



Education
Endowment
Foundation

Mathematical Reasoning

Evaluation report and executive summary

December 2018

Independent evaluators:

Lucy Stokes, Nathan Hudson-Sharp, Richard Dorsett, Heather Rolfe,
Jake Anders, Anitha George, Jonathan Buzzeo, Naomi Munro-Lott



NatCen
Social Research



The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education. This evaluation was co-funded by the Worshipful Company of Actuaries

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.



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

Danielle Mason
Head of Research

Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP
p: 020 7802 1679
e: danielle.mason@eefoundation.org.uk
w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a NIESR-led team in partnership with NatCen who were responsible for the administration of assessments. The evaluation team included Lucy Stokes, Nathan Hudson-Sharp, Richard Dorsett, Heather Rolfe, Jake Anders, Anitha George, Jonathan Buzzeo and Naomi Munro-Lott. The NatCen team included Rakhee Patel, Lydia Marshall, Hannah Piggott, and Katie Drysdale.

The lead evaluator was Lucy Stokes.

Contact details:

Lucy Stokes
National Institute of Economic and Social Research
2 Dean Trench Street
Smith Square
London
SW1P 3HE

Tel: 020 7222 7665

Email: l.stokes@niesr.ac.uk

Contents

| | |
|--|-----------|
| Executive summary | 4 |
| Introduction | 6 |
| Methods | 11 |
| Impact evaluation | 18 |
| Implementation and process evaluation | 35 |
| Conclusion | 51 |
| References | 54 |
| Appendix A: EEF cost rating | 55 |
| Appendix B: Security classification of trial findings | 56 |
| Appendix C: Recruitment materials | 57 |
| Appendix D: Randomisation code | 62 |
| Appendix E: Histograms of pre-test scores | 64 |
| Appendix F: Analysis code | 65 |
| Appendix G: Histograms of PTM raw score and subscores | 66 |
| Appendix H: Interacting treatment with prior attainment | 69 |

Executive summary

The project

The Mathematical Reasoning programme aims to improve the mathematical attainment of pupils in Year 2 by developing their understanding of the logical principles underlying maths. The programme was previously tested in an EEF-funded efficacy trial (**Improving Numeracy and Literacy in Key Stage 1**) which suggested that it had a positive impact. The efficacy trial examined the programme under developer-led conditions. This report describes a follow-up effectiveness trial which examined the impact of the programme under everyday conditions in a large number of schools and with less involvement from the original developer.

Mathematical Reasoning lessons focus on developing pupils' understanding of number and quantitative reasoning. They cover principles such as place value and the inverse relation between addition and subtraction. The programme consists of ten units delivered to pupils by their teachers as part of their usual mathematics lessons. It is designed to be taught over a 12- to 15-week period, with each unit taking approximately one hour. Learning is supported by online games, which can be used by pupils both at school and at home. The intervention was originally developed by a team at the University of Oxford, led by Professor Terezinha Nunes and Professor Peter Bryant. The National Centre for Excellence in the Teaching of Mathematics (NCETM) contributed to the development of the training model used in this trial and coordinated the delivery of the training through the network of Maths Hubs (partnerships of schools created to lead improvements to maths education).

In this trial, the teacher training was delivered using a 'train-the-trainers' model through eight Maths Hubs. Each Maths Hub was asked to recruit two 'Work Group Leads'. The University of Oxford programme developers trained these Work Group Leads who then trained the teachers in participating schools to deliver the programme. To prepare them to train the teachers, Work Group Leads received an initial day of training, used the materials in their own teaching, and then received a further two days' training. Teachers delivering the programme then received one day of training from a Work Group Lead as well as a visit from the Work Group Lead during programme delivery. They were also able to seek additional support directly from the Work Group Lead or ask questions through an online Maths Hub community.

The impact of the programme on maths attainment was evaluated using a randomised controlled trial involving 160 schools. Schools were randomly allocated either to receive Mathematical Reasoning or to be in the control group, the latter having the opportunity to take part in the programme in the following school year. A process evaluation used observations of training sessions, teacher interviews, lesson observations, and an online survey of treatment and control schools to examine implementation and the factors influencing impact. The trial began in August 2015 and analysis and reporting of the trial completed in December 2018. The project was co-funded by the Worshipful Company of Actuaries.

