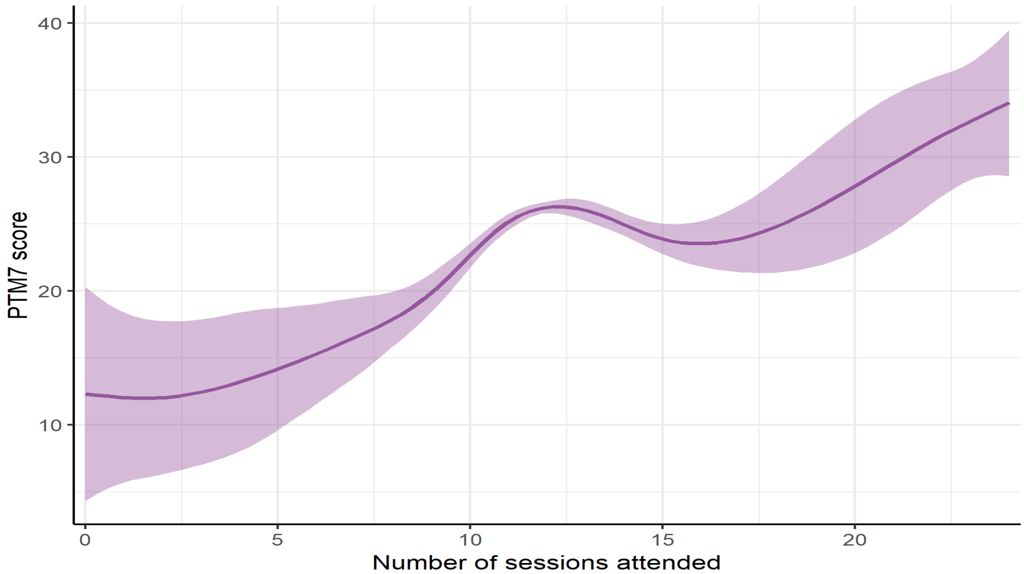
[24] This effect size was obtained by dividing the estimate from the multiple imputation by the denominator used for the primary analysis effect size. This is likely to be close to (but not exactly) correct.

[25] Specifically, the line was estimated using a generalised additive model, which is the default approach used by the ggplot2 function 'geom_smooth' when there are over 1,000 data points (see Wickham *et al.*, no date)).

Table 18: Impact of the number of Mathematical Reasoning sessions attended among intervention pupils

Group	Total n (n missing)	Predictor	Estimate (95% CI)	P-value
Intervention pupils	2,214 (1,067)	No. of sessions attended	0.368 (0.151, 0.586)	<0.001
Intervention pupils in the FSM-eligible subgroup	572 (344)	No. of sessions attended	0.653 (0.338, 0.970)	<0.001

Figure 6: The relationship between the number of Mathematical Reasoning sessions attended and PTM7 score among intervention pupils



Impact of computer game participation (RQ2.4)

Table 19 shows the results from models investigating the impact of the number of computer games played (Model 1), the number of game subject topics (Model 2), and the potential for an interaction between these (Model 3), as well as the corresponding models in the FSM-eligible subgroup (Models 4–6). The 'number of computer games played' counted the same game multiple times if it was played by a pupil multiple times. There was a positive association between the number of games played and PTM7 score seen in Model 1, with an estimated increase of 0.039 points per game (95% CI: 0.021 to 0.056). For a pupil that played 21 games (the median) this would translate into an average benefit of 0.819 points on the PTM7, compared to playing no games. In Model 2, a positive association was also observed between the number of subject topics and PTM7 score, with an estimated increase of 0.421 points per subject topic (95% CI: 0.217 to 0.624). In Model 3, there was no firm evidence of an interaction between the number of games played and the number of subject topics, in terms of their relationship with PTM7 score. Estimates obtained when restricting these models to the FSM-eligible subgroup (Models 4–6) were broadly similar: the Model 4 result was similar to Model 1 result, and so on.

Further models investigated the association between the number of games with 100% scored (Models 7–8) and the number of games with at least 50% scored (Models 9–10), with PTM7 score (see Table 20). The number of games with 100% scored was positively associated with PTM7 score (Model 7) and this result was robust to the inclusion of number of games played as a covariate (Model 8). Similarly, the number of games with at least 50% scored was positively associated with PTM7 score (Model 9) and the result was robust to the inclusion of number of games played as a covariate (Model 10). These results suggest that the positive associations seen in Models 7 and 9 are not just due to correlation with the total number of games played. This indicates that pupils who play large numbers of games need to score highly in at least some of them to have a higher PTM7 score (on average); it is not sufficient to play large numbers of low-scoring games.

As discussed in the 'Methods' section above, Models 1–10 included intervention pupils with a 'green' or 'amber' quality match between their computer game and other data and excluded N=11 pupils with 150 or more computer games played. Sensitivity checks were carried out for Models 1–3 in which the N=11 pupils with 150 or more games played were added to the model (see Appendix H in Further Appendices). These outlying pupils were excluded from the main analysis, as there was evidence of non-effortful game playing and we did not expect a linear relationship between number of games played and PTM7 score to hold above 150 games. The 11 outliers had a notable impact on the estimated increase in PTM7 score per game played from Model 1, lowering it to 0.013 (95% CI: 0.004 to 0.022). This suggests that the association estimated in Model 1 is unreliable for pupils who play 150 or more games, which is a relatively minor caveat given how few pupils appear to do so in practice. The estimate from Model 2 was unaffected, unsurprisingly given that the 11 pupils were outliers in terms of the number of games they played, not the number of subject topics.

Further sensitivity checks were carried out for Model 1–3 in which only pupils with a 'green' quality computer game match were included in modelling (see Appendix H in Further Appendices). Estimates for the number of computer games and number of subject topics increased somewhat compared to Models 1–3. This suggests that some 'amber' pupils may have been incorrectly matched, slightly attenuating the estimates observed in Models 1–3. However, the overall conclusion that number of games played and the number of subject topics are associated with PTM7 score remains the same whether or not 'amber' pupils are included.

Overall, these results provide some support for the conclusion that the impact of the intervention is moderated by computer game use, increasing with the quantity of games played (in particular, high-scoring games) and the number of subject topics. Prior attainment is adjusted for via PTM6 score in all models, which greatly reduces the threat of prior attainment as a confounder. However other potential confounders remain: engagement with the games (and enthusiasm for maths content generally); and receipt of other intervention content (e.g. number of Mathematical Reasoning sessions). As these would cause positive bias, there remains a strong possibility that all computer game results are positively biased to some degree. It is worth noting that the effects reported for different intervention components (number of Mathematical Reasoning units, number of Mathematical Reasoning sessions, number of computer games) are not additive, as these components are highly correlated.

Table 19: Results from six models exploring the relationship between the quantity and diversity of Mathematical Reasoning computer games played by a pupil and their PTM7 score

Group	Model	Predictor	Estimate(95% CI)	P-value
Intervention pupils	1	Number of games played	0.039 (0.021, 0.056)	<0.001
	2	Number of game subject topics	0.421 (0.217, 0.624)	<0.001
	3	Number of games played	0.051 (0.011, 0.091)	0.018
		Number of game subject topics	0.422 (0.113, 0.731)	0.007
		Number of games played × Number of game subject topics	-0.005 (-0.014, 0.003)	0.222
Intervention pupils in the FSM-eligible subgroup	4	Number of games played	0.043 (0.011, 0.076)	0.008
	5	Number of game subject topics	0.416 (-0.018, 0.847)	0.060
	6	Number of games played	0.081 (0.005, 0.156)	0.038
		Number of game subject topics	0.499 (-0.151, 1.147)	0.113
		Number of games played × Number of game subject topics	-0.011 (-0.029, 0.006)	0.203

Note: Models included baseline PTM6 score as a covariate, in addition to the predictors listed.

Table 20: Results from four models exploring the relationship between the scores obtained in the Mathematical Reasoning computer games played by a pupil and their PTM7 score

Group	Model	Predictor	Estimate(95% CI)	P-value
Intervention pupils	7	Number of games with 100% scored	0.086 (0.058, 0.113)	<0.001
	8	Number of games with 100% scored	0.087 (0.046, 0.127)	<0.001
		Number of games played	0.000 (-0.025, 0.024)	0.970
	9	Number of games with at least 50% scored	0.056 (0.037, 0.076)	<0.001
	10	Number of games with at least 50% scored	0.088 (0.043, 0.132)	<0.001
		Number of games played	-0.031 (-0.070, 0.008)	0.125

Note: Models included baseline PTM6 score as a covariate, in addition to the predictors listed.

Prior attainment moderator analysis (RQ2.5)

We explored the differential impact of Mathematical Reasoning at different levels of prior attainment by adding an interaction between the intervention indicator and baseline PTM6 score to the primary analysis model. Results for all Table 21 below, with the 'Intervention group × PTM6 raw score' interaction being the main predictor of interest. The negative interaction suggests that the intervention effect decreases as prior attainment increases (so pupils with lower prior attainment received more benefit from the intervention), with the interaction coefficient estimated at -0.045 (95% CI: -0.091, 0.000). To put this in context, a pupil with the lowest possible PTM6 score (0) could on average expect to gain 1.395 more PTM7 points from Mathematical Reasoning, compared to a pupil with the highest possible PTM6 score (31). This result reflects the higher impact of Mathematical Reasoning estimated in the FSM-eligible subgroup (see 'FSM-eligible subgroup' section above), as FSM status is known to be correlated with lower prior attainment. A degree of caution is warranted when interpreting these results: the models assumed the differential impact to be linear, which is useful for identifying an overall trend, but may overlook more complex patterns within certain bands of prior attainment. For example, if Mathematical Reasoning was effective for pupils with moderately low prior attainment, but ineffective for a small number of pupils with very low prior attainment (e.g. 0–5 PTM6 score) the results here would not capture this. Furthermore, pupils with lower prior attainment on the PTM6 were more likely to have missing PTM7 scores and pupils with SEND were more likely to have missing PTM6 and PTM7 scores (see 'Missing data analysis' and Appendix G). This means that pupils with lower prior attainment are likely to be less well represented in the analytic sample, which weakens the strength of inference that can be drawn for this group.

Table 21: Models estimating the differential impact of Mathematical Reasoning on pupil PTM7 scores at different levels of prior attainment (PTM6 score)

Group	Total n(n missing)	Predictor	Estimate(95% CI)	P-value
All pupils	5,373 (918)	Intervention group	1.574 (0.454, 2.694)	0.006
		PTM6 raw score	0.971 (0.938, 1.004)	<0.001
		Intervention group × PTM6 raw score	-0.045 (-0.091, 0.000)	0.052
FSM-eligible subgroup	1,424 (378)	Intervention group	1.316 (-0.419, 3.053)	0.138
		PTM6 raw score	0.969 (0.901, 1.037)	<0.001
		Intervention group × PTM6 raw score	-0.002 (-0.099, 0.095)	0.968

Impact of teacher training attendance (RQ2.6)

The impact of teacher training attendance, which we defined as completion of the first five training modules and two out of three training webinars, is shown in Table 22. There was no evidence of an impact from teacher training attendance, either for all pupils or the FSM-eligible subgroup. However, it is worth recalling that we defined 'training attendance' as attendance of modules 1–5 and attendance of two of the three webinars. However, of the N=2,018 intervention pupils whose teacher did not attend the training it was just module 5 that was not attended for N=257 pupils. Only a small number of pupils were taught by a teacher that failed to attend modules 1–4 (N=202) or two out of three webinars (N=257). This means the 'training attendance' variable was largely just measuring the impact of attending module 5. It may be that attending modules 1–4 was impactful, but there are too few teachers that failed to attend these modules to quantify the impact in isolation.

Teacher training was not randomised, and results were therefore susceptible to bias from unmeasured confounders. However, it seems that most obvious confounders (e.g. teacher illness, lack of teacher enthusiasm, or staff turnover) would be negatively correlated with both training attendance and pupil outcomes, and so would bias results positively. We therefore do not expect confounding to explain the lack of a positive result in this case: it is more likely to be explained by how training attendance was defined, as described above. We did include the number of pupil Mathematical Reasoning units attended as a covariate, although when we removed this covariate as an exploratory check, the resulting effect size remained similar at -0.045 (95% CI: -0.191, 0.100).

Table 22: Models estimating the impact of teacher attendance of the Mathematical Reasoning training on pupil PTM7 scores

Group	Total n (n missing)	Predictor	Effect size (95% CI)	P-value
Intervention pupils	2,214 (1,067)	Pupil's teacher attended training	-0.031 (-0.190, 0.127)	0.702
Intervention pupils in the FSM-eligible subgroup	572 (344)	Pupil's teacher attended training	-0.128 (-0.341, 0.084)	0.242

CFA on the PTM7 subscales

Confirmatory factor analyses (CFA) were performed to test the suitability of a four-factor structure of the PTM7 items against a one-factor structure. The fit indices of the four-factor CFA model showed that a four-factor structure was a good fit to the data ($\chi2(521) = 6646.232$ (scaled), CFI = 0.937, TLI = 0.933, RMSEA = 0.047, SRMR = 0.059), and factor loadings for each item were sufficient on each factor (fluency in facts and procedures; 0.365 – 0.764, fluency in conceptual understanding; 0.484 – 0.733, problem-solving; 0.376 – 0.642, mathematical reasoning; 0.427 – 0.846). However, latent correlations between the four subscales ranged between 0.942 and 1.023.[26] Correlation values this close to 1 suggest very high overlap and poor discrimination between the subscales. They also reduce the confidence we can have in the accuracy of the four-factor CFA results (e.g. fit indices) in general. A one-factor CFA model was performed on the data, to understand whether this was a better fit to the data than the four-factor model. The fit indices of the one-factor CFA model were similar to those of the four-factor model ($\chi2(527) = 6790.435$ [scaled], CFI = 0.936, TLI = 0.932, RMSEA = 0.047, SRMR = 0.059), and factor loadings for each item were sufficient with loadings ranging from 0.360 to 0.839. However, statistical comparison of the four-factor and one-factor model suggest that the four-factor model provides a better fit to the data than the one-factor model ($\chi2_{\text{difference}}(6) = 157.51$, $p<0.001$). The apparent contradiction between these findings can be explained by the fact that small non-substantive differences in model fit can still produce a small p-value when the sample size is sufficiently large.

Alongside the CFA, reliability estimates were calculated for the four subscales and the overall PTM7 scale. The overall PTM7 scale showed excellent reliability (e.g. Cronbach's $\alpha = 0.90$), however reliability estimates of the individual subscales were varied, ranging from $\alpha = 0.51$ (fluency in facts and procedures subscale) to $\alpha = 0.84$ (mathematical reasoning subscale). Correlations between items ranged between $r = 0.064$ to $r = 0.518$ but were generally lower than expected with multiple

[26] Correlations between factors that are above 1 are possible when performing CFA: these are known as 'Heywood cases' and are often (though not necessarily) due to some form of model misspecification. In this case, it seems likely that the issue is that the underlying factors are very highly correlated, to the point of being almost colinear. It is therefore, unlikely that this issue can be solved, except by specifying a different CFA model (e.g. fewer factors), which would go against the purpose of this analysis.

correlations close to zero. It was expected that there would be stronger correlations between items within the same subscale, supporting the proposed four-subscale structure, however this was not observed in the current data. Altogether, while the CFA results do provide support for the four-factor model of the PTM7, the results should be interpreted in the context of the subscale-level reliability statistics and correlations between items. Analyses are reported in full in Appendix E of the Further Appendices.

Addition of further covariates to the primary analysis

Pupil-level characteristics that may have been imbalanced post-randomisation by chance were added as predictors to the primary analysis model as a sensitivity check. Estimates for all predictors in this model are shown below in Table 23. The pupil- and school-level components of the conditional variance were 27.4 and 9.1, respectively, only a slight reduction compared to 28.4 and 9.5 in the primary analysis model. This means that while several predictors were associated with the PTM7 outcome to some degree, in combination they provided little additional value over and above the PTM6 baseline in terms of explaining PTM7 variance. Most importantly, the intervention group effect size of 0.093 (95% CI: 0.002 to 0.184) was the same as for the primary analysis. This means that the primary analysis result is robust to any chance post-randomisation imbalances in the pupil-level characteristics reported here.

Table 23: Results from a model in which additional pupil-level characteristics are added as covariates to the primary analysis model

Total n (n missing)	Predictor	Estimate(95% CI)	Effect size(95% CI)	P-value
5,365 (926)	Intervention group	0.841 (0.014, 1.668)	0.093 (0.002, 0.184)	0.048
	Eligible for FSM	-0.747 (-1.107, -0.39)	-0.082 (-0.122, -0.043)	<0.001
	Male sex	0.186 (-0.107, 0.478)	0.020 (-0.012, 0.053)	0.214
	Speaks EAL	0.406 (-0.030, 0.841)	0.045 (-0.003, 0.093)	0.068
	Has SEND	-2.702 (-3.130, -2.274)	-0.298 (-0.345, -0.251)	<0.001
	PTM6 raw score	0.899 (0.875, 0.922)	Not applicable	<0.001

Estimation of ICC

The ICCs for the primary and secondary outcomes are given below in Table 24 for all pupils and in Table 25 for the FSM-eligible subgroup. Both conditional ICCs from the full model (e.g. the primary analysis model for PTM7 total score) and unconditional ICCs from a model with no predictors are shown. These figures show that the proportion of variance in the primary outcome explained by differences between schools (before conditioning on covariates) was about 21%. The proportion of variance in the secondary outcomes explained by differences between schools was somewhat lower, ranging from 14% to 21% in Table 24.

Table 24: ICCs for the primary and secondary outcomes, calculated for all pupils

Outcome	Unconditional ICC			Conditional ICC		
	Between-school variance σ_B^2	Within-school variance σ_W^2	ICC	Between-school variance σ_B^2	Within-school variance σ_W^2	ICC
PTM7 total score	17.43	64.91	0.212	9.48	28.39	0.250
Fluency in facts and procedures score	0.35	1.64	0.177	0.28	1.40	0.168
Fluency in conceptual understanding score	2.15	8.33	0.205	1.24	5.19	0.192
Problem-solving score	0.14	0.85	0.140	0.11	0.74	0.133
Mathematical reasoning score	4.05	18.26	0.181	2.58	10.88	0.192

Table 25: ICCs for the primary and secondary outcomes, calculated for the subgroup of FSM-eligible pupils

Outcome	Unconditional ICC			Conditional ICC		
	Between-school variance σ_B^2	Within-school variance σ_W^2	ICC	Between-school variance σ_B^2	Within-school variance σ_W^2	ICC
PTM7 total score	17.04	67.79	0.201	11.02	31.55	0.259
Fluency in facts and procedures score	0.34	1.98	0.147	0.30	1.74	0.146
Fluency in conceptual understanding score	2.26	8.92	0.202	1.52	5.66	0.212
Problem-solving score	0.12	0.65	0.159	0.11	0.62	0.146
Mathematical reasoning score	3.71	19.09	0.163	2.63	11.75	0.183

Implementation and process evaluation results

Usual practice before and during the trial

This section draws on data from the surveys completed by teachers from all schools at baseline and control schools at endpoint, asking about their usual maths teaching practice and recent CPD.

IPE RQ5: *What was business as usual (BAU) in relation to maths teaching, and to what extent did it differ from the Mathematical Reasoning programme?*

Almost all schools followed an external scheme for teaching Year 2 maths, most commonly from White Rose Maths

Almost four-fifths of all respondents to the teacher survey (79%) reported at baseline that their school followed White Rose Maths schemes in Year 2.[27] Almost two-fifths (38%) said they used schemes from NCETM. A small number reported using Power Maths (9%) or another scheme or published resource (14%). Only 2% reported using no externally produced schemes. There was no difference observed between control and intervention groups in this respect.

It was common for both control and intervention schools to already be implementing similar approaches to the Mathematical Reasoning programme at baseline

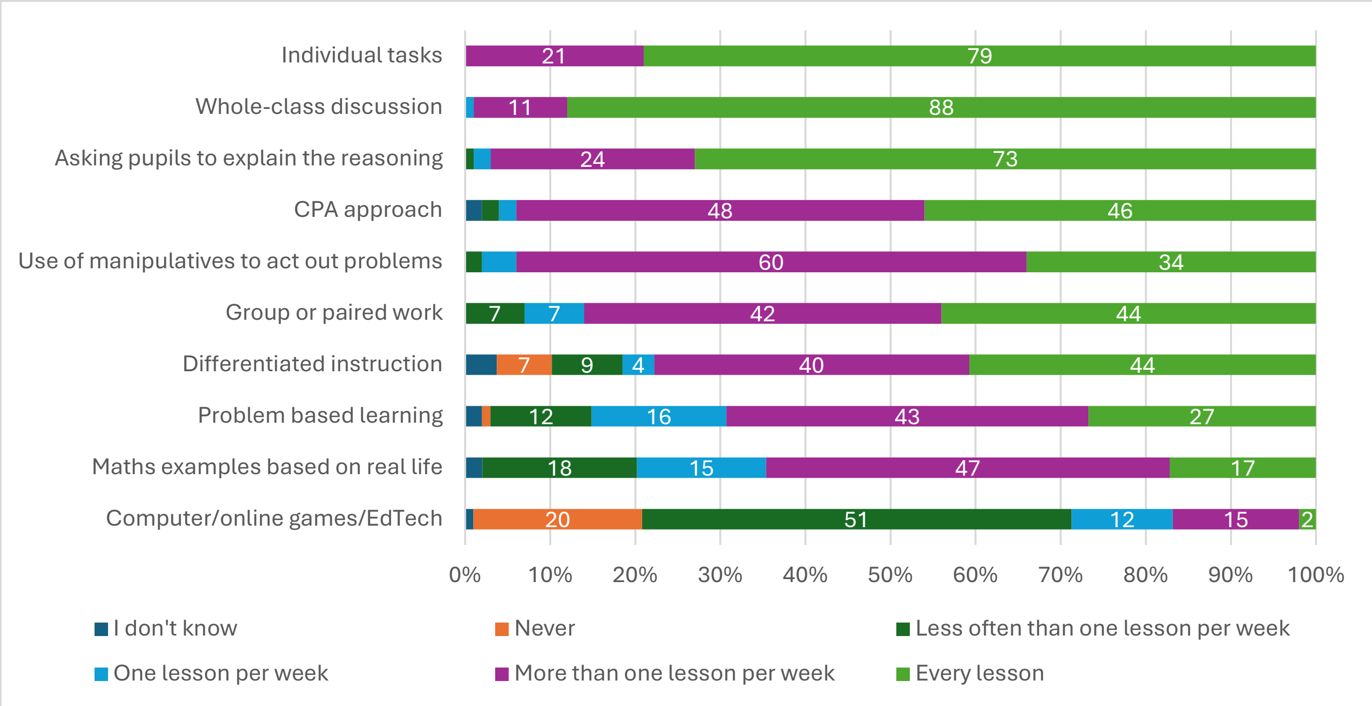
Almost all teachers reported that prior to the trial they already used individual tasks, whole-class discussion, asking pupils to explain their reasoning, concrete pictorial abstract (CPA) approach, using manipulatives and group/paired work in at least one lesson per week (Figure 7).

Problem-based learning, differentiated instruction, and maths examples based on real life were also used slightly less frequently than the other approaches. Less than a third of teachers reported using computer/online games/EdTech in at least one lesson per week (29%).

Teachers in the intervention and control groups reported largely similar use of these approaches at baseline.

[27] Future references to numerical proportions of teachers or TAs should be understood to be based on the survey results, with percentages out of total survey respondents rather than the full trial sample.

Figure 7: Frequency with which different approaches featured in maths lessons at baseline across all schools (intervention and control)



Source: QD.1 – D.10 (How often, if at all, do the following approaches feature in your lesson when teaching maths to your Year 2 class?) from the baseline BAU survey. N=243 (control group n=122, intervention group n=121).