Key conclusions

1. Pupils who received Mathematical Reasoning made the equivalent of one additional month's progress in maths, on average, compared to children who did not. This result has high security.
2. Among pupils eligible for free school meals, those who received Mathematical Reasoning made an average of one additional month's progress compared to those who did not. This result may have lower security than the overall finding because of the smaller number of pupils.
3. There was some evidence that the programme also had a positive impact on mathematical reasoning.

4. The intervention was generally well received by schools. Teachers reported positive experiences with the training and materials, and were positive about the programme's focus on fundamental mathematical principles.
5. The process evaluation found that there was some variation in how schools implemented aspects of the programme, particularly in relation to the use of the online games.

EEF security rating

These findings have a high security rating. The trial was a well-designed and well-powered randomised controlled trial. The pupils in mathematical reasoning classes were similar to those in the comparison classes in terms of prior attainment. However, 14% of the pupils who started the trial were not included in the final analysis. The main causes of pupils not being included were schools dropping out of the trial and pupils moving school between randomisation and the post-test.

Additional findings

Exploratory analyses investigated the impact of the programme on the different components of the Progress Test in Maths score used to measure pupils' maths attainment. This suggests there may have been a positive impact on some sub-components of mathematical knowledge and understanding, notably mathematical reasoning.


The process evaluation revealed that the intervention was generally well received by schools, although the extent to which the programme was implemented as intended varied. This was partly because some schools did not have access to the necessary IT equipment or teaching assistant (TA) support.

The previous efficacy trial estimated that the programme had a larger positive impact than in this effectiveness trial. There are several differences between the two trials which may explain the smaller effect size in the effectiveness trial. The introduction of the train-the-trainers model might be expected to reduce fidelity because the programme developers are no longer directly training teachers. Also, although a precise comparison is difficult, there was evidence that control group schools in the effectiveness trial were more likely than in the efficacy trial to use other materials or resources to support children's reasoning in maths. This could have diluted the impact of the programme.

Cost

The cost per pupil per year, averaged over three years, is estimated to be £8. Access to computers or tablets for pupils to play the online games is not included in the cost, but is an important prerequisite for the programme. In terms of staff time, teachers are required to attend one day of training. While the programme is delivered as part of normal maths lessons, teachers had to spend some time preparing lessons and attending the support visit from the Work Group Lead.

Table 1: 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 | 0.08 (-0.03, 0.18) | 1 |  | 6,353 | 0.156 | £££££ |
| Maths FSM pupils | 0.09 (-0.07, 0.25) | 1 | N/A | 1,323 | 0.288 | £££££ |

Introduction

Background evidence

The Mathematical Reasoning programme develops children's understanding of the logical principles underlying mathematics. Developed by Professor Terezinha Nunes and Professor Peter Bryant at the University of Oxford, the programme focuses on quantitative reasoning, that is, the ability to understand the relationships between numbers and to use them to solve problems. It is designed for children in Year 2.

Previous research by the developers of the Mathematical Reasoning programme has demonstrated the importance of logical reasoning for mathematical understanding (Nunes, Bryant, et al., 2007), and that both number sense and quantitative reasoning are key predictors for how well children perform in KS2 and KS3 mathematics (Nunes et al., 2011). The programme focuses on developing these two abilities, which although interdependent, can be promoted through different types of activities (Nunes and Bryant, 2015). A previous EEF efficacy trial of the Mathematical Reasoning programme has shown the intervention to have a positive impact on mathematical attainment, with an effect size of 0.2 (Worth et al., 2015). Such efficacy trials are designed to test whether an intervention can work under ideal conditions, with considerable support from the original developer of the intervention. This effectiveness trial provides the opportunity to test whether the intervention can work at scale. As such it is particularly important for the evaluation to explore those factors which have changed in order to implement the intervention on a larger scale, for example, whether the introduction of the 'train the trainers' model has consequences for the effectiveness of the programme.

Intervention

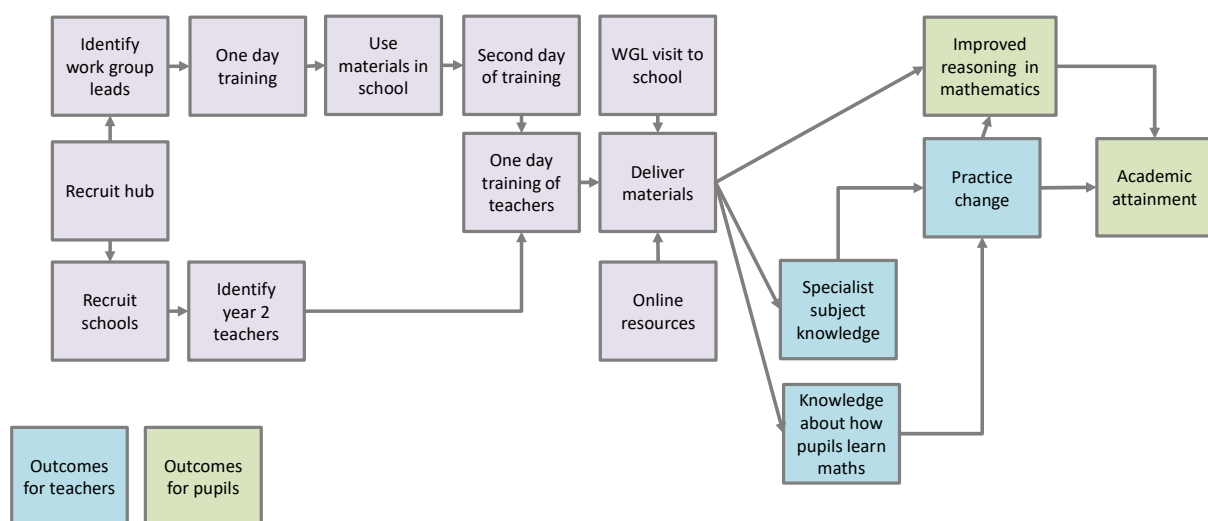
The Mathematical Reasoning programme consists of ten teaching units, designed to last approximately 12 to 15 weeks, with children receiving approximately one hour of instruction per week as part of their normal mathematics lessons. This estimation is based on the recommendation given to teachers implementing the programme that if they are unable to complete a unit in one session, they should continue from where they left off next time. The programme has two strands. The first focuses on promoting the children's understanding of numbers or number sense and the second focuses on quantitative reasoning or problem solving. The number sense strand included activities looking at additive composition of number and place-value, and inverse relation between addition and subtraction. The quantitative reasoning strand involved thinking about the different relations that can be established between quantities, including part-whole or additive relations, and one-to-many correspondence or multiplicative relations. The programme does not introduce new subject content—it is based on the national curriculum—but instead the focus of the programme is on reasoning and understanding. As noted above, the intervention does not involve additional maths teaching—the lessons from the programme are intended to replace the maths lessons that the teachers would otherwise be giving.

Each new concept is introduced through teacher-led activities during which the children sit on the carpet and use manipulatives to enact a story problem. The demonstrations with objects are followed by teacher-led PowerPoint presentations in which the children solve further problems. After the whole class activity the class is separated into two groups labelled Level 1 (L1) and Level 2 (L2). Level 1 activities are for children who need some more time and support with learning new concepts. Level 2 activities are for children who are quicker to grasp new ideas and can be helped to extend their thinking. The intention is that children are allocated to these groups depending on how they react to the content of the lesson being taught (that is, it is not the case that children are assigned to each group at a particular stage). During this time one group (L1 or L2) focuses on consolidation of what has been taught in the lesson through online games. The other group has more intensive, differentiated teaching with the teacher which might be an extension to greater depth, some extra support with a concept, or some pre-teaching to support the next lesson. The groups alternate each week, meaning over the ten units each

group has five turns each with the games. The games are linked to the resources and designed to build on the reasoning skills developed in the classroom session. Due to the use of online games, access to ICT facilities was identified as an essential prerequisite for participation in this trial. Having support from teaching assistants to facilitate classroom differentiation was also identified as preferable. Figure 1 presents a logic model for the intervention, showing how the programme aims to ultimately improve children’s maths attainment both through a direct impact on their mathematical reasoning abilities as well as through bringing about change in teachers’ practice.