Many schools delivered maths interventions in Year 2, most targeted at pupils requiring additional support

Almost three-quarters of teachers (72%) indicated that their school delivered small group or one-on-one maths interventions for Year 2 pupils. This was reported by a slightly higher proportion of intervention teachers (76%) than control teachers (67%). Almost all teachers in both the control (99%) and intervention groups (98%) reported that pupils requiring additional support were targeted for these interventions. Some schools also reported targeting Pupil Premium pupils (25%) and those requiring extension opportunities (22%).

Prior to the trial, some teachers had recently completed training on similar topics to the Mathematical Reasoning programme, but this did not differ between intervention and control

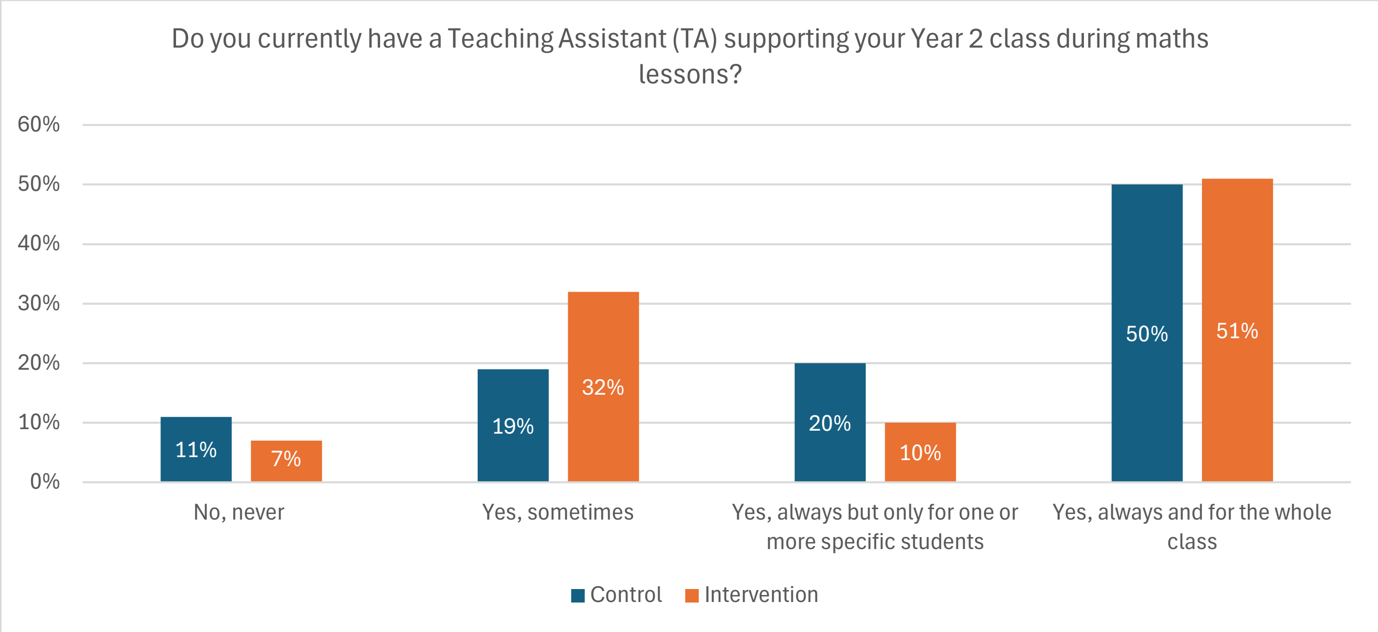
Half of all teachers (51%) reported that they had participated in maths-related CPD opportunities in the 12 months prior to participation in the trial. This proportion was similar across the control and intervention groups. The most common focus of the CPD was mastery learning (35%), followed by problem-solving (13%) and use of manipulatives (12%). These are all areas of particular focus within the Mathematical Reasoning programme, although given the wide variety of ways in which 'mastery' and 'problem-solving' can be interpreted, it is difficult to estimate the extent to which this CPD aligned with that of the programme.

Maths CPD prior to the trial was typically delivered in person (82%), lasted less than a day (51%), and delivered by an external specialist (61%).

Most teachers reported some form of TA support for their usual Year 2 maths lessons at baseline, but the extent and consistency differed slightly between control and intervention

Overall, the prevalence of a TA to support the class was similar across the intervention and control groups at the start of the trial, with only 11% of control teachers and 7% of intervention teachers reporting they 'never' had the support of a TA during maths lessons (Figure 8). Half of teachers at baseline reported that they 'always' have a TA available to support the whole class during maths lessons (51%; Figure 8).

Figure 8: Teacher report of whether they have a TA supporting their Year 2 class during maths lessons at baseline



Source: Question I from the Baseline BAU Survey: 'Do you currently have a Teaching Assistant (TA) supporting your Year 2 class during maths lessons?' (control n=122, intervention n=121).

Practice in control schools over the course of the trial was largely similar to that reported at baseline, although there was more frequent use of some approaches used in the Mathematical Reasoning programme

White Rose Maths remained the most commonly used externally produced scheme to support maths teaching in Year 2 in control schools (used by 76% of teachers), followed by NCETM (used by 41% of teachers).[28] A few of the case study schools that followed White Rose Maths programmes said that as a result reasoning was already incorporated into their maths lessons through their scheme of work, but that the Mathematical Reasoning programme provided additional materials for exploring this topic and focused more on developing pupils' skills.

The frequency with which control teachers drew upon some of the approaches also used in the Mathematical Reasoning programme in their Year 2 maths lessons remained largely unchanged from baseline over the course of the trial, with the exception of slight increases (11–14 percentage points) in the frequency with which problem-based learning and examples based on real life were used. As outlined above, some control teachers completed CPD on these topics in the 12 months prior to the trial, and it is natural that these teachers would want and/or be expected to apply this in their teaching. However, a wide range of approaches can be used in problem-based learning, so this change in control practice between baseline and endpoint does not necessarily indicate a greater similarity to the Mathematical Reasoning programme.

Pupils requiring additional support were most likely to receive maths interventions in both control and intervention schools

The proportion of control teachers who reported that their school delivered maths interventions to pupils in Year 2 was largely unchanged between baseline and endpoint (69%). All teachers who said their school delivered interventions for maths reported they had been targeted at pupils requiring additional support, with similar proportions also reporting interventions targeted at Pupil Premium pupils and those requiring extension opportunities as at baseline. This suggests that teachers did not change the level of formal additional support provided to pupils during the trial.

This appears to be aligned with practice in intervention schools, albeit on the basis of the much smaller sample of case study schools. Several case study teachers reported that small group maths interventions continued to be delivered alongside the Mathematical Reasoning programme. These interventions generally targeted lower attaining pupils and/or those pupils who had been observed to find the main maths lesson challenging. One school had delivered booster sessions

[28] Multiple responses were permitted.

in preparation for SATs (Scholastic Assessment Tests), targeted at pupils 'on the cusp' of meeting expected standards. However, there were also several case study teachers who reported that no additional maths interventions had been implemented.

Some control teachers may have received CPD in topics relevant to the Mathematical Reasoning programme during the trial period
A similar proportion of surveyed teachers (58%) in the control group indicated at endpoint that they had participated in maths-related CPD over the past 12 months as at baseline. The focus of this CPD also remained very similar to what was seen at baseline, with the most common being mastery learning (39%), followed by use of manipulatives (13%), reasoning (10%), and/or problem-solving (8%). A small increase (12 percentage points) was seen over this period in the proportion of teachers who reported having completed training with an external specialist.

Several case study intervention teachers reported having participated in additional maths-related CPD over the trial period, including in relation to reasoning, problem-solving, mastery learning, and manipulatives. This suggests that both control and intervention teachers received relevant training over and above what they did or did not receive as part of the trial.

Overall, while it does appear that control teachers may have somewhat increased their focus on mathematical reasoning over the trial period, there is no indication of significant bias or contamination that could undermine this study's results.

Fidelity, compliance, and adaptation: Training

IPE RQ1: *To what extent was the Mathematical Reasoning programme delivered as intended at scale?*

Training and materials for the TLs, teachers and TAs were delivered as intended

The TL training was designed to involve a combination of in-person days, online sessions, programme delivery, and asynchronous online training courses, and was all delivered as intended (full outline available in Appendix I of Further Appendices). Based on the training day observation, the TL training appeared to be an open, collaborative process, focused not just on teaching the TLs about the programme but also a co-construction of the TL role and webinar content through a workshop approach. The TLs were active participants in the training by asking questions, suggesting improvements to the teacher training and programme requirements, and raising potential issues or concerns. In some cases, the delivery team responded to TL feedback by making changes to programme materials or specific messaging. For example, the logistical challenge of only some pupils using the extension sheets during the whole-class component was raised by the TLs as part of their training and their use changed to be explicitly optional as a result.

Teachers and TAs were required to complete the online training course, which covered both the theory behind the programme and the practicalities of delivering it. The course consisted primarily of four to five videos per module delivered in a lecture-style format—each one lasting between three and ten minutes. These videos were interspersed with interactive activities that required each teacher to post a response sharing their reflections. The training platform also provided a number of example videos of TLs delivering Mathematical Reasoning sessions to groups of children. Additional resources included 'research briefs' which summarised the relevant research to support further thinking about the topics and as a future resource. These were not a required part of the training course, but some were covered as part of the webinars.

Each TL delivered three webinars for their cohort of teachers and TAs as planned. The webinars were intended to provide support for practical challenges, create a community of practice, and support integration of learning into practice beyond the programme. The observed webinars primarily covered the logistics and practicalities of delivery. They also offered the opportunity for teachers to have their questions answered by TLs particularly in relation to challenges and potential adaptations.

In each webinar, the TL went through the slides prepared by the delivery team and facilitated discussion and questions from the teachers—both directly and via the chat function. Evidence from the observations suggested that there was a broadly high level of consistency between the webinars delivered by different TLs, with only minor tweaks to the slides (e.g. more visuals) and some variation in time spent on different parts of the webinar, depending on what the TL perceived to be of greatest concern for the attending teachers. TLs often drew on their own experience of delivering programme sessions with Year 2 pupils to provide examples.

Finally, in addition to presenting the webinars, TLs were described by themselves and by teachers as having four primary functions that were aligned with the programme's intentions:

1. Supporting with technical and administrative issues, particularly getting teachers set-up on and familiar with the different platforms prior to the start of delivery.
2. Providing schools with encouragement and reassurance.
3. Giving advice on acceptable adaptations to overcome challenges while also promoting fidelity.
4. Acting as a first port of call for questions from teachers and liaising with the delivery team and evaluation team for answers as needed.

Completion rates of the core modules were high, but most teachers did not complete all compulsory modules, and TA participation was low

Teachers and TAs were required to complete modules 1–5 of an online training course before starting delivery. Modules 6–8 were optional, but module 9 was a requirement ahead of the final webinar. Each class had one training platform account, with the expectation that teachers and TAs complete all the modules together (using the same account). Some schools also chose for teachers from more than one class to complete the training through a single account. Each module included one or more interactive activities that required the participant to submit a response that their TL and other schools supported by the same TL would be able to see.

TLs were required to confirm that modules 1–4 had been completed on an account before allowing that account access to the programme materials. As a result, completion rates for these modules were very high (92% of schools).[29] However, due to account sharing, this does not necessarily mean that both the teacher and TA for each participating class each covered all the content. The 8% of schools that did not complete modules 1–4 withdrew from delivery.

Completion rates were lower for the other 'compulsory' modules (5 and 9), with only 27% of school accounts completing all 'compulsory' modules (see Table 26). While each account had to submit a response to at least one activity per module for it to be judged 'complete', TLs reported that limited interaction took place via the training platform as participants generally did not return to look at any responses given by the TL or their peers, with one TL reporting that teachers approached the activities as a 'tick box' exercise.

Table 26: School-level completion rates for the compulsory modules of the training course

	Modules 1–4	Module 5	Module 9	All compulsory modules
Completion rate (n=118)	92%	37%	53%	26%

Source: Training completion data provided by the delivery team (n=118).

Completion rates of the optional modules (6–8) by the case study schools were low, primarily due to time constraints and competing priorities. The few teachers who completed the optional modules had mixed views in terms of how worthwhile they found them.

As outlined above, it is not possible to gauge TA participation in the training course from the training completion data. Moreover, while 87% (55 out of 63) of the TAs who completed the endpoint survey reported having done at least some of the training course, only around half of the TAs (63 TAs out of the 128 intervention classes) completed this survey. As a result, the overall TA training participation rate may be different. Of those TAs who reported doing at least some of the training course, 66% of them (36 out of 55) reported always doing it with the teacher, as intended.

[29] A module was judged to be 'complete' if the teacher had responded to at least one of the interactive questions. This means it is possible a teacher may have read through the module without responding to any questions and would not be marked as having completed it. Given that some teachers within the same school shared accounts, training completion data is only available at the school level.

Participating schools were advised that teachers and TAs would require approximately a day and a half (7.5 hours) of protected time to complete the online training course. Of those surveyed teachers who did at least some of the training course, around half (53%) reported spending between one and one and a half days on the course (see Table 27), as intended. Around one-third (29%) spent more time than this.

Table 27: Time spent on the training course by role

	Less than one day (five hours)	One to one and half days (five to seven and a half hours)	More than one and a half days (eight or more hours)
Proportion of TAs (n=55)	49%	42%	9%
Proportion of teachers (n=101)	18%	53%	29%

Source: Question B.1 from the Endpoint Teacher and TA Survey: 'Approximately how long did you spend doing the online training course (Canvas)?' (N=55 TAs, N=101 teachers).
Note: Only those teachers and TAs who responded to the Endpoint Teacher and TA Survey and indicated having completed at last some of the training course are included in these numbers.

Teacher attendance at early webinars was high but levels of active participation in them varied

Webinar attendance data showed that staff from 59% of intervention classes attended all three webinars. Although intended practice was for attendees to take part in the live webinars so they could access the interactive elements of the sessions, staff from a further five classes were only able to watch recordings. Overall, TA attendance was low, with 21% (n=13) of the 63 TAs who responded to the endpoint survey reporting that they attended all three webinars, while almost a third (n=19) said they attended none (Table 28). Of the 44 TAs who reported in the endpoint survey that they attended at least one webinar, two-thirds (n=28) reported always attending with the teacher.

The observations of the webinars captured varying levels of teacher participation, which appeared to be related to which webinar functions were available to be used: breakout sessions resulted in more interactive discussions than sessions where only the 'chat' was used. Some TLs reported that they found it hard to get the teachers to speak during the sessions, that they felt engagement levels were low or that they felt unable to gauge this in a remote environment.

Table 28: Number of webinars attended, by role

	No webinars	One webinar	Two webinars	Three webinars
Proportion of TAs (n=63)	30%	14%	35%	21%
Proportion of teachers (n=102)	1%	10%	25%	64%

Source: Question B.2 from the Endpoint Teacher and TA Survey: 'How many webinars with your Teacher Leader did you attend?' (N=63 TAs, N=102 teachers).
Note: Only those teachers and TAs who responded to the Endpoint Teacher and TA Survey are included in these numbers.

All interviewed TLs were very positive about the training they received and felt it had prepared them to feel confident in their role

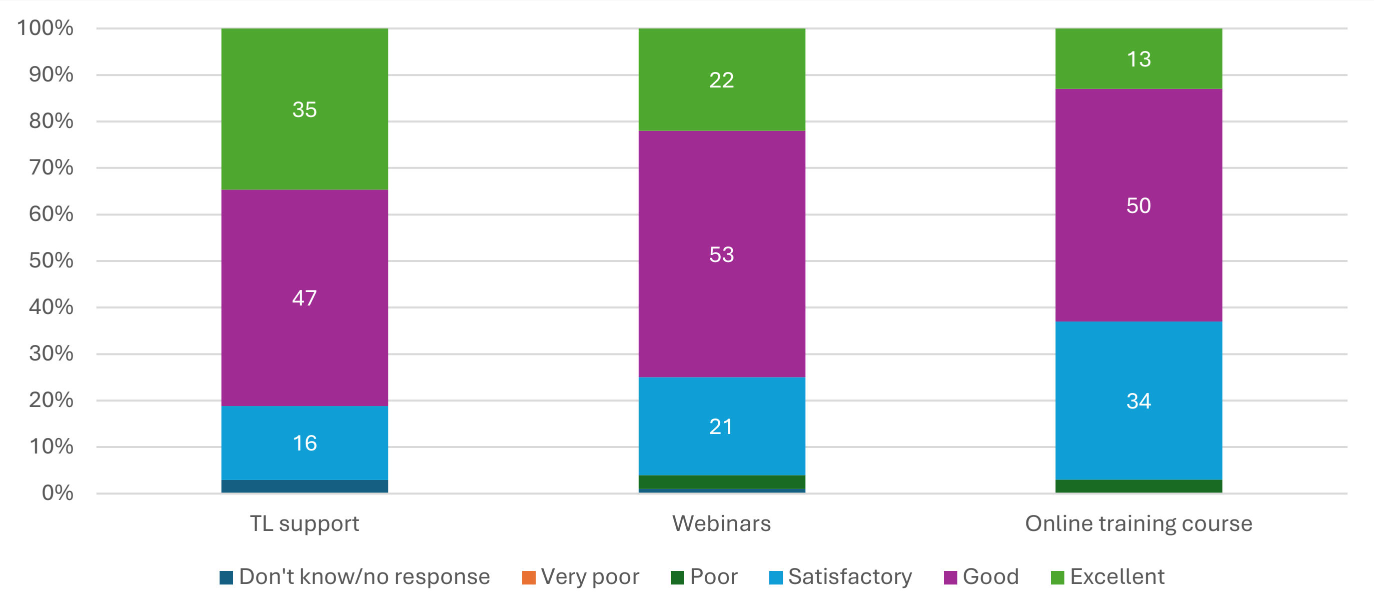
Most of the TLs valued having the opportunity to teach some of the sessions to groups of pupils outside of the trial as part of their training. Some TLs felt they would have benefited from teaching the whole programme and with a whole class (rather than just a group of pupils). TLs also reported that the group virtual check-ins with the Oxford University team before each set of webinars were helpful preparation for webinar delivery, while the in-person training days provided opportunity for discussion and development of subject knowledge. The opportunity for TLs with less experience of teacher training to shadow more experienced TLs was also highlighted by TLs as a valuable aspect of the TL training.

TLs made a few minor suggestions for improvements to their training. Suggestions included that there could be more emphasis on the practicalities of the role over the theory and that it would have been helpful for the delivery team to provide an icebreaker for use in the webinars, rather than each TL having to spend time coming up with their own.

The majority of staff were positive about the quality of the training and support they received

Teachers were mostly positive about the TL support, with 47% describing it as 'good' and a further 35% as 'excellent' (Figure 9). Three-quarters of teachers described the quality of the webinars as at least 'good' (53% 'good', 22% 'excellent'; Figure 9). Finally, almost two-thirds of surveyed teachers reported the online training course to be of at least 'good' quality (50% 'good', 13% 'excellent'; Figure 9).

Figure 9: Teachers' perceptions of the quality of training and support



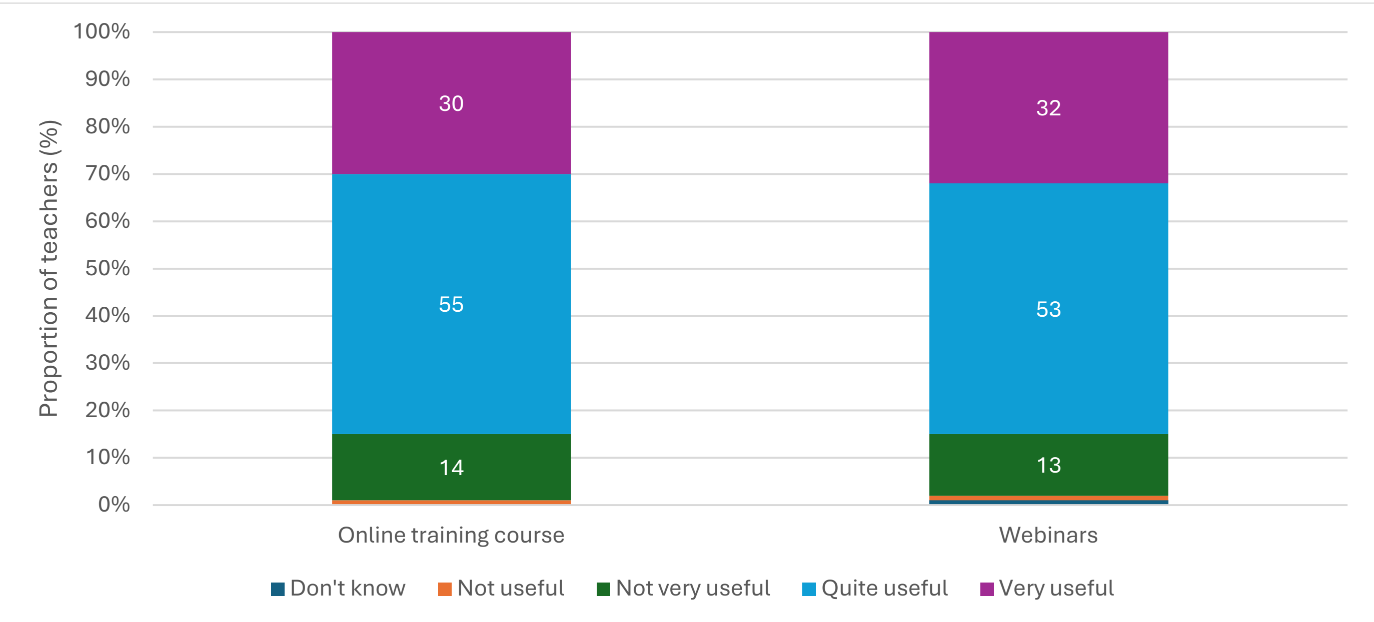
Source: Question F ('How would you describe the quality of...(1) the online training course (Canvas), (2) the webinars with your Teacher Leader, (5) the nature and level of support you received from your Teacher Leader') from the Endpoint Teacher and TA Survey (N=102 teachers).

Most teachers and TAs found the training and webinars useful

Most of the teachers that completed the endpoint survey (85%) felt the training was 'quite useful' or 'very useful' for preparing them to deliver the programme (55% and 30%, respectively; Figure 10). Teachers expressed very similar views on the usefulness of the webinars, with 53% describing them as 'quite useful' and 31% as 'very useful' (Figure 10).

TAs had similar perceptions of both the training course and webinars as the teachers. Half (n=28) of the 55 surveyed TAs who reported having completed at least some of the training course reported finding it 'quite useful', and a further third (n=19) reported finding it 'very useful'. Very similar results were seen from the 44 surveyed TAs who reported attending at least one webinar, with half (n=24) finding the webinars 'quite useful', and a further third (n=13) finding them 'very useful'. The case studies delved further into how participating schools felt about the training and support, described below.

Figure 10: Teacher perceptions of the usefulness of the training



Source: Question E.1 ('How useful did you find the online training course (Canvas) for preparing you to deliver the programme?') and E.2 ('How useful did you find the webinars with your Teacher Leader for supporting programme delivery?') from the Endpoint Teacher and TA Survey (N=101 teachers).

Several of the case study teachers reported finding the training course enjoyable and helpful in preparing them for delivery. One teacher explained:

> *I found the training has been very, very good [...] it's not my most confident area [...], so some of it I found quite challenging but still useful and enjoyable at the same time*. (Teacher, CS8)

Several case study teachers reported finding the videos of delivery particularly helpful as they allowed the teachers to see what delivery should look like in practice. A few also reported appreciating the broader CPD element, noting that the training had made them think about the mathematical concepts in a different way to how they usually would and that they had appreciated understanding more about the theory behind what they do:

> *It was useful to have a look at why we do things and how things are put into place and, I suppose, the 'behind the scenes' of the training, and like getting more of the children's understanding.* (Teacher, CS11)

Some of the case study teachers who completed the training with a TA, also noted finding this helpful as it provided an opportunity to 'bounce ideas off each other' and to understand how people approach things in different ways, as well as for helping the TAs to better support the pupils.

However, several case study teachers noted that the training course was very demanding and/or time consuming. A few teachers also reported finding the format challenging and/or felt there was too much time spent on the theory:

> *[T]he training videos were quite boring, it was difficult to keep focus. [...] I just felt it was too much talking at you.* (Teacher, CS6)

> *The amount of theory was unnecessary. I learnt all that at university. I just thought, 'why are you talking about all this, you're teaching it as if we're not teachers'.* (Teacher, CS9)

The case study TAs who completed at least some of the training course reported finding it useful and enjoyable, although one noted that it was somewhat repetitive. However, one teacher at a different school noted that their TA had not completed the training course because it was not pitched at the appropriate level for them.

The webinars were felt to be useful for providing practical advice and an opportunity for discussion

Most case study teachers appreciated the opportunity to discuss the programme with other teachers and to hear about others' experience with this during the webinars. This included unpicking challenges and gaining clarification from the TL and from each other, as well as gaining reassurance that other teachers were facing similar situations to their own:

> *It was good talking to other schools, knowing where they were at and how they were adapting the lessons.* (Teacher, CS10)

A minority of teachers were less satisfied with the webinars. For example, one teacher did not find the first webinar very useful as they found it only recapped information that had already been provided in the training course, and another felt that an in-person session would have better facilitated discussion.

There were a small number of researcher-observed cases where information about the programme communicated during a webinar conflicted with other training documentation. For example, in one of the webinars attendees were advised to keep children in the same groups for every session so that they have equal time on the computer games, with the justification that there is not much difference between the groups in terms of the content covered, which is not in line with the adaptive grouping practices advised by the training course videos.

There are also indications that some of the messaging around the programme requirements may not have been sufficiently clear in the training material. For example, a case study teacher reported that they and other teachers had thought they had to complete a whole unit each session, until they were informed in one of the webinars that this was not the case—once they were already partway through delivery. Furthermore, the teacher subsequently completed as much of the unit as they could in each hour-long session but reported in the interview that they then moved onto the next unit in the next session, suggesting that the webinar did not fully clarify this point either. These are anecdotal examples and so we do not have evidence of whether this was a widespread issue across the intervention group.

Teachers were positive about the TL role, but there were mixed views among TLs themselves about the importance of it

Case study teachers were positive about the nature and level of support received from TLs, particularly appreciating their responsiveness and the fact they knew they had someone to ask for help if needed.

However, the TLs varied quite considerably in how they saw their role and its relative importance for programme delivery. Some saw it as being about community and knowledge-building, while others felt it was just delivering the webinars and supporting with logistics. TLs were unsure how useful their responses to teachers' posts on the online platform were, and whether they were used by course participants. Their perception was that in most cases the teachers did not go back onto the platform to look at the responses TLs provided to their posts, although there was at least one example of a case study teacher who found their TL's responses useful.

Fidelity, compliance, and adaptation: Delivery

Most teachers delivered all the units across the intended number of sessions

Of the 104 teachers that returned their Session Delivery Logs, nearly all (97%) delivered the intended number of sessions (between 12 and 15). Most teachers (84%) reported delivering all 12 units. This suggests that some teachers were spreading units across more than one session, as intended, but also that some teachers stopped delivering sessions before they had delivered all the units. At least one case study teacher confirmed they chose to stop delivering sessions before completing the programme:

> *Each session was just too long. We didn't complete the last few sessions, I just gave up. [...] Every session was just as long as the next.* (Teacher, CS5)

Based on the Session Delivery Logs, the average (mean) duration of the programme (including any breaks) was 15 weeks. However, in some cases the duration was up to 23 weeks, aligning with reports from some teachers in the interviews and webinars that they had not been able to deliver a session every week. More deprived schools spent an average of 1.5 weeks longer delivering the programme than less deprived schools. This suggests they may have faced additional challenges fitting the sessions into their schedule as there was no difference in the average number of sessions delivered (also see the findings below, which indicate that it often took longer than the intended hour to deliver the sessions). Based on the webinars, some teachers occasionally delivered two sessions per week to 'catch-up' after missing a week.

In total, 14 schools did not return their Session Delivery Logs.

Most pupils attended 12 to 15 Mathematical Reasoning sessions, and around half attended all 12 units of the programme

Based on the Session Delivery Logs, pupils attended an average (mean) of 11.5 sessions, although this ranged from 0 to 24 sessions. The largest proportion (67%) attended between 12 and 15 sessions, as intended, and 86% of pupils attended at least 11. On average, FSM-eligible pupils attended half a session less than their peers (non-FSM average = 11.7 sessions; FSM-eligible average = 11.1 sessions).

Just over half of pupils (54%) completed all 12 units as intended. On average (mean), pupils completed 10.8 units. As with session attendance, FSM-eligible pupils completed on average half a unit less than their peers (non-FSM average = 10.9 units; FSM-eligible average = 10.3 units). Overall attendance to individual units was high—around four in five pupils. This was highest for units 2 and 3 (attended by 87% of pupils) and lowest for unit 12 (attended by 79% of pupils). Based on the case studies, this likely reflects some teachers stopping delivery before they had completed all the units (see above).

Teachers often took longer than an hour to deliver each session or skipped material to fit the unit within an hour

Programme sessions were expected to last 60 minutes, comprised of a 30-minute session for the whole class led by the teacher, followed by another 60 minutes on group work (half the class working with the teacher, the other half playing the computer games). The programme guidance indicated that if it was not possible to cover all of the content in the whole-class session, the teacher should still move onto the group work for the second half, and cover any remaining content in the next Mathematical Reasoning session.

Half of surveyed teachers (49%) who delivered at least one session reported that their average session length was over 60 minutes. A further third (35%) reported their average session length was 60 minutes. The range varied from an average of 30 minutes to over 120 minutes. More deprived schools were more likely to report delivering sessions lasting more than 60 minutes (54%, 27 out of 50) compared to less deprived schools (41%, 12 out of 29). It is not clear whether this is because less deprived schools skipped more content, and/or if teachers and/or pupils in more deprived schools found the sessions more challenging and consequently progressed through the content more slowly.

Most of the observed sessions lasted around 60 minutes with approximately half the time spent on group activities. However, there were some exceptions to this, both shorter and longer than the intended length. Moreover, many of the case study teachers did not complete the full unit in the hour during the observations—either skipping parts of the unit material or stopping before the end. These teachers explained during their interview that they did not intend to cover the missed content in a later session as they did not have the time. Other case study teachers reported that there was too much content to cover in each session, particularly when they were also required to cover multiple different concepts and types of activities. One teacher felt that this was too challenging for the pupils' focus, while another explained that this meant they felt they did not have the time to unpick some of the concepts as they were just 'rushing' to get through all the content. This contrasts with the delivery team's explanation to TLs that the purpose of the sessions is to give children 'space to think'.

These findings align with results from the survey where 16% of teachers reported skipping questions, examples, and/or units in their programme delivery.

Based on feedback in the webinars, the first few sessions were seen to be the longest and most challenging to deliver, with some teachers reporting that covering the content in 60 minutes became more manageable as they and their pupils became more familiar with the programme.

Most schools did not follow intended practice in relation to splitting the class for the group work component

Each programme session was intended to be comprised of teacher-led work as a class followed by half the class doing group work with the teacher and the other half working on the computer games. This split was intended to be determined by whether the teacher felt each pupil would benefit more from additional support or extension activities for the material covered in that session.

Nearly half of the surveyed teachers (46%) reported having delivered the teacher-led group activities as a whole class (i.e. not splitting into groups) at least once, although only 11% reported *never* having split into groups. Teachers in the interviews and webinars explained that scenarios that resulted in them not splitting the class into groups were typically because they had run out of time after the whole-group session or did not have the necessary TA support to enable group work (more information about the role of the TA, and implications of the absence of a TA on delivery is covered under 'Context and moderators' section). Where they did not split into groups, teachers often reported delivering the teacher-led group work to the whole class and providing the whole class with the opportunity to play the computer games at a different time (although at least one teacher noted that this did not always happen in practice).

Other observed variations around grouping practice included not involving some children (e.g. those with lower ability levels) in the sessions at all, splitting higher ability children into a separate third group to do the extension activities, or getting both groups to complete both activities each session, resulting in limited time on each. Some teachers (10%) reported not using the computer games at all.

Where group work did occur, only a quarter of surveyed teachers (24%) reported grouping pupils on the basis of how they responded to the whole-class component, as the programme intended. Instead, the most commonly reported approach (35%) was to put higher attaining pupils in one group and lower attaining pupils in the other. This suggests that the guidance provided in the TL training about grouping pupils based on the current session (rather than grouping based on prior information about pupils' abilities) was not followed by schools. A quarter of surveyed teachers (26%) opted for two mixed attainment groups, with some case study teachers explaining that it was not feasible for all the lower attaining pupils to be supported to access the computer games at the same time. Another case study teacher explained that one of their reasons for retaining the same groups was that it helped to create a routine for the pupils.

A few case study teachers reported grouping the pupils based on how they responded to the session. However, in two of these cases the teacher implied that the pupils who needed extra support always did the teacher-led activities, while the other group always played the computer games, rather than these cohorts switching between the two activity types, each session.

Observed sessions demonstrated scaffolding and encouragement of pupil discussion

Case study data collection included observations of session delivery, with the aim of gaining insight into the extent to which teachers delivered the material in the manner recommended by the programme training and self-evaluation tool.

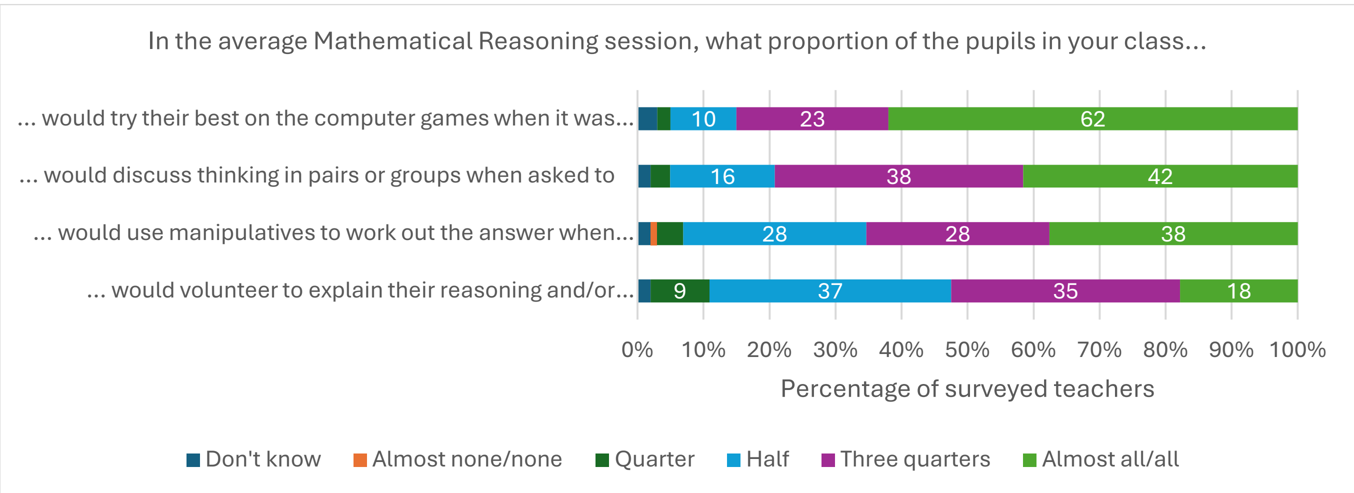
Some of the observed sessions displayed active pupil participation, with the teacher asking questions and prompting pupils to explain and discuss the reasoning behind their answers. In other cases, however, the teacher was observed to focus more on the accuracy of pupils' answers than on their reasoning. In most cases, the discussion was facilitated by the teacher and/or the pupils were allowed to talk to each other while they worked.

Most of the observed sessions demonstrated verbal scaffolding by the teacher. For example, the teacher would encourage their pupils to first try to explain their reasoning before they provided any support, and in some instances the teacher would repeat a pupil's explanation in clearer terms for the rest of the class. In at least one case, however, the level of scaffolding dropped off over the course of the session as the teacher started to move through the content more swiftly. Teachers also used the language from the programme to verbalise reasoning in a variety of ways.

The extent of pupil participation in elements of the teacher-led work differed across classrooms

Most of the surveyed teachers (80%) reported that at least three-quarters of pupils would discuss their thinking in pairs or groups when they were asked to as part of the programme sessions, with 42% reporting this to be the case for all or almost all pupils (Figure 11). However, most teachers (72%) reported that only between half and three-quarters of pupils would volunteer to explain their reasoning and answer questions.

Figure 11: Teacher perspectives on the proportion of pupils who engaged with various elements of the programme



Source: Question M (In the average Mathematical Reasoning session, what proportion of the pupils in your class...) from the Endpoint Teacher and TA Survey (N=101 teachers).

Case study staff also reported that most pupils had participated in the collaborative elements of the teacher-led work, and this was also observed during the case study visits. NFER researchers observed pupils listening to the scenarios which teachers read out, participating in whole-class counting activities and discussions, answering questions (both verbally and in workbooks), discussing with their peers and using the manipulatives (if available). Several staff (teachers and TAs) reported that, overall, pupils had enjoyed the sessions and had been excited for them each week. Staff felt this had been driven by the programme offering pupils something new and different to their usual maths teaching practice. Staff in most case study schools said that pupils had responded particularly well to the workbooks, which they felt was due to the coloured visuals and the activities/games.

Pupils involved in the focus groups shared their own positive views on the teacher-led work. Most of the pupils indicated that they liked the whole-class activities. Pupils reported particularly enjoying the games involving manipulatives (e.g. cards, money, counters), such as the caterpillar game, which was played as a whole class, and the shop game on the back page of the workbook. Most pupils also indicated that they enjoyed the smaller group work with their class teacher. The pupils liked the lower pupil–teacher ratios when the class was split and the additional attention and personalised support that they consequently received.

A few staff members perceived lower levels of enjoyment and participation in the teacher-led work among some pupils, which they attributed to the session content being too advanced for these pupils and/or the sessions being too long. A small number of pupils involved in the focus groups reported mixed feelings relating to the whole-class activities, and mixed or negative feelings relating to the group work with their teacher. These pupils mentioned getting frustrated because they found the questions too challenging, and that some of the activities could be more fun and practical.

Teachers were not always consistent in their use of manipulatives to support pupils' mathematical reasoning

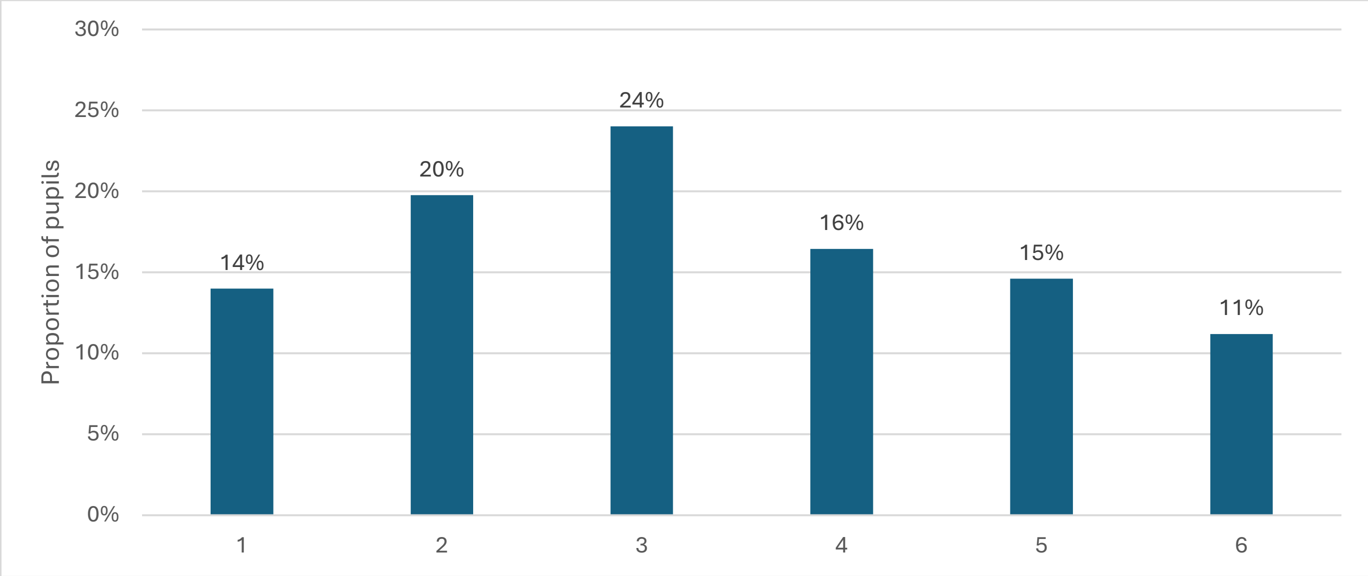
Use of manipulatives by both teachers and pupils was emphasised by the delivery team as an essential aspect of the programme. However, only around 38% of teachers responding to the survey indicated that almost all or all pupils in the class would use manipulatives to work out the answer when asked to in the average programme session. This was supported by the observations of programme sessions in the case studies where for example in some cases only pupils or only the teacher used manipulatives, or they were only used for some activities. No manipulatives were used in one of the observed sessions. Manipulatives were not used by pupils during any of the computer games sessions observed. Despite this, the value of using the manipulatives was raised by several teachers and pupils during the case study visits and webinars.

Few pupils took advantage of the number and variety of computer games that were available

The Mathematical Reasoning programme included 60 computer games, across six subject topics. Pupils were free to select the games they played, and they could play each game as many times as they wished. On average (median), pupils played a total of 21 computer games, including repeats of the same game. On average, FSM-eligible pupils played fewer games than their peers (19 vs 23 games, respectively). Similarly, pupils in more deprived schools played, on average, fewer games than pupils in less deprived schools (17 vs 24 games, respectively). These differences did not appear to be a consequence of playing the computer games outside of school hours. A quarter (23%) of teachers said they allowed pupils to play the computer games at home or at other times of the school day. Just 3% of all pupils played the games outside of school, a proportion comparable across FSM-eligible pupils and their peers. On average (median), these pupils only played four games outside of school hours, although this ranged from 1 to 59 games. As there was a higher proportion of FSM-eligible pupils in more deprived schools, structural challenges, such as the availability of TA support (see 'Context and moderators' section below), may have driven the difference in number of games played.

Of the pupils who played at least one game (N=2,924), just 11% played at least one game from each of the six topic areas (see Figure 12). The largest proportion (24%) played games from just three of the topic areas.

Figure 12: Number of topic areas covered by pupils on the computer games



Source: Computer games data analysis (N=2,924 pupils). Only includes pupils who are recorded to have played at least one computer game.

No pupil played more than three-quarters (45 out of 60) of the available games, and only 6% played over half. A significant minority (40%) played ten different games or fewer. The average (median) number of different games played by pupils who played at least one game was 12.5. This suggests that pupils were often repeating the same games.

Most teachers (62%) reported that all or almost all pupils would try their best on the computer games when it was their turn, and a further quarter (23%) said this was the case for three-quarters (also see Figure 11 earlier in this report).

Maths leads, teachers, and TAs in most case study schools reported that pupils had particularly enjoyed the opportunity to play the computer games on laptops or tablets; most pupils involved in the focus groups were positive about the computer games too. Pupils reported appreciating the variety of games that were available and the range of levels they could work through. They also liked the reward/bonus games that they could play at the end of the sessions. School staff perceived the computer games to be beneficial for giving pupils the chance to practice and develop their maths skills, without fully realising they were learning maths.

Pupils and staff did, however, raise several frustrations and limitations in relation to the computer games, many of which were also observed during case study visits. Pupils did not always read the questions or instructions correctly, limiting their ability to understand the aim of the games and answer correctly. One TA related this to reading ability levels. Researchers observed pupils guessing the answers to games, and pupils themselves told researchers they had to 'guess the answers', indicating a lack of understanding of the requirements of the games to apply their reasoning skills and knowledge from the teacher-led work. This may at least in part account for the quarter of games played (26%) that recorded a score of 50% or less.

School staff and pupils in a few schools reported that the requirement of achieving 100% to access the bonus games resulted in pupils losing interest in the computer games over the course of the programme. This was a particular issue for pupils with lower maths attainment and SEND. One in 20 children who played at least one game never scored 100%.

> *We just found it hard to engage some of them in the computer games after the first couple of weeks when they weren't able to get the bonus games—almost like they didn't want to bother and would ask to do the group work instead, because they could see that the group were doing a game that they were getting positive feedback for straight away. Whereas on the computer games those children were just always getting a 'try again'.* (Teacher, CS4)

Staff suggested that the computer games needed to be differentiated for different ability levels, for example, by enabling a lower threshold for the bonus games for lower attaining pupils so that they could still get a sense of achievement, and/or by providing more visual prompts to support accessibility for SEND pupils and those with lower reading abilities.

Other challenges with the computer games included 'glitches', which were reported to be frustrating for the pupils. For example, pupils and teachers in two schools reported that the games did not always accept the correct answers. Some pupils also faced challenges with the fine motor skills required for playing the games on a tablet, and with navigating the computers. In two case study schools, pupils reported a preference for the teacher-led small group work, which included practical games, over the computer games. Staff in a few other schools reported that pupils had been given the opportunity to play the games in school time, outside of Mathematical Reasoning sessions, and some had taken advantage of this.

Some teachers made further minor adaptations to the programme material

A small number of surveyed teachers (5%) reported started working with larger numbers than those specified in the programme materials, and one teacher reported having made 'tweaks' to the sessions to bring in some elements from their usual curriculum to make sure they were covered.

One case study teacher reported having made adaptations to the narratives around the questions to make them more engaging and accessible for the pupils. While some teachers in the webinars reported that the children appreciated receiving the certificates for the computer games, a few case study teachers reported using the certificates for the computer games differently to instructions. In some cases, teachers did not use the certificates at all because they felt they were not necessary in addition to the bonus games, or because they did not align with the school's rewards system; others wanted to provide a reward to those children who did not score 100% and so gave out additional certificates that they created themselves.

TAs played a key role in supporting specific pupils and enabling group work, but many schools were not always able to have one present

Most of the surveyed teachers (78%) reported having a TA present for at least 'most' of the programme sessions, but only 38% reported having one for all of them. Nearly one in five teachers (18%) had a TA for just a few or none of the sessions. Teachers at more deprived schools were much less likely to report having a TA for most or all of the sessions (69%, 36 out of 52) compared to those in less deprived schools (87%, 27 out of 31). Based on the Session Delivery Log data, TAs were present for an average (mean) of ten sessions (out of the intended 12 to 15), and for 9.5 sessions in more deprived schools. The case study data showed that during the observed whole-class session component, TAs usually supported specific pupils who were seen as requiring additional support, for example, by providing extra input, keeping them on task, supporting behaviour, and/or helping with a language barrier. In some case study schools, teachers were observed helping children from a range of ability levels.

The case studies also showed that during the group work, TAs primarily supported the children on the computer games (as intended). A large part of this was supporting with the technology, including logging the pupils in to their games account (see 'Context and moderators' section below). In addition, the TAs provided additional support, reading out questions and supporting with language if needed, and monitoring behaviour. One case study TA explained that their role involved stopping the pupils from just guessing the answers to the computer games and instead they encouraged them to try and work them out.