In this effectiveness trial, delivery operated through a scalable train-the-trainers model (also shown in Figure 1). Implementing schools were trained by Work Group Leads (WGLs), who were themselves trained by the programme developers at the University of Oxford and supported by the National Centre for Excellence in the Teaching of Mathematics (NCETM) through the Maths Hub programme. This is as opposed to the efficacy trial (Worth et al., 2015) where implementing schools were trained directly by the programme developers. Eight Maths Hubs were identified by the NCETM to take part in the trial, with Work Group Leads then identified by the participating Maths Hubs. The Maths Hub Network is a Department for Education (DfE) funded initiative, coordinated by the NCETM, to support schools in leading improvements in maths education; there are currently 35 Maths Hubs in England. Each hub is a partnership, led locally by an outstanding school or college and bringing together maths education professionals to develop and spread best practice.¹ The NCETM advertised the opportunity to take part in the trial to Maths Hubs that could then express interest in taking part. While twelve hubs initially expressed interest, one withdrew because of the time commitment required and a further three offered to step aside once it was clear a sufficient number of hubs were interested. This resulted in eight hubs to participate in the trial, as originally intended. Each hub was asked to recruit two WGLs, generally Year 2 teachers or teachers able to use the materials with a Year 2 group. Involvement from a local higher education institution was also encouraged to provide the opportunity for WGLs to benefit from ongoing discussions about the pedagogy and research base for the project. In some cases the WGLs were themselves based in higher education institutions. In order to facilitate the WGL recruitment process, the NCETM provided hubs with an information sheet about the project and the role of the WGL along with an expression of interest form for potential WGLs to return, including a statement about why they were interested in the role and their previous relevant experience as well as a statement from their headteacher confirming the school’s commitment to the WGL taking on the role.

Figure 1: Logic model for the intervention



The initial training for Work Group Leads took place in February 2016. The initial training event was hosted by the project team at the University of Oxford and the NCETM at the Department for

¹ <http://www.mathshubs.org.uk/>

Education, University of Oxford. During this event, WGLs were provided with background research, were given the opportunity to become familiar with the materials and how to use them, and were given detailed information regarding their role as WGLs and the project (although it was planned just to have one initial training event, in practice a second event was held in April 2016 for a small number of WGLs who had been unable to attend the first event). It was at this initial training event that WGLs were provided with the programme materials, including the project handbook, online access to all materials required to implement the programme, and access to an exclusive online NCETM forum managed by the NCETM Assistant Director, Ione Crossley. WGLs then used the mathematical reasoning programme materials in the schools in which they work before attending a two-day follow-up training session to share their experience of teaching the units and to prepare to train implementing schools.

The follow-up two-day training event for WGLs was held in June/July 2016. The first day focused on WGLs sharing their experiences of implementing the programme in their own classrooms and to learn more about how they would be evaluated by the NCETM. The second day focused on preparing WGLs to deliver their own training to treatment schools and how to facilitate trial schools' use of the NCETM community and other programme expectations, such as school visits. WGLs were able to ask questions and share their experiences of supporting treatment schools through the NCETM online community (which was set up and active from the first training event in February 2016). WGLs were also able to make contact with the NCETM directly if they required additional support.

Training days for treatment schools were held throughout September and October 2016 at one of the WGLs' schools. WGLs were provided with a series of PowerPoint presentations by the NCETM and Oxford to use as the basis for the training to ensure consistency. These slides largely followed WGLs' own training but had been adjusted slightly and more notes included after feedback at the June/July training days. Each day began with an introduction to the intervention, which included background theory and evidence, and was followed by a description of the specific concepts addressed in the programme and an explanation of the structure and detail of the teaching units. This was followed by a session to explore the use of online materials (including the games) as well as some discussion regarding the nature of the intervention, including arrangements for visits and testing.

After the training, Work Group Leads provided further support to the teachers through a school visit during the period in which they were delivering the programme. Where possible this included a lesson observation and gave teachers an opportunity to discuss any issues they were facing in implementation. Visits occurred between November 2016 and March 2017. Teachers in the intervention group were also able to access a local Maths Hub online community managed by Work Group Leads, where they could ask questions, raise issues, and share experiences.

One of the aims of the process evaluation was to assess fidelity, and later in this report we discuss the extent to which the programme was implemented as intended. A small number of schools withdrew from delivering the programme. Most participating schools delivered all the programme units, although this often occurred over a longer period of time than anticipated. In some schools, there were issues with playing the online games that formed part of the programme. All of these issues are discussed in greater detail in later sections of this report.

Recruitment to the trial was successful with slightly more schools recruited than initially planned. Some concerns were raised during the testing for the evaluation, focusing particularly on the appropriateness of such testing for children of a relatively young age. These concerns were investigated by all parties involved and all schools were contacted to provide an opportunity to air any issues and to ensure lessons were learned for future evaluations. In terms of interpreting the results regarding the impact of the programme, we have no reason to believe that the prevalence of such issues would have differed systematically among the treatment and control groups.

Evaluation objectives

The primary research question the impact evaluation was designed to answer is:

- What is the effect of the Mathematical Reasoning programme on children's mathematical attainment (as measured by the Progress Test in Maths) at the end of Year 2?