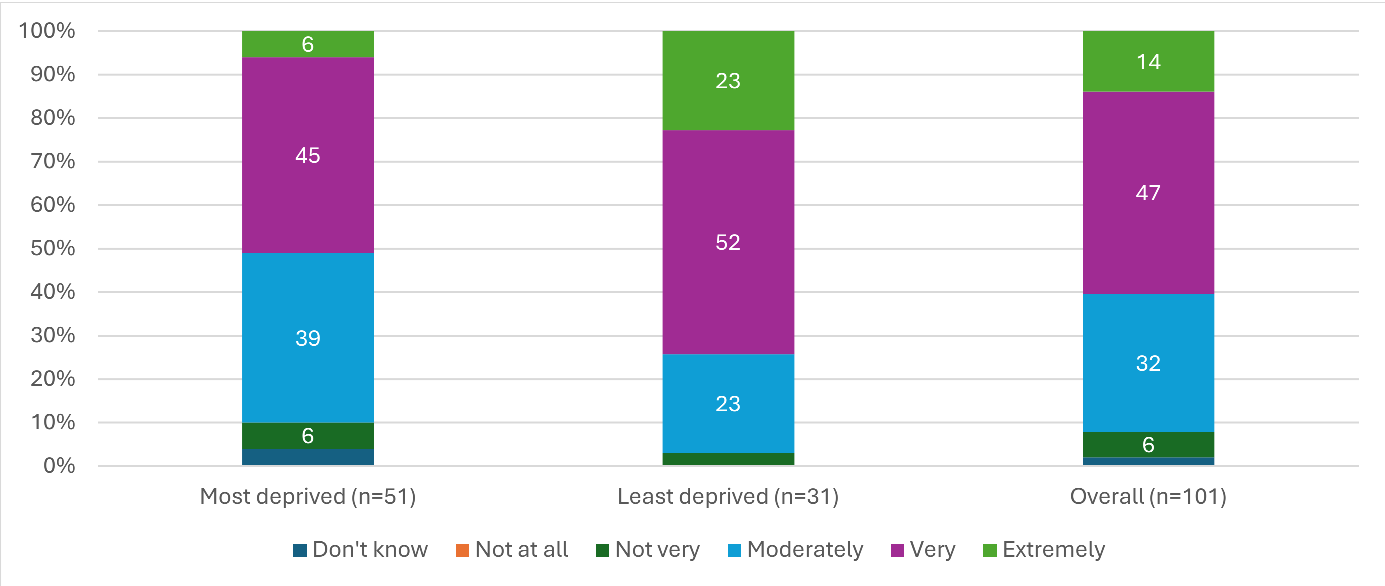
One school mentioned that the teacher's wider responsibilities sometimes resulted in the TA delivering the session. A small proportion of the surveyed TAs (9 out of 61) reported that they would not have usually been supporting the class in the time allocated for the programme sessions. This aligns with the findings from the Baseline Survey, where 9% of the teachers who responded reported never having a TA supporting their Year 2 class during maths lessons. Two case study teachers reported that they would not usually have TA support for maths lessons, but the TA who they were allocated had their time protected specifically so they could support delivery of the Mathematical Reasoning programme. While only occurring in a small proportion of cases, this additional support for pupils during maths lessons can be seen as a positive unintended consequence of the programme for the pupils involved but may also mean that pupils in another class were missing out on usual support as a result.

Case study teachers and TLs emphasised the importance of having the support of TAs for effective programme delivery, particularly for the group work, and not having TA support was reported to be a significant challenge (see 'Context and moderators' section below for more information).

Most intervention teachers perceived the Mathematical Reasoning programme to be very different to their usual practice particularly in terms of the structure and the explicit dedicated focus on reasoning

To give some context to the findings described above, most (61%) of the surveyed intervention teachers reported that the Mathematical Reasoning programme was 'very' (47%) or 'extremely' (14%) different to their usual maths lessons. However, teachers in less deprived schools were more likely to report this (52% and 23%, respectively) than those in more deprived schools (45% and 6%, respectively) (Figure 13).

Figure 13: Teacher perceptions of how different programme is to usual maths lessons



Source: Question O.5 from the Endpoint Teacher and TA Survey: 'How different were the Mathematical Reasoning programme sessions from your usual maths lessons?' (most deprived n=51, least deprived n=31, overall n=101).

Several case study teachers reported that their usual maths lessons tended to focus on fluency (including counting, calculations using the four operations, and applied fluency tasks) and included fewer opportunities for pupils to engage with reasoning and problem-solving activities compared to the programme sessions. In some schools, prior to the trial, mathematical reasoning had been an add-on activity at the end of the lesson, rather than an integrated part of regular lessons or the regular focus of a whole lesson, as in the Mathematical Reasoning programme. Teachers reported that participating in Mathematical Reasoning had differed from their usual practice by creating dedicated time to focus on reasoning and problem-solving concepts and providing direct, explicit approaches for teaching pupils these skills.

Staff and pupils also reported that it was unusual for the teacher to work with just half the class at the time. Some teachers also reported the use of computer games as different to usual practice, which tended to provide pupils with activities via the board or worksheets. The extent to which use of manipulatives constituted part of usual practice varied substantially between schools.

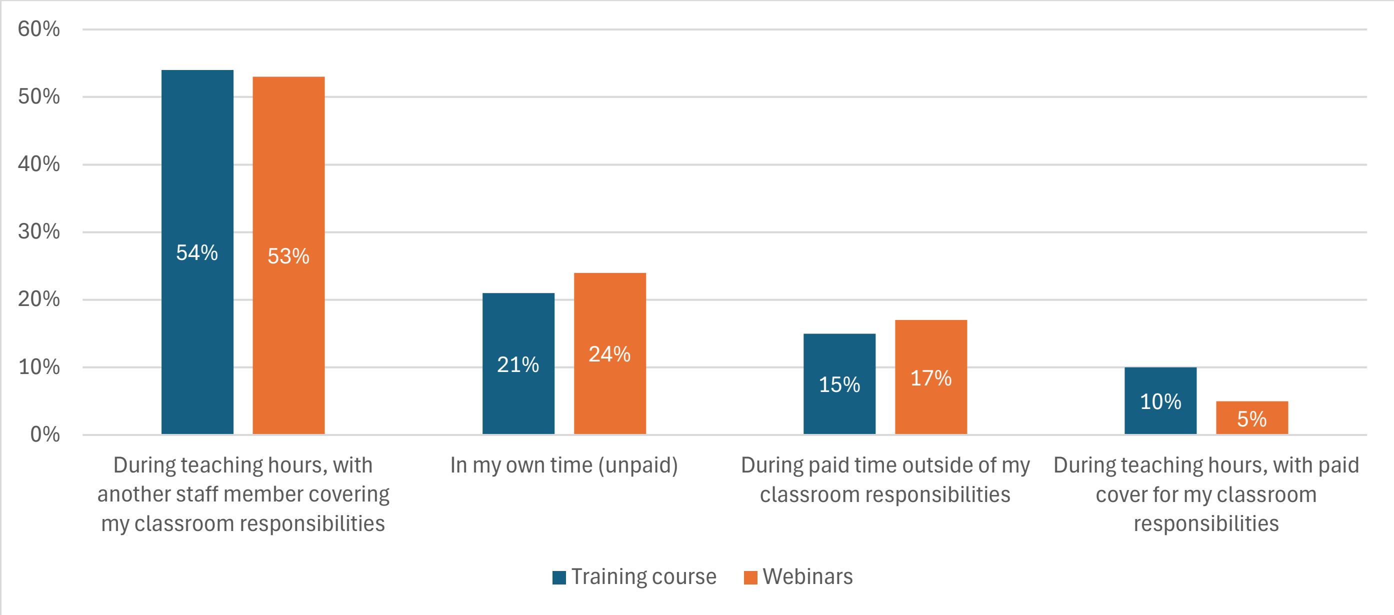
Context and moderators

IPE RQ2: What are the key moderators and contextual factors that influenced how effectively the programme was delivered?

Finding the time and cover for teachers and TAs to participate in the training and webinars was a key challenge

Half of the surveyed teachers reported completing the training courses and webinars during teaching hours with another staff member covering their teaching responsibilities (54% and 53%, respectively; see Figure 14). Just over one in five reported completing them, unpaid, in their own time (training course: 21%; webinars: 24%).

Figure 14: When teachers completed the training course and webinars



Source: Question C in the Endpoint Teacher and TA Survey: 'For the most part, when did you (1) do the online training course (Canvas), (2) attend the webinars with your Teacher Leader'? (N=101, including one no response).

Teachers in more deprived schools were more likely to complete the training course and webinars outside of their usual teaching hours (36%, 18 out of 51) compared to teachers in less deprived schools (26%, 8 out of 31).

Most of the surveyed TAs (56%, 31 out of 55) who completed at least some of the training course did so during teaching hours with another staff member covering their classroom responsibilities. A further quarter (24%, 13 out of 55) did so during paid time outside of their classroom responsibilities. Only five TAs reported doing the training course in their own time. Similar results were reported for the webinars.

Evidence from the case study teachers indicated that it was very challenging to get the cover they needed to complete the training course during class time, as one teacher explained:

> *I was doing them [the training modules] over my Christmas holidays and that's not what I should have been doing. [...] That's just how it ended up with us because we can't be released. We can't have two teachers out of class. That's just not a thing we can manage in our school.* (Teacher, CS6)

Another case study teacher reported the timings of the webinars to be particularly challenging for paid cover as this would have to cover the full afternoon, not just the duration of the webinars (typically 60 minutes to 90 minutes).

Some of the TLs noted that it was challenging for the TAs to attend the webinars as they were often covering for the teacher to attend and did not have paid time outside of class hours during which they could attend as the teachers did. One TL noted that the low levels of TA attendance at the webinars was a significant change from when this training model was piloted.

Many teachers and one TL emphasised the importance of having a supportive SLT that was willing and able to provide the necessary cover to allow the staff to fully engage with the programme, whether this be through paid cover or support staff. One teacher also highlighted the importance of a dedicated quiet space for completing the training course.

Constraints around TA availability created significant challenges for session delivery

Under IPE RQ1 it was highlighted that many teachers were not always able to have a TA supporting the programme sessions. Case study teachers reported that this was primarily the result of wider school staffing demands and constraints. Many of the case study teachers emphasised the challenges of delivering the sessions without a TA, primarily in relation to providing additional support and/or facilitating extension opportunities for specific pupils, and the feasibility of splitting into groups for the second half of the session. This latter was seen to be particularly challenging in classes with a high level of need, as one teacher explained:

> *The reason I split them into the two different groups is more behavioural than maths ability, which, if I had a teaching assistant, I would do it the other way round, but because there's only me as the adult, I have to make sure that the learning that's happening rather than them sitting there talking to each other.* (Teacher, CS8)

Several TLs also noted staffing challenges in relation to teacher availability, primarily due to sickness.

Most schools had the technology infrastructure to support programme delivery, however many pupils needed support to log on to the computer games

The large majority of surveyed teachers (89%) reported having already had enough devices for the pupils to play the computer games. However, there were a few cases where issues with the technology was observed, including laptops running out of charge during the session, poor Wi-Fi™ quality and slow laptop speed. For example, one case study teacher reported a good IT set-up in the school as an important facilitator, while another felt poor quality laptops had been a key challenge.

Over three-quarters of teachers (78%) reported that it was at least 'moderately' easy for pupils to access the technology to place the computer games. However, one in five surveyed teachers (21%) reported that it was a challenge for pupils to access the games. Several case study teachers mentioned that the pupils found it challenging to get started on the games each session, with substantial support from the TA required to get them logged in—particularly when the level of need in the class required the TA to log each pupil in individually. In some cases, pupils were observed to spend longer getting logged in to the platform than they spent playing the games. Where no TA was available or the TA was not comfortable with the technology, the teacher had to choose between taking time away from the teacher-led group to support the computer games group or not splitting the groups. Some teachers mentioned in the webinars that they had developed quick response (QR) codes™ or laminated log-in cards to support with this process.

Preparation requirements were seen to be burdensome, particularly in relation to manipulatives

More than half of surveyed teachers (57%) said it took them longer to prepare for each Mathematical Reasoning session than for a usual maths lesson, with 13% saying it took 'much' longer. Only 15% said that it took less time than usual.

Several case study teachers reported the most time-consuming element was preparing a large number of manipulatives for each pupil for some of the sessions. This was reported to be a particular issue when the teacher had no TA support. Some teachers also reported spending additional time checking they were happy with the content and flow of the slides and making adaptations to the session materials to better meet pupils' needs, as well as returning to the training as a reminder for how to play the in-class games. Some case study teachers reported that it took some time to 'get your head around' the programme at the start, but that preparation time reduced as they became more familiar with the programme. A few TLs also reported some teachers had fed back that the handbook was 'a lot to digest' and noted that the teachers were required to process a lot of information per session, given it only represented 60 minutes of lesson time per week.

On the other hand, several teachers reported in the webinars that the pre-prepared resources had reduced the amount of time they spent on preparation compared to a usual maths lesson.

Characteristics of the pupil cohort were perceived to be a moderator for effective implementation

Several case study teachers observed that the programme was very challenging to deliver when there was a large range of maths ability levels within the class or a 'high level of need' in the class. Some teachers reported finding it difficult to sufficiently adapt the programme to meet the needs of lower performing children who needed to go at a much slower pace; these teachers would have liked additional guidance around this.

One teacher also reported finding it difficult to provide higher attaining pupils with sufficient level of challenge when they had to focus on helping other pupils to access the programme. Use of the extension sheets was rarely observed during the case study visits, and in one case they caused confusion as the pupils were not able to work out what to do without teacher support.

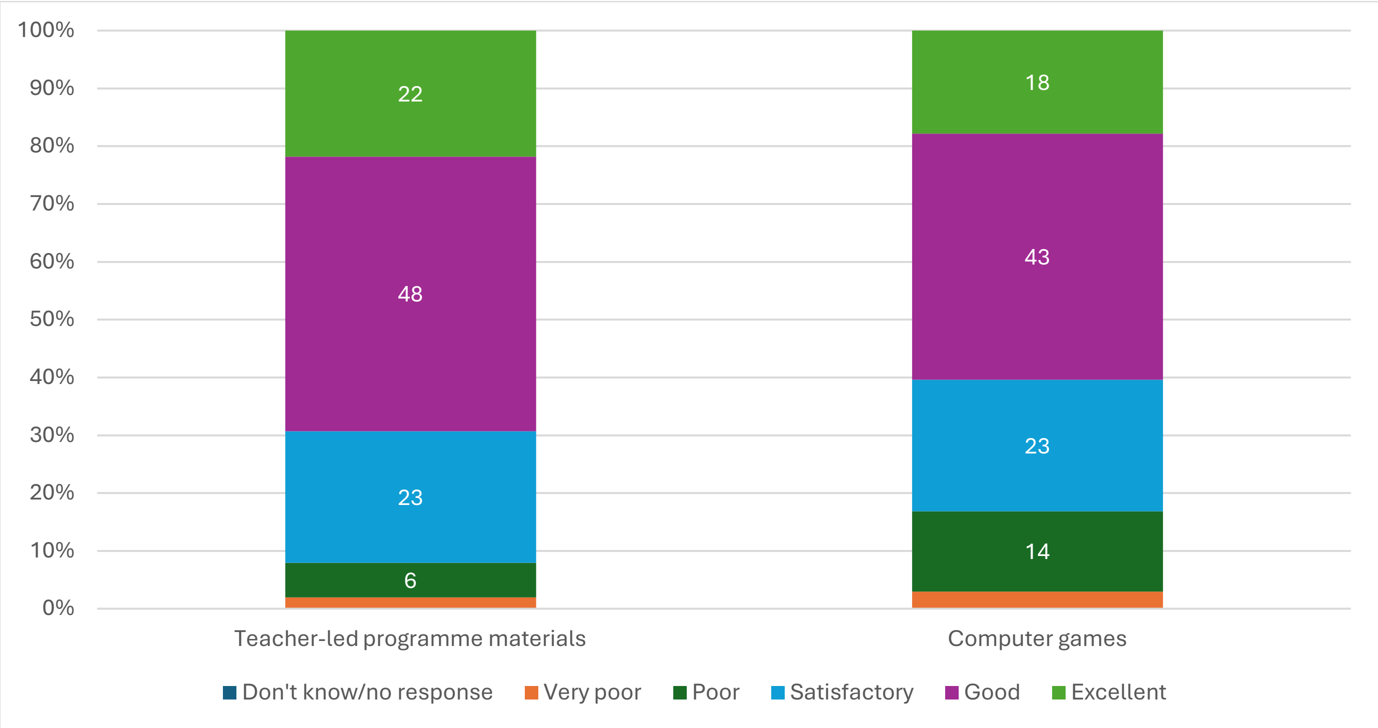
Some case study schools also mentioned pupil behaviour as a potential challenge for delivery as the use of games and manipulatives could make the children 'very excitable', and noise from the pupils on the computer games could be disruptive to the teacher-led group work.

Most teachers judged the programme materials to be of high quality, although views on the computer games were more mixed

Teachers were provided with materials for the teacher-led components, including a handbook, digital slides, and Pupil Workbooks. Most of the surveyed teachers (70%) considered the quality of these materials to be 'good' (48%) or 'excellent' (22%) (Figure 15), while a further quarter (23%) considered them 'satisfactory'.

Similar results were seen for the computer games—although a slightly lower proportion considered the quality of them to have been 'good' or 'excellent' (43% and 18%, respectively) compared to the other Mathematical Reasoning materials.

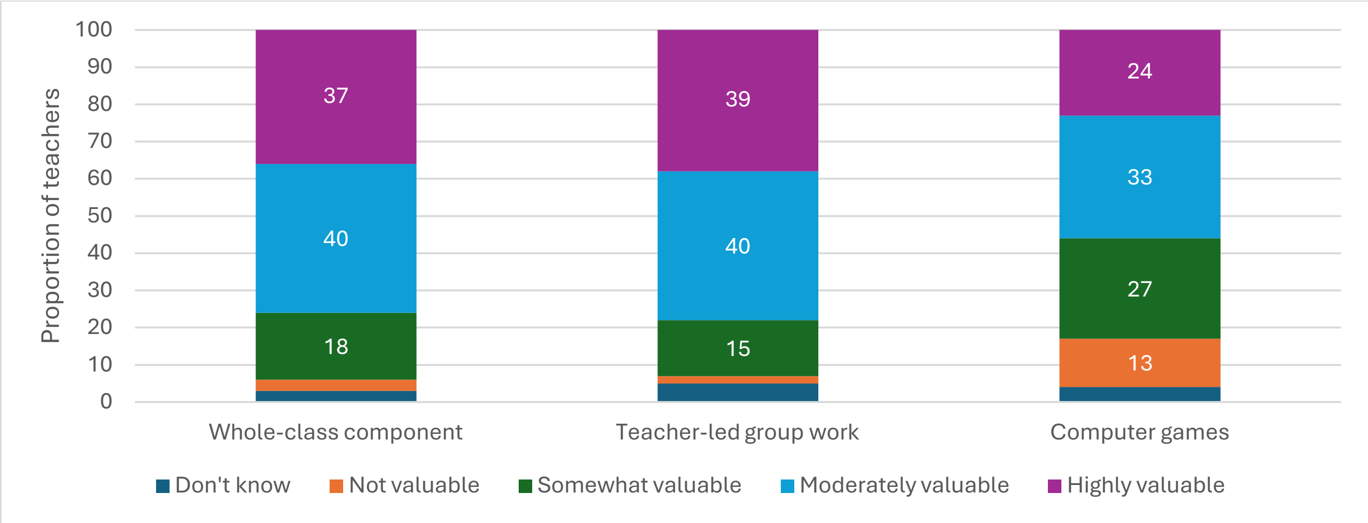
Figure 15: Teacher perceptions of the quality of teacher-led programme materials and the computer games



Source: Question F ('How would you describe the quality of...(3) the materials provided to support delivery of the programme (e.g. the teaching handbook, lesson presentations, pupil workbooks) and (4) the computer games') from the Endpoint Teacher and TA Survey (N=102 teachers).

Similarly, while most surveyed teachers (76–78%) reported that the teacher-led components of the programme were at least 'moderately' valuable for pupil learning (Figure 16) fewer—just over half (57%)—considered this to be the case for the computer games.

Figure 16: Teacher perception of the values of different programme components for pupil learning



Source: Question R from the Endpoint Teacher and TA Survey: 'Based on your experiences of delivering the programme, how would you rate the following components of the Mathematical Reasoning programme in relation to your pupils' learning? (1) whole-class component, (2) teacher-led group work, (3) computer games (N=101).

Case study schools described the resources as high quality, engaging, and accessible. Several teachers also emphasised the value of having all the resources provided and in one place (on the website) as a facilitator for preparation and delivery. Most case study teachers were positive about the quality of the digital slides, describing them as easy to follow, and the handbook was seen to be helpful and self-explanatory. One case study teacher also mentioned that they found the structured nature of the sessions helpful for creating routine and shared expectations for each session.

However, one case study teacher described the session approach as an '*old way of doing things*' (Teacher, CS9), while another noted that they had tried to move away from using worksheets in teaching. They also found juggling the slides and handbook at the same time during the whole-class session a challenge.

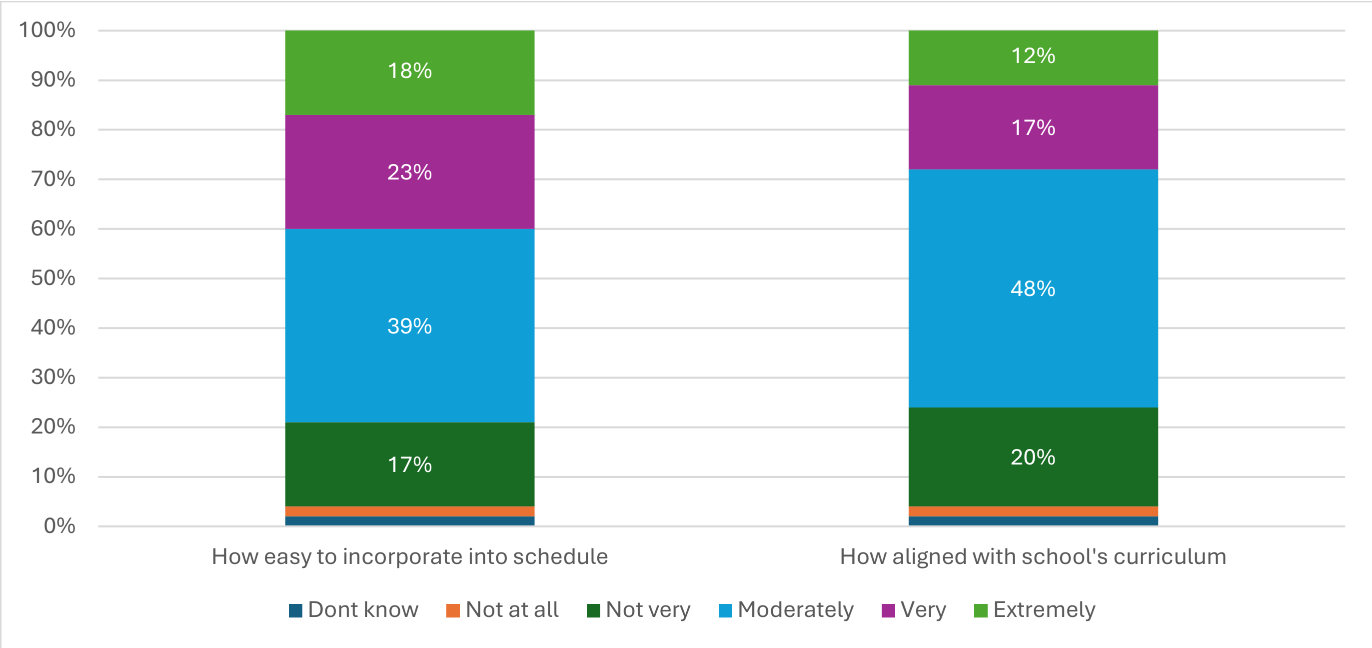
Some case study schools mentioned the importance of particular teacher qualities for effective session delivery, including experience, enthusiasm, higher maths ability, or a strong understanding of the year group. Both teachers and TLs observed that the programme would have been 'too much' for an Early Career Teacher to take on due to the amount of content and materials to manage.

Some case study teachers described the computer games as 'fun' and 'simple', providing a good opportunity for independent work to consolidate learning while allowing the teacher to work with a smaller group. However, others felt the games were not accessible for all pupils and that it could be demoralising for those unable to achieve the 100% required for a bonus game or certificate. One in 20 children who played at least one game never scored 100%, and this proportion was slightly higher for FSM-eligible pupils compared to their peers (5% vs 3%).

Most schools found that the sessions were broadly aligned with their curriculum and that they were able to fit the sessions into their schedule

Most schools (77%) reported that the Mathematical Reasoning programme was at least 'moderately' aligned with their school's maths curriculum (see Figure 17). One in five (22%) said it aligned 'not very' or 'not at all' well. Four out of five respondents reported that it was at least 'moderately' easy to incorporate Mathematical Reasoning sessions into the class schedule (80%).

Figure 17: Teacher perspectives on how aligned the programme was with their curriculum and how easy it was to incorporate into their class schedule



Source: Question O.2 ('How easy was it to incorporate the Mathematical Reasoning programme sessions into your class schedule?') and O.3 ('How aligned was the Mathematical Reasoning programme with your school's maths curriculum?') in the Endpoint Teacher and TA Survey (N=101).

Case study schools generally reported that the programme had aligned well with their usual curriculum and existing maths schemes, although several mentioned finding it a challenge to fit the sessions into their class schedule given the many competing priorities.

A few teachers reported that delivering the programme had caused them to 'fall behind' their usual curriculum delivery timeline and they felt unable to dedicate as much time to some maths topics as they would have liked.

> *That is the only negative, that we haven't been able to get through [the rest of] our maths lessons as quickly. But to be fair [...] they were much more able to recognise coins straight away because we had done the basics in the Mathematical Reasoning lessons, so we have covered some of it, but it did mean we fell behind with the other lessons.* (Teacher, CS3)

One teacher noted that always delivering the sessions at the same time of the week created a routine that was helpful for the pupils. Another teacher mentioned that they deliberately scheduled the sessions for the morning when they knew the pupils would be best able to focus.

Staffing requirements and competing school priorities may be key barriers to effective scale-up

Only around a third of surveyed teachers (35%) reported that they would choose to continue delivering the programme in the next academic year. Just over a quarter (28%) said they would not continue to deliver the programme, if it was up to them, while 38% were not sure if they would choose to continue to deliver the programme next year or not.

The most common reasons for not wanting to continue with the programme were not having the necessary staffing capacity and the programme not fitting with the school's maths priorities (39%, 11 out of 28 for each reason). This concern about staffing may reflect the previously reported challenges around enough TA support time.

Case study schools felt that in order to support future delivery it would be important for the programme to be better integrated into a school's maths strategy and curriculum, rather than as separate, stand-alone sessions.

Eight teachers also used the free-text entry option in the survey to indicate that the time demands of the programme were a barrier for continuing the programme, both in terms of the length of the sessions and competing school priorities. These reflect the concerns reported above, around there being too much content to deliver for each session, as well as the challenge some schools reported fitting the sessions in alongside their standard maths curriculum. Five teachers used the same survey function to report that the programme had not been suitable and/or sufficiently engaging for the pupils.

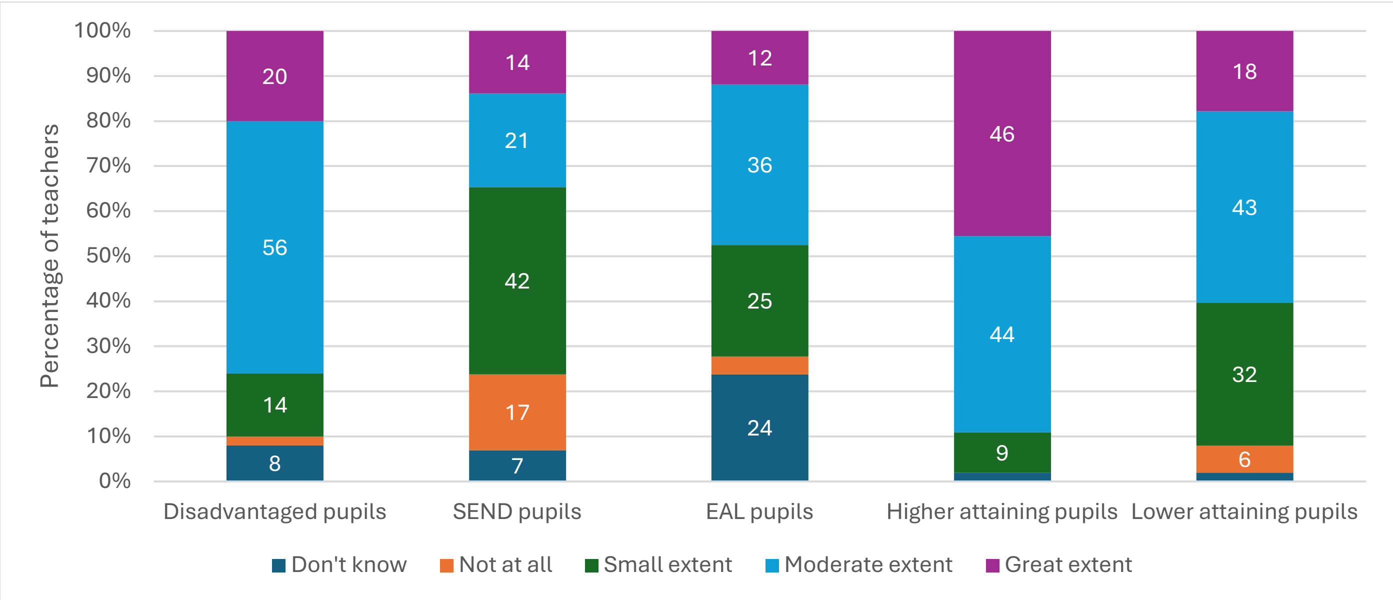
Although equipment requirements were only cited by one teacher as a specific barrier to continuing with the programme, evidence from elsewhere suggested that this would be an ongoing consideration for participating schools. For example, 13% of surveyed teachers in control schools reported not having had enough personal devices in the 2024/2025 academic year for half of their Year 2 class to be using one at the same time, and a further 17% said that while these devices are available in theory, it is very hard to get hold of them. This suggests that access to the necessary technology could still be a challenge for scale-up more widely.

IPE RQ3: *How did pupils' experiences of the programme vary depending on pupil characteristics?*

Most teachers felt that the programme had met the needs of disadvantaged pupils

Three-quarters of teachers felt that the programme met the needs of disadvantaged pupils (i.e. those eligible for FSM) to a 'moderate extent' or 'great extent' (56% and 20%, respectively; Figure 18). Analysis by FSM quintile did, however, reveal some differences within these proportions: teachers in less deprived schools were more positive about the suitability of the programme for disadvantaged pupils than teachers in more deprived schools.

Figure 18: Teachers' perceptions of the extent to which the Mathematical Reasoning programme met the needs of different pupil subgroups



Source: QN (To what extent did you find the Mathematical Reasoning programme met the needs of the following groups of pupils?) from the Endpoint Teacher and TA Survey (N=101).

Staff in case study schools perceived the experience of disadvantaged pupils to depend on their ability, rather than their eligibility for FSM alone. For example, they reported that while disadvantaged pupils who were higher attaining could access and benefit from the Mathematical Reasoning programme, disadvantaged pupils who were lower attaining were more likely to face challenges. Staff in one school did however, report a knock-on effect of the impact disadvantage had on pupils' vocabulary, which made it challenging for pupils to access the question wording and verbalise their reasoning.

Only a minority of teachers felt that the programme fully met the needs of pupils with SEND

Nearly one in five (17%) of surveyed teachers reported that the programme did not meet the needs of pupils with SEND 'at all', while a further 42% said the programme did so to a 'small extent' (see Figure 18). Only 14% of surveyed teachers reported that the programme met the needs of pupils with SEND to a 'great extent', and 21% that it did so to a 'moderate extent'.

Several case study interviewees reported that the sessions were pitched too high and too 'pacy' for pupils with SEND, leading to challenges around access and participation for these pupils. Teachers reported that to be suitable for pupils with SEND, the programme required additional scaffolding and differentiation. A small number of teachers made adaptations to their delivery for these pupils, such as using more resources, working with small numbers, and dedicating TA support to these pupils for the content to be explained in greater detail and delivered at a more appropriate pace. In one case study school, pupils with SEND did not participate in the Mathematical Reasoning programme at all, with the research team observing pupils being given alternative activities.

Those case study teachers who reported that the programme had worked well and benefited pupils with SEND attributed this to the use of resources and manipulatives, which were seen to support pupils with SEND to access reasoning concepts.

Most teachers reported that the programme met the needs of EAL pupils to a 'small' or 'moderate' extent

Surveyed teachers were most likely to report that the Mathematical Reasoning programme had met the needs of EAL pupils to a 'moderate extent' (36%) or 'small extent' (25%) (Figure 18). Similar to pupils with SEND, case study teachers reported elements of the Mathematical Reasoning programme such as the pictorial representations and manipulatives had supported pupils with EAL to participate in the sessions. However, the central role of language in the programme meant that challenges arose for children with very limited English language abilities in terms of accessing and understanding the reasoning questions, participating in discussions, and verbalising their reasoning. In one school where particular challenges were experienced, pupils with EAL were given another activity to complete with the TA to support the development of their basic number skills.

Almost a quarter of teachers were unsure regarding the extent to which the programme met the needs of this group of pupils (24%), which this may relate to whether or not there were children with EAL in their class, as well as the high level of variation in the English language abilities of pupils with EAL.

Teachers felt that the programme was best suited to higher attaining learners

Teachers were more positive about the suitability of the programme for higher attaining learners than for lower attaining ones. Nearly half (46%) of surveyed teachers indicated that the programme had met the needs of higher attaining pupils to a 'great extent', compared to 18% of teachers for lower attaining pupils (Figure 18). A comparable proportion indicated that the programme had met the needs of higher (44%) and lower (43%) attaining pupils to a 'moderate extent'.

Staff in case study schools perceived the Mathematical Reasoning programme to be more accessible and beneficial for higher and middle ability pupils with stronger fluency and number facts knowledge. They felt that these pupils could dedicate more cognitive capacity to the demands of the reasoning questions and apply their prior knowledge to answering questions and explaining their reasoning. Teachers reported that higher ability pupils were able to work through the activities more quickly and independently. The extension activities also presented a good opportunity to 'push' these pupils, although teachers did note that it could be logistically challenging to support pupils to access these during the whole-class work.

> *The children that have got the basics and the foundations of the fluency, they have the ability because they're quite secure in their number facts…The children who are quite higher attaining children already, I think that benefited them the most because they could really think deeper about what the questions were asking them and they could choose which operation it was asking, and they had a [sic] quite a really good understanding.* (Teacher, CS2)

A few staff reported that higher ability pupils were able to 'just see' or 'just knew' the answers to questions, in some cases without needing to work through their reasoning or use the manipulatives or resources. Similar comments had also been noted by a TL, who said that this scenario constituted a lack of understanding about the Mathematical Reasoning programme from teachers regarding the importance of the reasoning elements, rather than pupils just reaching the correct answers. A few teachers did however, raise that, higher and middle ability pupils had faced unexpected challenges with elements of the programme, including with money and with computer games which they felt was related to pupils' rushing the activities. Despite this, a couple of teachers felt that, overall, with increased access and ability came increased enjoyment of the Mathematical Reasoning programme.

The majority of case study staff also found that pupils with lower prior attainment required additional support to access the programme. They noted the importance of having a TA present to provide these pupils with the necessary support, but as discussed in the 'Context and moderators' section of this report, TAs were not always available to support session delivery. Teachers and TAs reported that they had made ability-based adaptations, such as delivering the content at a lower level (e.g. working with smaller numbers, or with a smaller number of coins), in more manageable steps, with more manipulatives and at a slower pace to support pupils to participate in the programme.

> *We sometimes have to slow the pace right down, get the objects in front of them, break it down further but they are still accessing it and last week, once they'd done the further group work with me, their confidence was growing because they were starting to get it so they are achieving in their own way. We scaffold but we try and get them all to the same point, it's just how you get them there that changes.* (Teacher, CS1)

Despite these efforts, several teachers reported that pupils who did not have foundational maths and arithmetic skills, or the reading and comprehension skills required to access and understand the demands of questions (both in workbooks and computer games), struggled to access the programme. The focus on money was perceived to be a particular challenge—for example, some teachers reported that some pupils struggled with 20 and 50 pence because they did not have strong number skills beyond ten. Some teachers felt that, while lower attaining pupils could access some of the content when working in a group, or when paired with higher attaining pupils, overall the programme was not suitable for them to engage with independently.

> *The children who were lower attaining, I think, to be honest, they probably didn't get a lot out of it because a lot of it was a bit overwhelming for them…We read the questions to them which was great but they don't have the comprehension skills to even understand if it was adding or subtracting, which operation to use.* (Teacher, CS2)

Many of the case study staff felt that the Mathematical Reasoning programme needed more options for differentiation, so that activities could be better aligned with the level at which the pupils were working. Teachers suggested that a different workbook and different games (both practical and computer games) pitched at a lower level would be beneficial so that pupils with lower prior attainment could better access the programme and experience the same successes and enjoyment in the programme as their higher ability peers.

However, several teachers also highlighted elements of the Mathematical Reasoning programme which were helpful for pupils with lower prior attainment. For example, the visual/pictorial and practical elements of the programme, such as drawing out objects in word problems, or using manipulatives to represent elements of the problems, were seen to have made the content more accessible for pupils and subsequently enabled them to participate in the activities.

In one case study school, some pupils did not receive the programme at all and instead spent the time working on more foundational maths content to secure their basic number skills. A class teacher in another school reported that if they were to deliver Mathematical Reasoning again, it would be more beneficial for this group of pupils to receive another intervention during the session to help close the attainment gap.

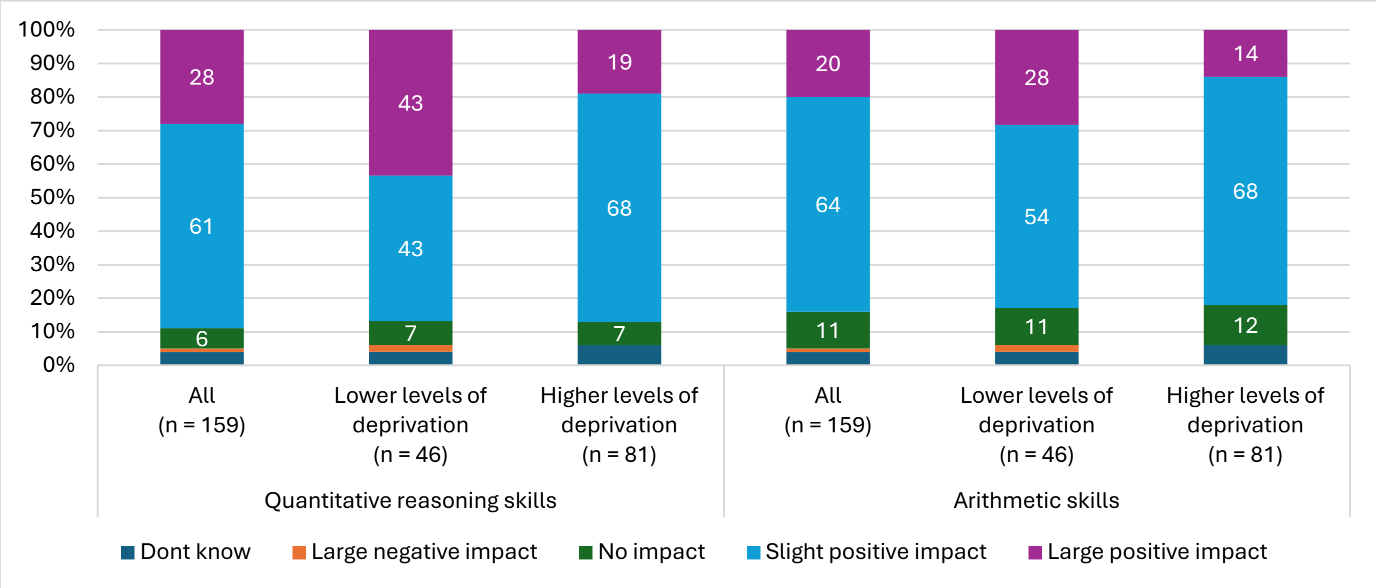
Perceived impact

IPE RQ4: *What was the perceived impact of the Mathematical Reasoning programme for: i) staff members; ii) pupils; and iii) disadvantaged pupils specifically?*

Teachers and TAs perceived the Mathematical Reasoning programme to have had a positive impact on pupils' quantitative reasoning and arithmetic skills

The large majority of surveyed teachers perceived a positive impact of participating in the Mathematical Reasoning programme on pupils' quantitative reasoning (92%) and arithmetic skills (86%) (Figure 19). Most TAs likewise reported a positive impact in these areas. However, both teachers and TAs were more likely to report a 'slight' positive impact than a 'large' one across both areas. Teachers in schools with lower levels of deprivation tended to report larger positive impact compared to those with higher levels of deprivation. This suggests that the programme may have been perceived to be less effective for more disadvantaged children, who would constitute a larger proportion of the cohort in more deprived schools.

Figure 19: Teacher and TA views of the impacts of the Mathematical Reasoning programme on pupils' quantitative reasoning and arithmetic skills



Source: Question Q.1 and Q.2 (Overall, what (if any) impact do you feel participating in the Mathematical Reasoning programme has had on your pupils in terms of their...) from the Endpoint Teacher and TA Survey (N=159).

Most of the case study schools reported having observed positive impacts for their pupils' maths abilities as a result of participating in the programme. This was also mentioned by several schools during the webinars. Improvements were noted in pupils' reasoning and problem-solving abilities, particularly in relation to money and inversion, which teachers attributed to participation in the Mathematical Reasoning programme. Some teachers reported that pupils had a better understanding of what the questions were asking for as a result of the programme's focus on relevant vocabulary and word problems. The Mathematical Reasoning programme was also seen to have given pupils the skills and strategies to approach reasoning and problem-solving questions, such as using annotations and drawing visual aids to support their thinking or making informed decisions about which manipulatives to draw upon to guide them. Pupils were consequently better equipped to attempt this style of questioning, which then developed their reasoning skills further.

> *We have a lot of resources available for use in the classroom and the children are now more willing to go and choose the appropriate resource for a task, not just one they want to play with. Because they used resources so thoroughly in the practice of the programme, they're now more competent in making the choice for themselves of which tools will actually help them.* (Teacher, CS4)

Staff in a few schools also reported that pupils were better able to verbalise, explain, and discuss their reasoning. They attributed this to the opportunities the programme created for group discussions in response to the questions being posed. One teacher felt 'reluctant speakers' and pupils with EAL had particularly benefited in this aspect.

> *I think the concept of the opportunities for discussion have been really powerful for the children. We've had children who have been very shy to talk as part of the class and as a school we've been focusing on oracy a lot but having something specific to discuss as part of learning, not a stand-alone lesson we're you're learning to talk as a class, was good. It was an opportunity for the children to explore discussion and how to incorporate manipulatives into developing their reasoning—that was really useful for them.* (Teacher, CS4)

Of the schools that reported these positive impacts, most reported that pupils were applying their new reasoning skills and problem-solving strategies to other maths lessons and topics, such as place value and times tables. They also reported that pupils were applying the skills/strategies they had learned to internal maths assessments and in optional SATS, and that they were more familiar with the reasoning-style questions and word problems that these involved. Staff in several schools felt that the programme had contributed to the improvements they had observed in pupils' performance in these assessments. In a small number of cases, however, pupils made references to 'guessing the answers' when doing the computer games, which suggests learning from the teacher-led work was not always transferred to other activities.

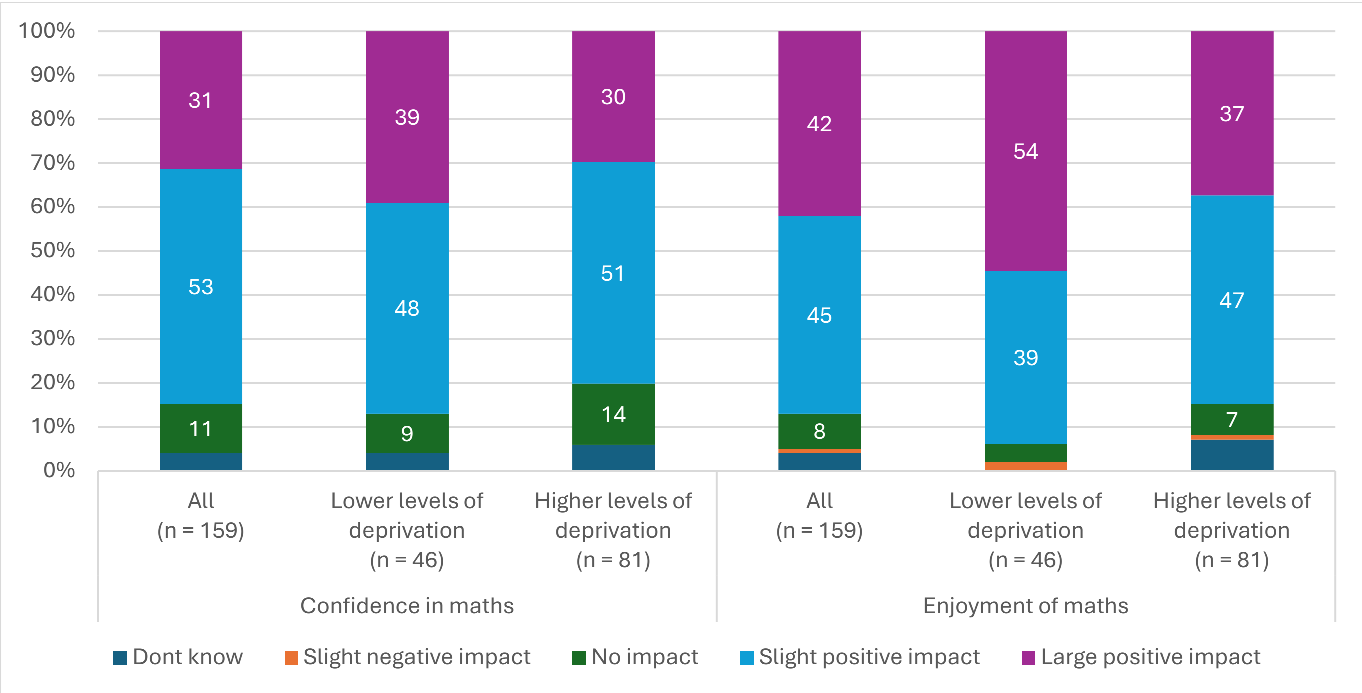
However, positive impacts for pupils were not universally observed. A few case study schools reported that they had not identified any notable differences in their progress compared to previous Year 2 cohorts. Others found it difficult to identify the 'added value' of Mathematical Reasoning compared to the gains their pupils would otherwise have made. As detailed above in relation to pupil-level moderators, some staff reported that the programme has less impact for lower ability pupils, who did not have secure foundational maths skills and arithmetic fluency required to participate in the programme.

Overall, most teachers perceived the programme to have had a positive impact on pupils' maths abilities, but the magnitude of this impact and the extent to which it benefited all pupils equally was more contested.

Teachers and TAs perceived the Mathematical Reasoning programme to have had a positive impact on pupils' confidence in and enjoyment of maths

The large majority of surveyed teachers reported perceiving the programme to have had a positive impact on pupils' confidence in (84%) and enjoyment of (87%) maths (Figure 20). Very similar results were seen among TAs. However, as with impacts on mathematical abilities, higher proportions of teachers and TAs reported a 'slight positive impact' than a 'large positive impact' in these areas, and greater proportions of teachers in less deprived schools reported larger impacts than those in more deprived schools.

Figure 20: Teacher and TA views of the impacts of the Mathematical Reasoning programme on pupils' confidence in maths and enjoyment of maths



Source: Question Q.3 and Q.4 (Overall, what (if any) impact do you feel participating in the Mathematical Reasoning programme has had on your pupils in terms of their…) from the Endpoint Teacher and TA Survey (N=159).

Staff in several case study schools similarly reported pupils' increased confidence in relation to mathematical reasoning, including attempting questions and activities that they may not have attempted in the past. Two teachers felt that because the workbooks were not formally marked, the pupils felt less pressure to answer the questions correctly so were more willing to 'have a go'. Staff in several schools highlighted that the practical and pictorial nature of the programme (such as the games, the use of manipulatives and resources) had made maths engaging and interactive for pupils and had contributed to their enjoyment of the programme.

Participation in the Mathematical Reasoning programme had a positive impact on intervention teachers' understanding of mathematical reasoning topics, particularly in the least deprived schools

Teacher understanding and practice in relation to teaching mathematical reasoning are a key objective of the Mathematical Reasoning programme. A long-term goal of the programme is for participating teachers to transfer their learning to other areas of their practice and share it with other colleagues so that as many pupils as possible are taught mathematical reasoning in an effective manner across the curriculum. In the short-term, the programme aimed to increase teacher and TA understandings of quantitative and numerical reasoning. This trial focused on teacher outcomes in order to reduce the data collection burden on TAs.

The baseline and endpoint surveys asked all teachers to rate their level of understanding on a variety of mathematical reasoning topics. Teachers' understanding of the topics improved for both groups between baseline and endpoint, with the greater change observed in the intervention group (Figure 21). The change in intervention teachers' understanding between baseline and endpoint was more positive than the same change for control teachers ($p<0.001$) in four areas[30]:

- doing and undoing an action;
- quantitative reasoning;
- the difference between additive and multiplicative reasoning; and
- one-to-many correspondence.

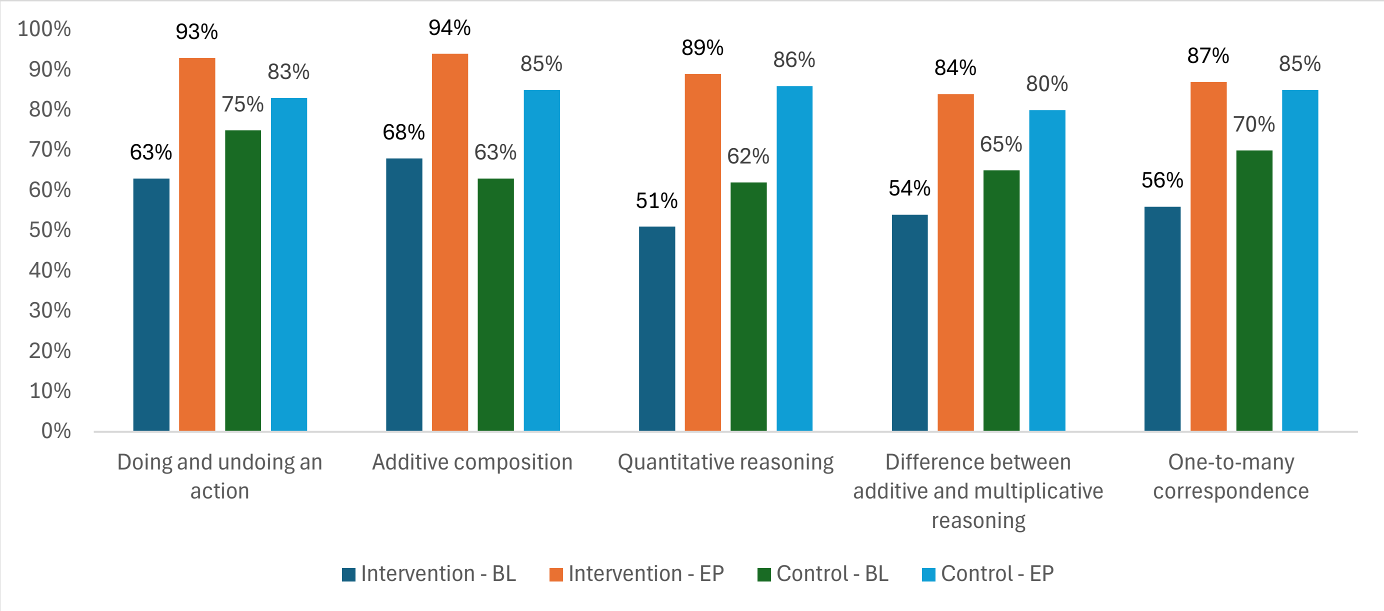
For the additive composition topic it was less clear whether the change in the intervention group was greater than for the control group ($p=0.086$).

The observed increases in self-reported understanding by control school teachers suggest that some control teachers developed their maths reasoning knowledge during the trial period, though not to as great an extent as teachers in the intervention group. This raises the question of whether the improved understanding of control teachers attenuated the programme effects observed in the impact analysis. Some or all of the increased understanding observed among control teachers could be due to the usual growth in professional knowledge that would occur anyway outside of the trial. Increased understanding of this sort would not be an artefact of trial participation and therefore would not attenuate the trial results. To some degree, this possibility is supported by the increase in understanding observed for maths topics unrelated to mathematical reasoning[31] group. Furthermore, improved understanding of mathematical reasoning topics among control teachers would also have to substantially change their teaching practice (despite the absence of content from the Mathematical Reasoning programme) for the impact results to be attenuated. Little or no attenuation therefore seems likely, although we cannot rule it out altogether.

[30] The statistical tests performed were Mann-Whitney U tests, treating the Likert-type scales as ordinal data. Therefore, we note that the comparison is not exactly the same as in the figure, which compares the proportion of teacher rating their understanding 'good' or 'very good'.

[31] In the control group, the proportion reporting a 'good' or 'very good' understanding of interpreting tally charts improved from 97% to 99%, the proportion for measuring lengths improved from 92% to 98% and the proportion for recalling multiplication facts went from 85% to 92%. It is hard to compare these increases with those seen in mathematical reasoning topics, due to the higher levels of understanding at baseline likely causing ceiling effects here.

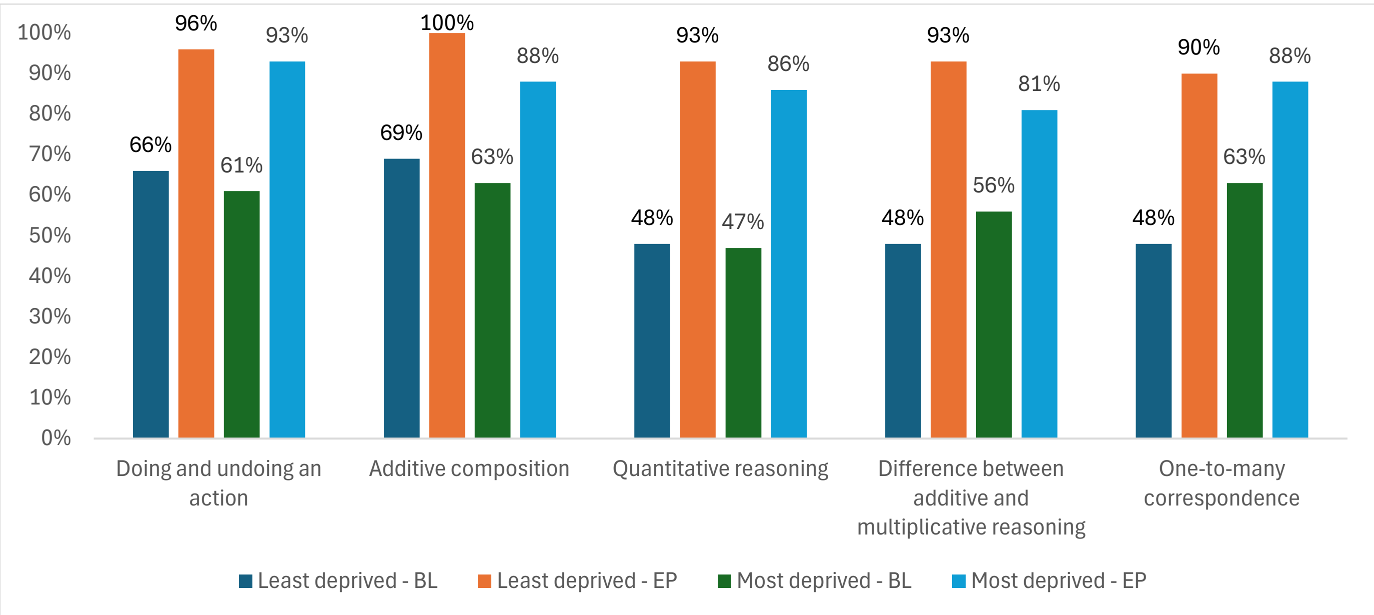
Figure 21: Proportion of control and intervention teachers at baseline and endpoint who rated their understanding 'good' or 'very good' across various areas of mathematical reasoning



Source: Question J in the Baseline BAU survey (intervention group n=88; control group n=96), Question K in the Endpoint BAU survey (control n=95), and Question U in the Endpoint Teacher and TA Survey (intervention n=87): 'How would you rate your understanding of the following maths topics?'
BL = baseline; EP = endpoint.

Looking just within the intervention schools, overall, the change in reported confidence levels was higher among teachers in schools with lower levels of deprivation for all topic areas except for 'doing and undoing an action' (Figure 22).

Figure 22: Proportion of intervention teachers who rated their understanding 'good' or 'very good' across various areas of mathematical reasoning, by deprivation level



Source: Question J in the Baseline BAU survey (least deprived n=29, most deprived n=43), and Question U in the Endpoint Teacher and TA Survey (least deprived n=29, most deprived n=42): 'How would you rate your understanding of the following maths topics?'
BL = baseline; EP = endpoint.

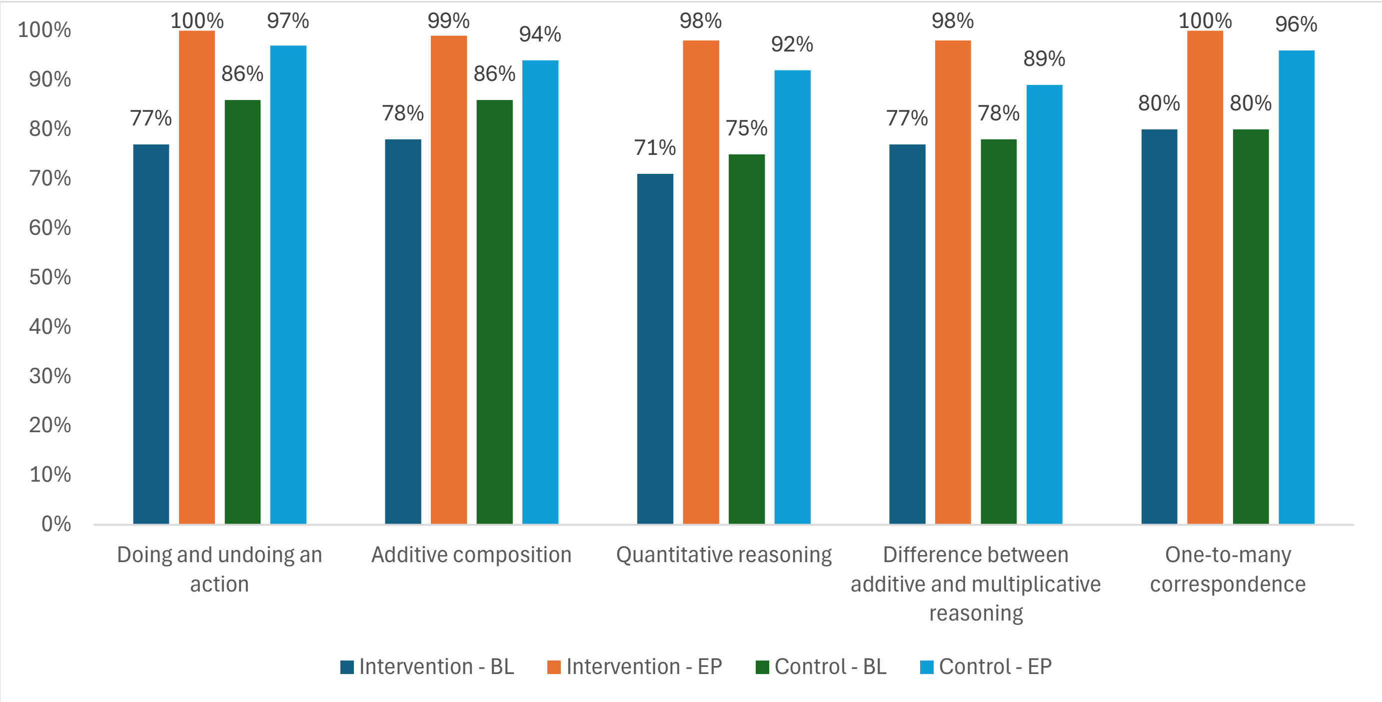
The Mathematical Reasoning programme had a positive impact on intervention teachers' confidence in teaching mathematical reasoning in Year 2

There is evidence to indicate that teachers with high levels of self-efficacy are more open to new teaching methods (Lazarides and Warner, 2020). Confidence is commonly used as a proxy for self-efficacy (Parchment *et al.*, 2025). As a result, teacher confidence in teaching different areas of mathematical reasoning is likely to predict the probability of them teaching these elements in future. This is an important outcome for any wider change in practice.

The baseline and endpoint surveys asked all teachers to rate their confidence in teaching various mathematical reasoning topics. Teachers in intervention schools demonstrated a larger increase between baseline and endpoint across all areas as compared to teachers in control schools (Figure 23). This difference was statistically significant in two areas: additive composition ($p = 0.045$); and quantitative reasoning ($p<0.001$). This suggests that the programme had an impact on teachers' confidence in teaching these aspects of mathematical reasoning.

However, as the figure shows, an increase of between 8 and 11 percentage points between baseline and endpoint was still observed for control teachers. This suggests, again, that control schools may have increased their focus on mathematical reasoning over the trial period. Note that we did not perform significance testing to determine whether the size of the increase was greater than zero among control teachers (and likewise we did not test this among intervention teachers). What we tested was whether the size of the increase was different between the groups (a 'difference-in-differences' approach).

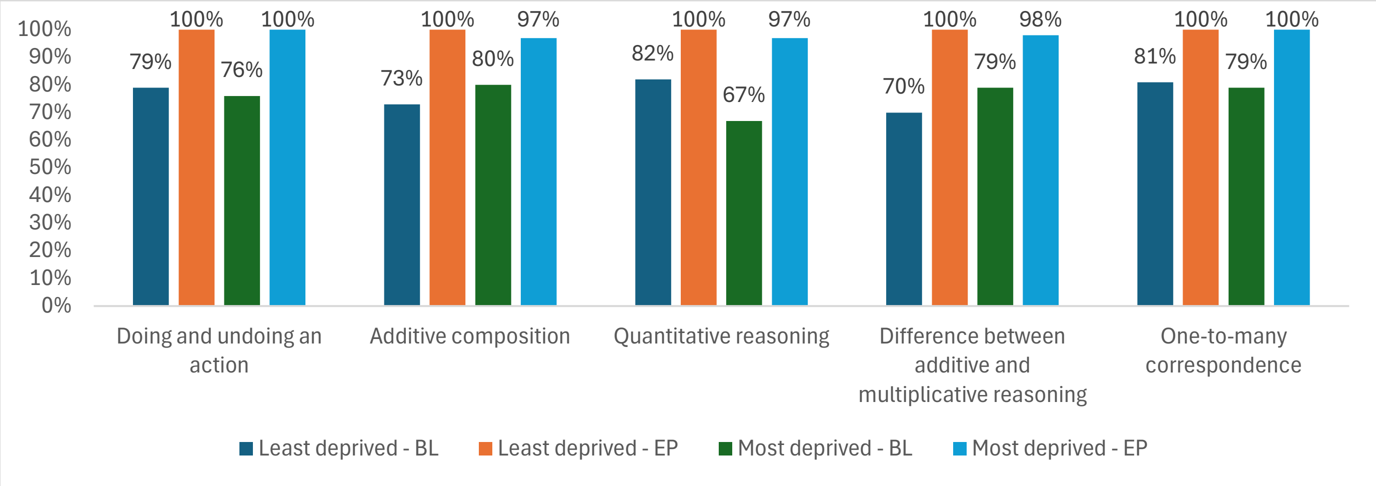
Figure 23: Proportion of control and intervention teachers who reported feeling 'fairly confident' or 'very confident' in teaching various areas of mathematical reasoning



Source: Question K in the Baseline BAU survey (control n=84–92, intervention n=73–82), Question L in the Endpoint BAU survey (control n=91–95), and Question V in the Endpoint Teacher and TA Survey (intervention n=86). 'How confident are you in teaching the following topics to your Year 2 maths class?'
Note: Staff who did not select 'I don't know what this is' in the previous question relating to each of these areas were not asked about their confidence in that area. The chart numbers only include those who saw the question, consistent with the reporting approach throughout this report.
BL = baseline; EP = endpoint.

While teachers in the least deprived schools saw a greater change in reported confidence for 'additive composition' and 'the difference between additive and multiplicative reasoning', a larger change was actually seen for teachers in the most deprived schools for 'quantitative reasoning' (Figure 24).

Figure 24: Proportion of intervention teachers who reported feeling 'fairly confident' or 'very confident' in teaching various areas of mathematical reasoning, by deprivation level



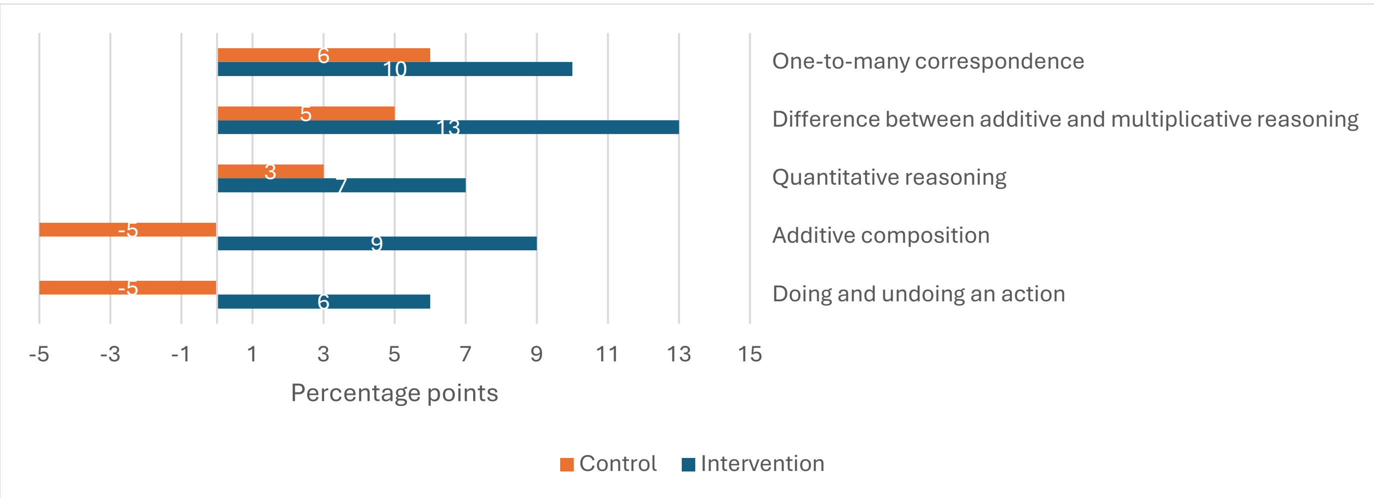
Source: Question K in the Baseline BAU survey (least deprived n=21–26, most deprived n=38–42), and Question V in the Endpoint Teacher and TA Survey (least deprived n=29, most deprived n=41). 'How confident are you in teaching the following topics to your Year 2 maths class?'
Note: Staff who did not select 'I don't know what this is' in the previous question relating to each of these areas were not asked about their confidence in that area. The chart numbers only include those who saw the question, consistent with the reporting approach throughout this report.
BL = baseline; EP = endpoint.

As expected, confidence in teaching for the reference items did not appear to be influenced by participation in the Mathematical Reasoning programme: almost all control and intervention group teachers reported being 'fairly confident' or 'very confident' at baseline and endpoint.

Most teachers already felt at baseline that mathematical reasoning topics should be first taught in Year 2 or earlier—but a larger increase was still seen in the intervention group over the course of the trial

Another key outcome identified in the Theory of Change for shifting teacher practice around mathematical reasoning was belief that it should be taught to pupils in Year 2 or earlier. The baseline and endpoint surveys listed mathematical reasoning topics and asked teachers to indicate in which year of primary school they felt the topics should first be taught. Of interest was whether intervention teachers would demonstrate a shift in their views towards teaching these topics in Year 2 or earlier by endpoint. While, at baseline, most teachers in the control and intervention groups (between 66% and 91%) already felt that mathematical reasoning topics should be taught in Year 2 or earlier, the proportion of intervention teachers that indicated this showed a slightly larger increase between baseline and endpoint than was the case for control teachers (Figure 25).

Figure 25: Change between baseline and endpoint in the proportion of control and intervention teachers who reported that various mathematical reasoning topics should be taught in Year 2 or earlier



Source: Question L in the Baseline BAU Survey (Control N=89, Intervention group N=79), Question M in the Endpoint BAU Survey (N=90) and Question W from the Endpoint Teacher and TA Survey (N=86): 'At what stage of the curriculum do you feel the following topics should be first taught?'

Staff in case study schools highlighted positive impacts of the Mathematical Reasoning programme for their teaching practice

Nearly two-thirds of surveyed teachers (61%) indicated that participating in the Mathematical Reasoning programme had ‘somewhat’ changed how they taught maths outside of the programme. A further 6% reported it had changed ‘substantially’. Impact on TA practice appears to have been less substantial, with just less than half of surveyed TAs (28 out of 58) indicating that participating in the Mathematical Reasoning programme had ‘somewhat’ impacted their maths teaching practice outside of the programme. This perhaps reflects the lower levels of TA participation in the training course and webinars.

Several teachers and TAs in case study schools reported that the Mathematical Reasoning training improved, or refreshed, their understanding of the importance of teaching mathematical reasoning so that pupils are encouraged and able to apply reasoning knowledge and strategies. Some teachers and TAs also reported that the training gave them new strategies for teaching reasoning and problem-solving concepts to their Year 2 pupils and gave them the confidence to implement the programme in their classrooms.

> *I do feel confident. It is one of those things where you always need refreshers. It is so important, but you can become so focused on the fluency you don’t spend as much time on the reasoning element so I have found it really useful to refresh my memory and teach me new things about it.* (Teacher, CS7)

Some teachers also reported finding the training helpful for understanding how pupils might approach reasoning concepts, the misconceptions they may display, and how pupils ultimately reach their level of understanding, while delivering the sessions gave them a better understanding of pupils’ mathematical reasoning capabilities. For example, in one case, delivering the programme highlighted to teachers that pupils could achieve more than they previously expected; in another, delivering the programme had revealed unexpected challenges relating to money. Teachers reported that the ‘in the moment marking’ during the whole-class teaching, and the chance for them to work with a smaller group of pupils when the class split, were key drivers of this increased understanding.

Teachers in most case study schools identified changes they had made or would make to their practice as a result of participating in the Mathematical Reasoning programme. Most often, teachers said they were better able to identify opportunities within maths lessons where they could incorporate elements of reasoning, using strategies from the programme to do so. While a few teachers were already doing this at the time of the case studies, others were planning to incorporate more reasoning into their maths teaching practice next academic year. Some teachers also indicated intentions to use manipulatives more, as well as increase their emphasis on repetition.

> *I think it has been really useful, it has given me a different way of looking at it, adding in those opportunities for reasoning that I wouldn’t necessarily think about before. I use resources [manipulatives] anyway but it reminded me how important that is and I will add that in next year.* (Teacher, CS3)

Many teachers reported that they planned to take specific elements from the programme to incorporate into their maths teaching practice longer term. This included: providing real-life contexts for pupils to practice using money; the structure of the Mathematical Reasoning sessions (i.e. whole-class teaching input, followed by two ability-based groups); and increasing the use of visuals to support pupils to access concepts and answer questions. Teachers also reported they would use some of the teacher-led activities again as the pupils had particularly enjoyed them—namely the Gremlins Game and the Caterpillar Game. Staff in one school suggested they may use Mathematical Reasoning as a resource across their Year 2 classes rather than as a 12-week programme, by drawing upon the Mathematical Reasoning materials over the course of the year to supplement their wider scheme of work.

Staff in a couple of case study schools reported no wider impacts on their maths teaching practice as they felt the specificity of the training and the programme limited wider application of any learning.

Around half of participating staff reported that they had shared learning from the programme with colleagues

One of the staff-level outcomes in the Theory of Change is for teachers and TAs to share insights from the programme across the school. Most teachers and TAs reported having shared or the intention to share learning with their colleagues. This most commonly took the form of informal sharing (47% of teachers). Just 5% of teachers indicated that they had shared learning through formal training. Just less than a third (29%) of teachers had not yet shared learning with colleagues but intended to do so. Very similar results were seen for TAs.

Some case study schools reported having shared learning from the programme with other colleagues, either by sharing relevant resources with another teacher or by presenting on the programme to all staff as part of a meeting.

However, more common was the intention to share learning or materials, either following completion of the Mathematical Reasoning programme in the 2024/2025 academic year, or in the following academic year, through Planning, Preparation, and Assessment (PPA) time and/or sharing the materials with other Key Stage 1 staff. Maths leads in two schools shared that participating in Mathematical Reasoning had encouraged them to think about how reasoning is taught in other year groups, with the suggestion that in the future they may approach this through delivering dedicated reasoning sessions in other year groups or consider how reasoning can be best incorporated into their existing maths lessons.

> *We've already factored it into our team meetings…in the Autumn Term [2025/2026]. They're going to share some of the training materials that were shared with them to other Key Stage 1 staff. I would normally lead it, but they're [teacher and TA] are going to do it. I think it's a good way they can demonstrate their own knowledge and understanding.* (Maths lead, CS1)

Maths leads in two schools shared that participating in the Mathematical Reasoning programme had encouraged them to think about how mathematical reasoning is taught in other year groups, with the suggestion that in the future they may approach this through delivering dedicated mathematical reasoning sessions in other year groups or consider how mathematical reasoning can be best incorporated into maths lessons.

Cost evaluation

Schools participating in the trial received the programme for free. As a result, during the trial, the programme cost to the average participating school was £2.63 (for printing; Table 30), plus any additional costs for those very few schools that required additional manipulatives or IT resources, or who paid for cover time for staff to complete the training (details on these costs are provided below). The information was gathered as part of the sixth IPE research question: How much does it cost a school to deliver the Mathematical Reasoning programme?

In line with the EEF cost evaluation guidance (EEF, 2023), the cost evaluation below describes the estimated full costs to schools if the programme was rolled out commercially, including the costs of purchasing the programme itself.
On average, teachers spent 11.25 hours completing the programme training and 14.5 hours on session delivery, over the course of the trial (Table 29). TAs spent an average of 7.75 hours on training and 12.75 hours of session delivery. It is assumed the session delivery time would be similar in the second and third years of delivery, but the training time would not be repeated.

However, as outlined in the 'IPE results' section above, not all classes had the support of a TA during Mathematical Reasoning sessions. If the classes without TA support are included in the calculations, the average time teachers spent on delivering the Mathematical Reasoning sessions reduces to 11.5 hours, suggesting that without the TA support available the teachers are less likely to be able to deliver the whole programme.

Preparation time for delivery varied widely but on average was the same as the preparation time for their usual maths lessons (Table 29).

The data in Table 29 is based on what respondents reported actually occurred in the trial. If the programme were delivered as intended, time spent on the webinars would increase to 4.75 hours for both the teacher and TA. Teacher and TA time on session delivery would still be between 12 and 15 hours, but they would be the same value as the TA would support every session.

Table 3: Time spent on Mathematical Reasoning delivery per class

		Year one		Year two		Year three	
		No. of staff members	Mean no. of hours (min, max)	No. of staff members	Mean no.of hours (min, max)	No. of staff members	Mean no. of hours (min, max)
Training course	Teacher	1	6.5 (1.5, 10.5)	0	0	0	0
	TA	1	5 (1, 10.5)	0	0	0	0
Webinars	Teacher	1	4 (2, 4.75)	0	0	0	0
	TA	1	2.75 (1.25, 4)	0	0	0	0
Liaising with the TL	Teacher	1	0.75 (0, 2)	0	0	0	0
Session preparation	Teacher	1	0 (-18.75, 6.25)[a]	1	0 (-18.75, 6.25)[a]	1	0 (-18.75, 6.25)[a]
Session delivery (for classes with teacher and TA)	Teacher	1	14.5 (6.25, 28.25)[b]	1	14.5 (6.25, 28.25)[b]	1	14.5 (6.25, 28.25)[b]
	TA	1	12.75 (1.25, 24.5)	1	12.75 (1.25, 24.5)	1	12.75 (1.25, 24.5)

Note: All amounts of time in this table have been rounded to the nearest quarter of an hour (i.e. 0.25).
[a] This refers to the difference in time spent compared to usual preparation time required for a maths lesson, as each session was intended to replace an existing maths lesson. Negative values represent time saved compared to usual practice.
[b] This time occurred during usual teacher time so did not represent any additional workload.

Sources: Endpoint Teacher and TA Survey (n=102 teachers, 63 TAs), case study cost pro formas (n=8), Session Delivery Logs (n=104).

Implementing the programme costs each school approximately £1,457.88 per class for three years (Table 30). Assuming 25 pupils per class, this means £19.44 per pupil over this period. The large majority of this cost is for purchasing the programme materials and training.

- There were some variations in delivery costs that may be of interest to note:10% of surveyed teachers reported having paid cover for the time spent completing the online training course, an average additional cost of £218 per class (cost per pupil over three years of £22.35).

- Of the surveyed TAs who did the training course, 7% (4 out of 53) reported having paid cover for it. If both the teacher and the TA were given paid cover, this would cost the school an additional £436 per class, increasing the cost per pupil over three years to £25.25.

- Paid cover for the time they spent attending the webinars was reported by 5% of surveyed teachers, at an average cost of £327 per class. This would increase the cost per pupil over three years to £23.80. Of the surveyed TAs who attended at least one webinar, 7% (3 out of 44) reported having paid cover for this, which would cost schools an additional £218 per class. If both the teacher and the TA were given paid cover to attend, this would cost the school an additional £545 per class (cost per pupil over three years of £26.71).

- If the teacher and TA completed all training course modules and attended all three webinars, as intended, and their school paid for cover for all of these sessions, the cost per pupil over three years would increase to £33.97. However, this still represents a very low-cost programme.

Table 4: Cost of delivering Mathematical Reasoning

Item	Type of cost	Cost (mean)	Total cost over three years	Total cost per pupil per year over three years
Purchasing materials (perishable, need to be purchased for each cohort)	Running cost per class	£400	£1,200	£16.00
Training attendance fee	Start-up cost per class	£250	£250	£3.33
Printing (certificates)	Running cost per class	£2.63	£7.88	£0.11
Total			£1,457.88	£19.44

Source: Case study cost pro formas (n=8), Oxford University delivery team.

Almost all schools already had the prerequisite materials for delivering the programme, with only 1% of surveyed teachers reporting that their school purchased any of the technology necessary for programme delivery (Table 30). If the school had to purchase devices for the programme, this would cost on average an additional £1,326 per class (increasing the per pupil cost over three years to £37.12) (see Appendix P of Further Appendices).

While 17% of teachers reported purchasing additional manipulatives (Table 30), this resulted in only a marginal cost increase: the cost pro forma indicated that on average £32.67 was spent on items such as coins and counters, resulting in a per pupil cost over three years of £19.88.

Table 5: Prerequisites for the implementation of Mathematical Reasoning

	Year one		
	Quantity required (mean)	Mean additional quantity required (min, max)	% of schools which purchased items
Computer (for staff use)	1	1	1%
Screen (for session delivery)	1	1	1%
Devices (for pupil use)	12	10	1%
Manipulatives	Variable	Variable	17%

Source: Endpoint Teacher and TA Survey (N=101 teachers).

The total cumulative cost to a school of delivering the Mathematical Reasoning programme over three years is £1,457.88 (Table 32).

Table 6: Cumulative costs of Mathematical Reasoning (assuming delivery over three years)

	Year one	Year two (cumulative)	Year three (cumulative)
Mathematical Reasoning	£652.63	£1,055.26	£1,457.88

Conclusion

Table 7: Key conclusions

Key conclusions
1. Pupils in Mathematical Reasoning schools made one additional month of progress in maths, on average, compared to pupils in other schools. This result has a high security rating.
2. Among pupils eligible for free school meals (FSM), those in Mathematical Reasoning schools made two additional months of progress in maths, on average, compared to those in other schools. This result has lower security than the overall finding due to a smaller number of pupils.
3. There was some cautious evidence that pupils with lower prior attainment benefited more from the programme than pupils with higher prior attainment. However, the IPE did not explain this finding, with teachers more likely to perceive the programme as better suited to higher attaining pupils.
4. The IPE showed that the programme had a positive impact on self-reported teacher understanding of and confidence in teaching mathematical reasoning, and most teachers felt it had a positive impact on pupils' mathematical reasoning and arithmetic skills, as well as their confidence in and enjoyment of maths.
5. Teachers reported that TA support was important for delivering group activities as intended. Although programme materials supported delivery, training and programme delivery required substantial time commitment, contributing to some issues with fidelity.

Impact evaluation and IPE integration

Evidence to support the logic model

The findings from this trial indicate that the overall path from inputs to outcomes expressed in the Theory of Change is broadly sound. There is sufficient evidence from the impact evaluation to indicate that the short-term outcome of participating children having enhanced mathematical performance was achieved, and there is some evidence from the IPE to suggest that the mediating outcomes of increased pupil understanding of quantitative reasoning, improved arithmetic skills, and greater confidence in and enjoyment of maths were observed. Pupils taking part in the programme also made more progress in the 'mathematical reasoning' domain as measured by the PTM7 subscale, on average, compared to those in other schools. The short-term teacher outcomes also appear to have been achieved, with an increase seen in teachers' self-reported understanding of quantitative and numerical reasoning and recognition of the importance of teaching mathematical reasoning to this age group. The additional teacher outcome of greater confidence in teaching mathematical reasoning to Year 2 pupils has also been sufficiently evidenced to be added to the Theory of Change (see Figure 26). However, the low rate of TA training completion suggests that the predicted TA outcomes in these areas are unlikely to have occurred in practice. The medium-term outcome of teachers and TAs sharing insights from the programme across the school occurred only to a limited extent within the trial time frame.

Although the training and support were delivered as intended (inputs), poor fidelity was observed in relation to a number of the activities and outputs, particularly in relation to training completion by teachers and TAs, TA involvement in the delivery of sessions, splitting the class, and use of manipulatives. Lower levels of fidelity appear to have been in part due to external challenges, such as staffing constraints, but there are also indications that teachers did not always have a strong understanding of the programme requirements. This was particularly demonstrated in the case study interviews where teachers were implementing approaches to splitting the class, extending the session length, and/or skipping some of the content, apparently without being aware that this was not intended practice. There was also evidence that not all teachers managed to deliver all of the content during the course of the programme.

Some of the TLs felt that the new training model enabled more teachers to complete the training compared to the previous iteration's requirement to attend in person, but they thought that by completing the training remotely, teachers may be subject to more distractions. Furthermore, the delivery team's intention for the programme to enable collaborative learning practice between the teacher and TA when completing the training course together does not appear to have been practical for most schools, primarily due to staffing constraints. As a result, most teachers completed the training course on their own, which meant the collaborative aspect was missing.

The role of the TL, which was introduced as part of this new training model, was well received by teachers.

Overall, there is evidence to suggest that at least some programme impacts can still be achieved even when deviations in delivery of some aspects of the programme occur. It may also be the case that greater fidelity to the intended delivery design would increase the size of impact, but this cannot be evidenced at this stage, and it seems that the time commitment for teachers and logistical and structural barriers (e.g. sustained availability of TAs) are not feasible for schools to meet in their entirety in the current model. The developer may wish to consider some minor adjustments in these areas to support greater fidelity if the programme is offered to schools in the future (Table 34).

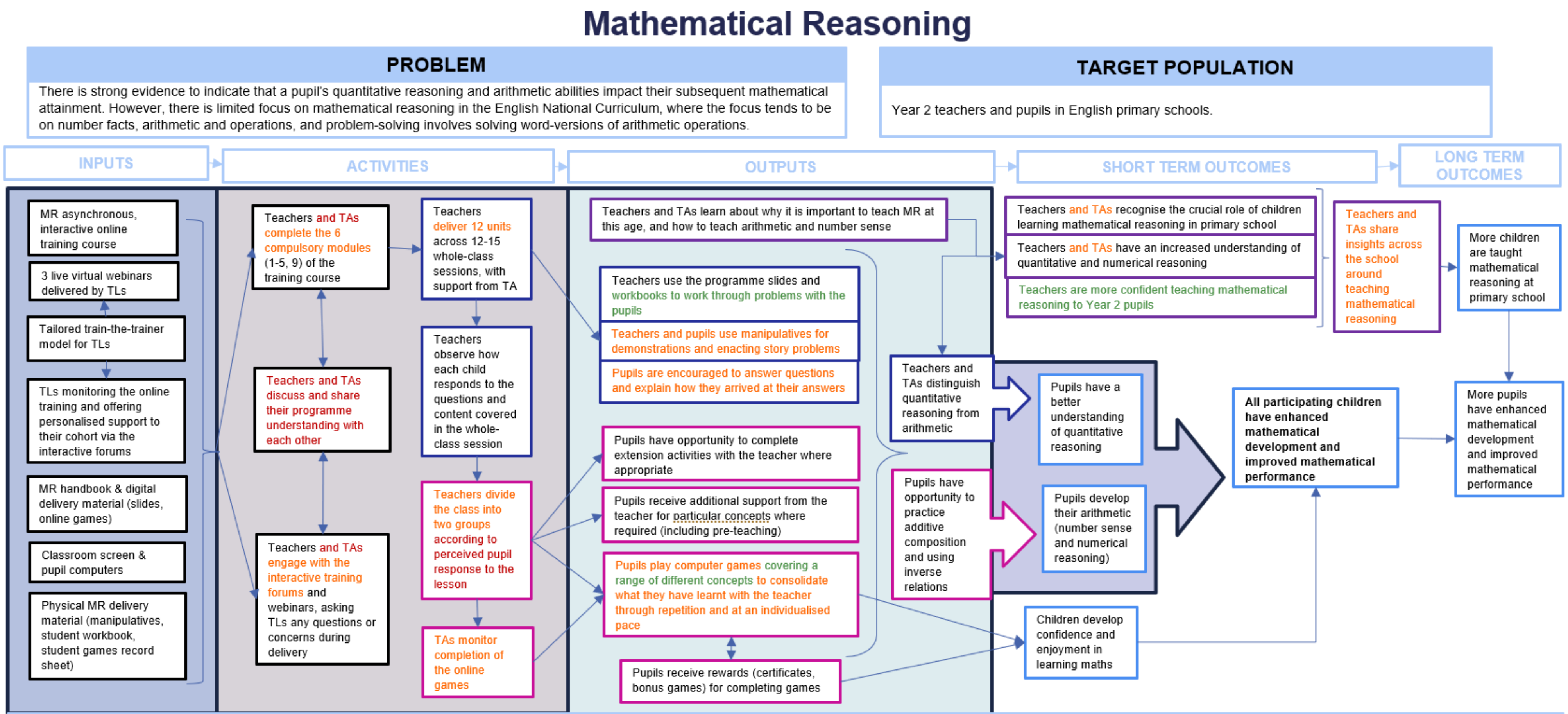
Table 8: Commentary on the Mathematical Reasoning programme Theory of Change

Areas of low fidelity	Implications
Teachers did not always deliver all of the programme material during the first, teacher-led, part of the session	The IPE revealed that teachers were often skipping content within units due to time pressures. Despite this, the impact analysis suggested that there may be a positive impact of the programme. As a result, developers may want to consider the amount of material the programme asks teachers to cover and whether this could be reduced. The amount of content to be covered is one of the key issues that has emerged in all three evaluations of this programme
Teachers did not complete all of the compulsory training modules	Consider condensing the training. As well as potentially enabling more teachers to complete all of the training sessions, this could reduce any negative unintended consequences of teacher displacement for pupils (i.e. additional cover teachers) and increased workload for teachers
Teachers and TAs did not always complete the training together	Consider how essential TA attendance at all training sessions is and whether TA attendance at key modules could be prioritised
Teachers and TAs often engaged only superficially with the interactive training forums	The time TLs spent checking and responding to participant contributions to the online training activities may be unnecessary as a key, or essential, component due to the limited interaction and follow-up observed
Teachers did not always divide the pupils into groups, and very rarely did so based on how the pupils responded to that lesson	The group work in the second half of the session appeared to be valuable to the programme: the computer games dosage analysis suggests that playing these games was beneficial for pupils, and the IPE indicates that teachers and pupils appreciated the opportunity for the teacher-led smaller group work However, the evidence suggests that it was not feasible for teachers to regularly and consistently group the pupils based on their responsiveness in each lesson, and sometimes not at all. This typically occurred due to challenges around TA availability, and logistical constraints within the classroom
Teachers and pupils did not consistently use manipulatives when problem-solving	The IPE suggests that most teachers valued the use of manipulatives as part of the programme, and pupils enjoyed using them, though their use was inconsistent. The time to prepare the manipulatives for sessions could be time consuming, which may have contributed to inconsistent use
Teachers varied in the extent to which they encouraged pupils to explain how they arrived at their answers	The role explanation, or the 'act of reasoning' plays in the Theory of Change suggests making this more consistent could increase the overall programme effect for pupils

In the updated Theory of Change (Figure 26) we have also suggested two additions to the programme activities that the evidence suggests should be included as core components. One is that teachers deliver Mathematical Reasoning sessions using the programme slides and workbooks to work through problems with the pupils. It is likely that the semi-scripted nature of the programme contributed to the consistent impact observed between the three trials, despite some aspects of low fidelity in each case. It appears that delivering *all* of the programme content is perhaps not necessary to achieve benefits for pupils, but the format in which it encourages teachers to approach problems with their pupils appears to have been important.

In addition, the computer games moderator analysis suggests that it is beneficial for pupils to play games covering a range of different topics or concepts. As a result, we have suggested adding this specification to the output relating to pupil use of the computer games.

Figure 26: Updated Theory of Change for Mathematical Reasoning



Moderators / contextual factors

Pupil-level: FSM status; **SEN**; *home access to and* facility with IT; *proportion of group sessions spent in teacher-led vs online games groups*; *confidence in maths; enjoyment of maths*

School-level: Teacher/TA's existing knowledge, experience, skill and commitment to MR; **SLT and broader staff support; human, physical, financial** & *IT* **resources**; approach to maths teaching outside of sessions; level of need and/or ability in the class

Implementation-related: number of sessions delivered; proportion of content delivered; *length of sessions*; *time spent on online games outside of sessions*; **level and nature of TA support;** *completion of mandatory training – including webinars (teacher, TA); implementation fidelity*; *nature of learning displaced by the sessions*; *quality of TL support*; *proportion of time spent in whole-class vs groups*; **number of computer games played**

Relating to long-term outcomes: *completion of 3 optional modules (teacher, TA);* school culture (e.g. around CPD, mathematical reasoning)

Notes: Elements in 'amber' and 'red' were seen to occur in practice to a limited extent or a very limited extent, respectively. Elements in 'green' have been added to the Theory of Change as a result of what was observed in the trial. Moderators/contextual factors in bold were observed to be key factors in the trial. There was no indication from the trial that those factors in italics constituted particularly important moderating or contextual factors.

Interpretation

The Mathematical Reasoning programme has previously been evaluated with both efficacy (Worth *et al.*, 2015) and effectiveness (Stokes *et al.*, 2018) trials. This, second, effectiveness trial of the programme was commissioned as a result of the training model changing from a train-the-trainer model to an online training course (supported by the TL role), developed by the core delivery team (see Appendix C). This change was intended to increase programme fidelity with the aim of bringing the effect size closer to what was seen in the efficacy trial. However, this second effectiveness trial found the same small effect for the primary outcome for all pupils as the first effectiveness trial (meaning both effectiveness trials found smaller effect sizes than the efficacy trial did), using the same outcome measure (see below for discussion on differences found in relation to the secondary RQs). While the associated p-value of 0.052 would not be 'statistically significant' at the 5% level in a hypothesis testing framework, we prefer to view the p-value as a continuous measure of evidence against the null hypothesis. This is especially true given that the MDES at the analysis stage of 0.131 was higher than the MDES of 0.117 anticipated in the **study protocol** (Flemons *et al.,* 2024), reducing the trial's ability to detect small or moderate sized effects.[32] In this context, the primary analysis result supports the conclusion that the programme does have a small positive effect on Year 2 pupils' mathematical attainment.

While this remote training model appears to have been broadly acceptable to schools, overall fidelity does not appear to have improved compared to the previous trials. The key fidelity issues reported in the previous trials related to teachers skipping some of the programme material in order to fit the units into an hour, limited use of the computer games due to staffing, technology or time constraints, and either not splitting into groups or doing so in a static manner based on prior attainment (Worth *et al.*, 2015; Stokes *et al.*, 2018). As described above, many of the same issues have been identified here. As most of these challenges relate to structural challenges faced by schools, it would not have been anticipated that changing the training model alone would have increased fidelity in these respects. A notable exception to this is the average number of computer games played, which increased in this trial compared to the previous one. However, as this increase is accompanied by an increase in school access to the necessary technology, this again appears to be the result of a broader shift in the structural constraints within teachers are working, rather than from any changes to the programme itself.

While it is likely that the programme does indeed have an effect, the effect size observed suggests that the mode of training does little to alter the magnitude of it. This may be in part due to low fidelity in terms of teachers and TAs completing the training together, which resulted in teachers completing the training on their own and often in their own time, which is likely to have resulted in lower levels of engagement with the material than in-person training, or for intended practice for this model. As discussed above, the amount of content and materials the programme requires still appears to be too great for teachers to cover in its entirety.

In the first effectiveness trial, it was hypothesised that the train-the-trainer model diluted the effect. However, as this caveat does not apply to the new training model, it is possible that the discrepancy is simply the result of a very small sample size at the efficacy stage (17 schools in the Mathematical Reasoning intervention group), leading to a relatively wide 95% CI of 0.02 to 0.37. Moreover, there is no evidence to suggest that the programme was delivered with higher levels of fidelity in the efficacy trial.

A difference between the two effectiveness trials is, however, the result for FSM-eligible pupils, with a slightly larger result found in this, second, effectiveness trial. It is not clear what has driven this difference, though there may be some relationship with another result that suggested the Mathematical Reasoning programme had greater benefits for pupils with lower prior attainment, as FSM status is known to be correlated with lower prior attainment. There was little evidence to contextualise this particular result from the IPE, though, as teachers felt that the programme was better suited for higher attaining learners and less suited for lower attaining pupils and those with SEND. While there may be some differences in terms of which pupils the impact analysis classed as 'lower attaining' and how teachers responding to the IPE grouped these pupils (e.g. case study teachers referred to those working below the level needed to be able to access the intervention), we

[32] This is not to say that the observed p-value should somehow be lowered to adjust for the trial being underpowered, just that if the real effect size is close to 0.117, a p-value above 0.05 will be more likely than the conventional probability of 0.2.

note that these pupils may also have been underrepresented at endpoint, as pupils with lower baseline scores and those with SEND were more likely to have missing endpoint scores (see Appendix G in Further Appendices).

There was some evidence to suggest that the computer game element of the programme was also beneficial, though inconsistency in terms of which pupils were allocated to the computer game groups within classes and varying approaches to how pupils approached playing the computer games suggests that further evidence may be needed to fully understand the role of this element of the programme. The IPE suggested that—in at least some schools—teachers were more likely to complete the teacher-led group work with the pupils that needed additional support, suggesting that the more able pupils were more likely to be playing the computer games. At the very least, IT access was not identified as a barrier in this second effectiveness trial, suggesting that schools have a much higher level of IT resources now compared to the previous trial, removing some of the barriers to implementation seen previously.

As with all trials that recruit schools, it may be expected that the schools that agreed to take part may be more interested in Mathematical Reasoning approaches, or may have identified this as an area that they wanted to improve—however, there is no data to compare whether the schools in this trial were different to the wider population of primary schools in this respect. It is possible to determine that the practice of control schools during the trial was similar to BAU across both groups at baseline (i.e. before the trial period), which suggests that the schools in the control group did not make significant changes to their practice in relation to Mathematical Reasoning training and practice in response to being allocated to the control condition.

The previous effectiveness trial reported 'a mixed opinion' among teachers about the impact on the programme on their practice (Stokes *et al.*, 2018). This second effectiveness trial found that self-reported teacher understanding and confidence in teaching mathematical reasoning increased significantly more among teachers in the intervention group compared to the control group. In both the first and second effectiveness trials, most teachers perceived the programme to have had a moderate impact on pupils' maths skills, supporting the impact findings. Despite these positive perceptions of outcomes, only a third of teachers indicated that they would continue with the programme in future. The main reasons for this were the programme not fitting with the school's maths priorities and staffing capacity challenges. However, nearly all of the case study teachers were planning to take some elements of the programme into their teaching and share these learnings with other staff members. This provides further evidence for the programme achieving its objectives in relation to CPD, which the delivery team identified as of equal importance to the short-term pupil outcomes. Nonetheless, these findings suggest that further modifications to the programme, including in relation to staffing requirements and unit length, may be required for wider take-up and longer-term scale-up to be successful.

Limitations and lessons learned

We observed that FSM-eligible pupils, urban schools, and LA-maintained schools were overrepresented in the trial sample compared to the national population. The overrepresentation of FSM-eligible pupils is perhaps of most concern, given that the effect size seen in the FSM-eligible subgroup was larger than for the primary analysis. A higher proportion of FSM-eligible pupils in the trial sample could mean the impact of Mathematical Reasoning in the national population is smaller on average than the impact seen in the primary analysis. That said, the difference in FSM proportions between the trial and national figures was moderate (28.6% vs 24.6%), so the impact on the national population would be reduced only slightly. Overall, the differences in pupil- and school-level characteristics between the trial sample and national population do not suggest a serious threat to the external validity of the primary analysis result.

The compliance analysis noted a high level of intervention take-up, with 91% of analysed intervention pupils attending ten or more units. However, this analysis was limited by the fact that 19% of intervention schools did not return the Session Delivery Log. Average compliance levels are almost certainly much lower among these schools, as they include 11 schools known to have withdrawn from the intervention. Compliance was therefore high only among pupils at schools that did not withdraw from the intervention. All but two of the withdrawn schools agreed to administer the PTM7 at endpoint, so the primary analysis included nine withdrawn schools. This is likely to have slightly reduced the primary analysis estimate, which could be viewed as a limitation if the 'true' impact of the intervention—that is, when delivered completely under ideal conditions—was of interest. As this is an effectiveness trial, we take the view that the reduction is appropriate: the challenges reported by withdrawn schools (e.g. lack of staff capacity) are also likely to arise when the programme is implemented in the real world.

Challenges relating to TA availability to support the sessions were found to be common and affected whether schools could implement the programme as intended. While most teachers delivered the intended number of sessions, the case study data suggests that some programme content may still have been missed. These indications of low fidelity mean that it is possible the effect size found for this trial is lower than it would be were the programme to be delivered fully as intended. However, as an effectiveness trial, these lower levels of fidelity are the consequence of real-world conditions (TA availability, time taken to deliver the material, and so on). This means that the effect size observed for this trial can still be considered a meaningful indication of the likely impact were the programme in its current iteration to be rolled out more widely.

Three of the four PTM7 subscales used as secondary outcomes had several inter-related issues: histograms suggested all three had floor or ceiling effects, while two of these three had a low range of possible values and a low Cronbach's Alpha. Only the 'mathematical reasoning' subscale was exempt from these issues, making the isolated positive impact seen for this subscale hard to interpret. Given that some of these issues were anticipated—floor and ceiling effects were observed in the previous effectiveness trial of Mathematical Reasoning—it might in retrospect have been better to include only the 'mathematical reasoning' subscale as a secondary outcome, while labelling the other three subscale results as exploratory or removing them entirely. Furthermore, there was some evidence that all four PTM7 subscales lacked discriminant validity in the trial sample. This means it is possible that the positive result for the 'mathematical reasoning' subscale is simply due to its correlation with total PTM7 score: it may be measuring general mathematical ability rather than a distinct construct 'mathematical reasoning'. The effect size for the 'mathematical reasoning' subscale (0.105) was higher than for the primary analysis (0.093), but only slightly so.

In the computer games impact analysis, we found that playing more games (especially high-scoring games) and playing games from more subject topics were both associated with a higher PTM7 score. However, this result could have been biased by factors that are correlated with both computer game use and PTM7 score, including receiving more Mathematical Reasoning content generally (e.g. attending more sessions) and enthusiasm for playing the games. These factors are likely to cause positive rather than negative bias, which means that our results could overestimate the true causal benefit a pupil received from playing more games.

There were further limitations to the computer game analysis, stemming from the data itself. The analysis included some pupils for whom a correct match between their computer game and PTM7 data was not confirmed by schools. A sensitivity check that excluded these pupils increased the estimated association of more games played and of more subject topics with the PTM7 score, suggesting these estimates may have been attenuated by some incorrect matches. We also took the step of excluding 11 pupils who played over 150 games from modelling, as we expected any benefit from playing more games to plateau (little or no further benefit) in this high range, which could influence the overall linear estimate of benefit per game played. A sensitivity check that added the 11 pupils to modelling confirmed our expectation, as the estimated increase in PTM7 score per game played dropped substantially, although it remained positive. This raises the question of whether the relationship between number of games played and PTM7 score is linear for pupils that played less than 150 games. If not, excluding pupils using a different threshold (e.g. over 100 games played) might have produced yet another different result. Adding non-linear terms to the regressions could have been attempted to explicitly model a non-linear relationship, but the results would probably have been inaccurate due to the scarcity of pupils who played a very high number of games.

Evidence from the IPE indicates that levels of understanding and confidence in relation to mathematical reasoning improved across both groups, but to a significantly greater extent in the intervention teachers. The slight increase in the understanding levels of control group teachers may be related to CPD received during the trial, however the amount of training received was consistent with usual practice reported at baseline. While it is hypothesised in the Theory of Change that increased teacher confidence and understanding of mathematical reasoning will result in more pupils receiving instruction in this area, we did not investigate in this trial whether these teacher outcomes mediate the pupil outcomes assessed. Overall, there is no clear indication that the effect observed in this trial represents an underestimate for reasons of contamination.

Control schools were permitted to deliver other mathematical interventions, including those relating to mathematical reasoning, over the trial period as part of usual practice. There was no difference in the proportion of control schools delivering mathematical reasoning interventions between baseline and endpoint. While a small number of control teachers completed CPD relating to mathematical reasoning over the trial period, as outlined above, a similar proportion reported having completed CPD on this topic in the 12 months prior to the trial, suggesting a continuation of usual practice in control schools over the trial period.

There was no indication from the case study data that intervention schools had changed their practice in relation to the teaching of mathematical reasoning during the trial period beyond implementation of the Mathematical Reasoning programme. However, as this data was not systematically collected across the intervention sample, there is still a small risk that the effect size identified here includes the impact of other measures implemented by intervention schools in addition to the programme, although the randomised nature of the trial means the likelihood of this is minimal.

The IPE findings are based on the responses of individuals who agreed to participate in additional data collection activities. As a result, they may be subject to response bias, where for example, participants with a more positive perception of the programme may have been particularly likely to respond to the endpoint surveys. We did not intend for the IPE data to be representative of the full trial population, but this does mean that the IPE may not have accurately captured the experiences of all participants. In addition, the training completion data may not fully represent the extent to which each member of staff engaged with the content of each module, as this is separate from whether or not they responded to the activities, which was the criterion used for determining module completion.

Future research and publications

With the second effectiveness trial producing very similar results for the primary outcome to the first, we would argue that there is clear evidence of positive impact, including for pupils eligible for FSM, and consequently there is no need for further effectiveness trials of this programme. Future research could take a more exploratory focus to understand how the programme could be adapted to increase the level of acceptability and feasibility of its delivery in schools, on the basis of the IPE findings that have largely aligned across the three trials, without compromising the core components that these trials have identified. Future research could also look to understand how the new training model could be further developed to improve programme fidelity, including in terms of consistency around the use of manipulatives and asking pupils about their reasoning.

References

Angrist, J.D. and Imbens, G.W. (1995) 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association*, 90(430), pp. 431–442. Available at: https://doi.org/10.1080/01621459.1995.10476535.

Barrett, P. (2007) 'Structural equation modelling: adjudging model fit', *Personality and Individual Differences*, 42(5), pp. 815–824. Available at: https://doi.org/10.1016/j.paid.2006.09.018.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software*, 67, pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Bishenden, O. (2023) 'RE: Progress Test in Maths in primary schools'.

Bloom, H.S., Richburg-Hayes, L. and Black, A.R. (2007) 'Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions', *Educational Eval;uation and Policy Analysis*, 29(1). Available at: https://journals.sagepub.com/doi/abs/10.3102/0162373707299550.

Bulus, M., Dong, N., Kelcey, B. and Spybrook, J. (2021) 'PowerUpR: power analysis tools for multilevel randomized experiments'. Available at: https://CRAN.R-project.org/package=PowerUpR.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) 'mice: Multivariate Imputation by Chained Equations in R', *Journal of Statistical Software*, 45, pp. 1–67. Available at: https://doi.org/10.18637/jss.v045.i03.

Ching, B.H.-H., Kong, K.H.C., Wu, H.X. and Chen, T.T. (2020) 'Examining the reciprocal relations of mathematics anxiety to quantitative reasoning and number knowledge in Chinese children', *Contemporary Educational Psychology*, 63, p. 101919. Available at: https://doi.org/10.1016/j.cedpsych.2020.101919.

Ching, B.H.-H. and Nunes, T. (2017) 'The importance of additive reasoning in children's mathematical achievement: a longitudinal study', *Journal of Educational Psychology*, 109(4), pp. 477–508. Available at: https://doi.org/10.1037/edu0000154.

Çiftçi, S.K. and Yildiz, P. (2019) 'The effect of self-confidence on mathematics achievement: the metaanalysis of Trends in International Mathematics and Science Study (TIMSS)', *International Journal of Instruction*, 12(2), pp. 683–694. Available at: https://doi.org/10.29333/iji.2019.12243a.

Data Protection Act 2018, C. 12 (2018). Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents/.

Department for Education (2025) *Schools, pupils and their characteristics - Academic year 2024/25*, *gov.uk*. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2024-25.

Department for Education (2026a) *Get Information about Schools*, *gov.uk*. Available at: https://get-information-schools.service.gov.uk/.

Department for Education (2026b) *Special educational needs in England - Academic year 2024/25*, *gov.uk*. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/special-educational-needs-in-england/2024-25.

Education Endowment Foundation (2022) *Statistical analysis guidance for EEF evaluations*. London. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1699621596 (Accessed: 27 January 2025).

Education Endowment Foundation (2023) 'Cost evaluation guidance for EEF evaluations'. EEF. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/Cost-Evaluation-Guidance-Feb_2023.pdf?v=1765786924.

Education Endowment Foundation (2026) *Teaching and Learning Toolkit*. Available at: https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit.

Flemons, L., Smith, A., Morton, C. and Poet, H. (2024) *Effectiveness trial of Mathematical Reasoning - Evaluation Protocol*. London. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/mathematical_reasoning_-_evaluation_protocol.pdf?v=1737980741 (Accessed: 27 January 2025).

Fox, J., Kleiber, C., Zeileis, A. and Kuschnig, N. (2021) ‘ivreg: instrumental-variables regression by “2SLS”, “2SM”, or “2SMM”, with diagnostics’. Available at: https://CRAN.R-project.org/package=ivreg (Accessed: 27 January 2025).

Gerber, A.S. and Green, D.P. (2012) *Field experiments: design, analysis and interpretation*. London: W. W. Norton & Company.

Hair, J.F., Anderson, R., Tatham, R. and Black, W. (1998) *Multivariate data analysis*. Fifth. Hoboken: New Jersey: Prentice Hall. Available at: https://books.google.co.uk/books?id=-ZGsQgAACAAJ (Accessed: 29 January 2026).

Hedges, L.V. (2007) ‘Effect Sizes in Cluster-Randomized Designs’, *Journal of Educational and Behavioral Statistics*, 32(4), pp. 341–370. Available at: https://doi.org/10.3102/1076998606298043.

Hewitt, C.E. and Torgerson, D.J. (2006) ‘Is restricted randomisation necessary?’, *BMJ*, 332(7556), pp. 1506–1508. Available at: https://doi.org/10.1136/bmj.332.7556.1506.

Hu, L. and Bentler, P.M. (1999) ‘Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives’, *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), pp. 1–55. Available at: https://doi.org/10.1080/10705519909540118.

Jordan, N.C., Devlin, B.L. and Botello, M. (2022) ‘Core foundations of early mathematics: refining the number sense framework’, *Current Opinion in Behavioral Sciences*, 46, p. 101181. Available at: https://doi.org/10.1016/j.cobeha.2022.101181.

Lazarides, R. and Warner, L.M. (2020) ‘Teacher Self-Efficacy’, *Oxford Research Encyclopedia of Education*. Available at: https://doi.org/10.1093/acrefore/9780190264093.013.890.

Li, C.-H. (2016) ‘Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares’, *Behavior Research Methods*, 48(3), pp. 936–949. Available at: https://doi.org/10.3758/s13428-015-0619-7.

Mardia, K.V. (1970) ‘Measures of multivariate skewness and kurtosis with applications’, *Biometrika*, 57(3), pp. 519–530. Available at: https://doi.org/10.1093/biomet/57.3.519.

Marsh, H.W., Hau, K.-T. and Wen, Z. (2004) ‘In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler’s (1999) findings’, *Structural Equation Modeling: A Multidisciplinary Journal*, 11(3), pp. 320–341. Available at: https://doi.org/10.1207/s15328007sem1103_2.

McMullin, C. (2023) ‘Transcription and qualitative methods: implications for third sector research’, *Voluntas*, 34, pp. 140–153. Available at: https://doi.org/10.1007/s11266-021-00400-3.

Morton, C. and Smith, A. (2025) ‘Effectiveness Trial of Mathematical Reasoning - Statistical Analysis Plan’. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/mathematical_reasoning_-_statistical_analysis_plan_-_v.1.0.0.pdf?v=1769170402.

Mujtaba, T. and Sheldrake, R. (2023) *Evaluation of the online professional development for delivering the Mathematical Reasoning Programme*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Mathematical-Reasoning_Study-Plan_v.1.1.pdf?v=1710936061usb.cloudfront.net) (Accessed: 21 June 2024).

Myers, N.D., Ahn, S. and Jin, Y. (2011) ‘Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: a monte carlo approach’, *Research Quarterly for Exercise and Sport*, 82(3), pp. 412–423. Available at: https://doi.org/10.1080/02701367.2011.10599773.

Nunes, T., Bryant, P., Barros, R. and Sylva, K. (2011) 'The relative importance of two different mathematical abilities to mathematical achievement', *The British Journal of Educational Psychology*, 82(1), pp. 136–56. Available at: https://doi.org/10.1111/j.2044-8279.2011.02033.x.

Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A. and Carraher, J. (2007) 'The contribution of logical reasoning to the learning of mathematics in primary school', *British Journal of Developmental Psychology*, 25(1), pp. 147–166. Available at: https://doi.org/10.1348/026151006X153127.

Parchment, A., Lawrence, W., Rahman, E., Townsend, N., Wainwright, E. and Wainwright, D. (2025) 'Increasing confidence and competence in supporting behaviour change in physiotherapy practice using Making Every Contact Count Healthy Conversation Skills: a before and after evaluation', *BMC Health Services Research*, 25(1), p. 797. Available at: https://doi.org/10.1186/s12913-025-12513-2.

Putwain, D.W., Becker, S., Symes, W. and Pekrun, R. (2018) 'Reciprocal relations between students' academic enjoyment, boredom, and achievement over time', *Learning and Instruction*, 54, pp. 73–81. Available at: https://doi.org/10.1016/j.learninstruc.2017.08.004.

R Core Team (2026) 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Regulation (EU) 2016/679 of the European Parliament and of the Council (2016). Available at: https://www.legislation.gov.uk/eur/2016/679.

Sheldrake, R., Mujtaba, T., Markwick, A., McCory, A., Geraniou, E. and Gokpinar, T. (2024) *Mathematical Reasoning - Pilot report*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Mathematical_reasoning_evaluation_report.pdf?v=1779374500.

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Matur, T., Higgins, S., Einbeck, J., Culliney, M. and Demack, S. (2023) *Improving power calculations in educational trials*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/Work_Package_2023-WP6_18_09_2023_FINAL.pdf?v=1696410358 (Accessed: 21 June 2024).

Stokes, L., Hudson-Sharp, N., Dorsett, R., Rolfe, H., Anders, J., George, A., Buzzeo, J. and Munro-Lott, N. (2018) *Mathematical reasoning: evaluation report and executive summary*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Mathematical_Reasoning.pdf?v=1696414461 (Accessed: 21 June 2024).

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D. and van den Brand, T. (no date) *Smoothed Conditional Means*, *ggplot2*. Available at: https://ggplot2.tidyverse.org/reference/geom_smooth.html.

Worth, J., Sizmur, J., Ager, R. and Styles, B. (2015) *Improving numeracy and literacy: evaluation report and executive summary*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Oxford_Numeracy_and_Literacy.pdf?v=1696414466 (Accessed: 21 June 2024).

Zeileis, A. (2006) 'Object-oriented computation of sandwich estimators', *Journal of Statistical Software*, 16(9), pp. 1–16. Available at: https://doi.org/10.18637/jss.v016.i09.

Zeileis, A., Köll, S. and Graham, N. (2020) 'Various versatile variances: an object-oriented implementation of clustered covariances in R', *Journal of Statistical Software*, 95(1), pp. 1–36. Available at: https://doi.org/10.18637/jss.v095.i01.

Appendix A: EEF cost rating

Appendix A Table 1: Cost rating

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

Appendix B: Security classification of trial findings

OUTCOME: *GL Assessment Progress Test in Maths (PTM) 7*

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%	4				4
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%			Adjustment for threats to internal validity **[0]**		
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%					
0	No comparator	>=0.6	>50%					

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Low	Well-conducted RCT with good balance across most pupil characteristics. While there are some minor differences (e.g. EAL and some school-level characteristics), these do not appear to be sufficient to introduce meaningful confounding.
Threat 2: Concurrent Interventions	Low	While there is some evidence of differential prevalence in concurrent support (76% vs 67%), this is not judged sufficient to materially threaten the trial's overall internal validity under the padlock framework. However, given the targeting of these interventions toward additional support pupils, some caution remains in interpreting subgroup effects.
Threat 3: Experimental effects	Low	While exposure to MR-related training and approaches appears broadly similar across groups, there is some evidence of increased use of relevant practices in control schools (e.g. 11–14% increase in problem-based learning and real-life examples). This likely reflects ongoing professional development but may introduce a limited risk of overlap between groups.
Threat 4: Implementation fidelity	Moderate	The overall delivery was strong in terms of training completion, materials, and pupil participation. However, the examples provided (e.g. skipping content, variation in TA involvement, inconsistent use of key components such as manipulatives and computer games) suggest that some core elements were not implemented consistently. While this reflects the realities of delivery across different school contexts, it indicates that the programme may be challenging to implement as intended in practice. This introduces some uncertainty about whether the evaluation tested a sufficiently consistent version of the intervention.

Threat 5: Missing Data	Moderate	The multiple imputation analysis produced an estimated impact of 0.72 (95% CI: -0.14 to 1.59), representing a reduction of around 14% compared to the primary analysis estimate. This suggests that, under the assumptions explored, missing data may have introduced some downward bias, but the magnitude of this bias appears limited. However, as these analyses rely on untestable assumptions about the missing data mechanism, they cannot fully rule out bias. The confidence interval from the imputed analysis includes both small negative and positive values, indicating increased uncertainty around both the presence and size of the effect. Overall, the findings suggest that the results are relatively stable under the assumptions tested but should be interpreted with appropriate caution.
Threat 6: Measurement of Outcomes	Low	The outcome assessment appears robust, with independent administration, blinding, and use of a well-established measure. We consider that the risk to validity is low.
Threat 7: Selective reporting	Low	The trial appears to have been reported transparently, with analyses clearly specified in advance and all key research questions addressed. We agree that the risk here is low.

- **Initial padlock score:** [NUMBER] Padlocks - Insert Description
- **Reason for adjustment for threats to validity**: [NUMBER] Padlocks - Insert Description
- **Final padlock score:** initial score adjusted for threats to validity = 4 Padlocks.

Appendix C: Changes since the previous evaluation

Appendix C Table 1: Changes since the previous evaluation

	Feature	Efficacy to effectiveness stage	First effectiveness trial to second effectiveness trial
Programme	Programme content	No change	No change
	Delivery model	One day of in-person training and coaching delivered by Oxford University developers changed to a train-the-trainer model. Oxford University trained Work Group Leaders on two days of in-person training, who then provided one in-person training day and one coaching day to teachers	Change from an in-person train-the-trainer model to one and a half days of training made up of an online course offered by the team of developers, three webinars, an online forum and support via email, if required, from TLs. TLs receive three days of training from the Oxford University team, in a blended format (some online elements and some in-person training)
	Programme duration	Schools were advised to deliver the programme's ten units over 12–15 lessons	No change
Evaluation	Eligibility criteria	No change	Change from only schools within the eight Maths Hubs to all English state primary schools
	Level of randomisation	Randomisation within blocks was defined based on hubs, the proportion of children eligible for FSM, and prior attainment at Key Stage 1	Simple randomisation without stratification
	Outcomes and baseline	The baseline and endpoint have been changed to an updated version of Progress in Maths 7 (PiM 7)—Progress Test in Maths (PTM) 7	No change
	Control condition	No change	No change

Appendix D: Effect size estimation

Appendix D Table 1: Effect size estimation for all pupils

Outcome	Unadjusted differences in means	Adjusted differences in means	Variance components obtained from a model with no predictors		
			Between-school variance σ_B^2	Within-school variance σ_W^2	Effect size denominator $\sqrt{\sigma_B^2 + \sigma_W^2}$
PTM7 total score	0.73	0.84	17.43	64.91	9.07
Fluency in facts and procedures score	0.07	0.09	0.35	1.64	1.41
Fluency in conceptual understanding score	0.25	0.21	2.15	8.33	3.24
Problem-solving score	-0.03	-0.02	0.14	0.85	0.99
Mathematical reasoning score	0.44	0.50	4.05	18.26	4.72

Appendix D Table 2: Effect size estimation for the subgroup of FSM-eligible pupils

Outcome	Unadjusted differences in means	Adjusted differences in means	Variance components obtained from a model with no predictors		
			Between-school variance σ_B^2	Within-school variance σ_W^2	Effect size denominator $\sqrt{\sigma_B^2 + \sigma_W^2}$
PTM7 total score	1.64	1.29	17.04	67.79	9.21
Fluency in facts and procedures score	0.16	0.15	0.34	1.98	1.52
Fluency in conceptual understanding score	0.55	0.37	2.26	8.92	3.34
Problem-solving score	0.02	-0.03	0.12	0.65	0.88
Mathematical reasoning score	0.91	0.88	3.71	19.09	4.78

Further Appendices

Appendices E – P can be found in an accompanying document named '*Further Appendices'* published on the project page.

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