The impact evaluation also aimed to explore:

- What is the effect of the Mathematical Reasoning programme on mathematical attainment at the end of Year 2 among pupils who are eligible for free school meals (FSM)?

The purpose of the process evaluation was to establish fidelity and to identify the factors influencing impact. A key aim was to understand the implications of the introduction of the train-the-trainers model.

The protocol for the evaluation is available at:

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_Maths_and_Reasoning_effectiveness_trial.pdf

Ethics and trial registration

Ethical review was undertaken by the University of Oxford, with the project receiving ethical approval from the Departmental Research Ethics Committee in February 2016. In addition, NIESR adheres to the Ethics Guidelines of the Social Research Association (SRA).

As the intervention was delivered within school hours, consent from the school was considered sufficient with regard to consent for the intervention; and as randomisation took place at the school level (rather than randomising individual pupils), the decision to enter into randomisation could also be made by the school. The requirements of participating in the trial were set out in the Memorandum of Understanding for the evaluation, a copy of which is provided in Appendix C.

A parental information sheet provided with an opt-out form (Appendix C) gave information on the aims of the research and the use of data in order that parents (or legal guardians) were able to make an informed decision about whether to withhold consent from data sharing.

International Standard Randomised Controlled Trial Number: ISRCTN89670776

Data protection

As noted above, schools were given an information sheet providing the details of the trial prior to deciding to take part, with the requirements of participating set out within the Memorandum of Understanding. Parents were also provided with a letter outlining the trial, explaining the reasons for seeking access to data and how this data would be used, along with an opt-out form for consent for data sharing.

Each of the organisations forming part of the evaluation and delivery teams have rigorous data security policies in place. In addition, a data sharing agreement for the project was drawn up between the evaluation and delivery teams setting out agreed processes for the secure storage and transfer of data. This included details of methods for secure transfer of data and the use of password-protection and encryption as appropriate.

Project team

The project team comprised a collaboration between the University of Oxford and the National Centre for Excellence in the Teaching of Mathematics (NCETM). The intervention was developed by a team

at the University of Oxford, led by Professor Terezinha Nunes and Professor Peter Bryant, along with Rossana Barros Baertl and Deborah Evans. The NCETM led on the delivery of the intervention and contributed to the training model used in this trial, with the NCETM's involvement led by Ione Crossley. The NCETM worked in partnership with Maths Hubs in order to deliver the intervention.

The evaluation was led by the National Institute of Economic and Social Research (NIESR) in partnership with the National Centre for Social Research (NatCen). The evaluation was led by Lucy Stokes (NIESR), working with Richard Dorsett (formerly NIESR, now University of Westminster), and in the earlier stages of the trial also involved Jake Anders (formerly NIESR, now UCL Institute of Education). The process evaluation was led by Nathan Hudson-Sharp (NIESR) with support from Heather Rolfe (NIESR). Anitha George (formerly process evaluation lead), Jonathan Buzzeo, and Naomi Munro-Lott also contributed to the process evaluation during their time at NIESR. NatCen were responsible for the administration of the assessments that formed part of the evaluation, with the NatCen team led by Rakhee Patel, working with Lydia Marshall, Hannah Piggott, and Katie Drysdale.

Methods

Trial design

| | | |
|---|--|---|
| Trial type and number of arms | Cluster randomised controlled trial, 2 arms | |
| Unit of randomisation | School | |
| Stratification variable(s) (if applicable) | Maths Hub; Proportion of pupils eligible for Free School Meals; Prior attainment (school-level KS1 results) ² | |
| Primary outcome | variable | Progress Test in Maths (GL Assessment), Level 7 |
| | measure (instrument, scale) | Overall standardised score |
| Secondary outcome(s) | variable(s) | No secondary outcomes |
| | measure(s) (instrument, scale) | Not applicable |

The evaluation used a cluster randomised controlled trial design, with randomisation taking place at the school level. Since the programme is a class-level intervention, the choice was between randomising classes or randomising schools. As class-level randomisation may have resulted in cross-contamination across trial arms, school-level randomisation was deemed preferable.

There were two arms of the trial, with schools randomly allocated to either the treatment arm (receiving the Mathematical Reasoning programme) or the control group. Schools in the control group were expected to deliver 'business as usual' mathematics teaching and were offered the opportunity to take part in the programme in the following school year (that is, for the pupils who were in Year 2 in the following academic year). This waitlist design was chosen with the aim of minimising attrition from the trial.

Participant selection

Schools were recruited to the trial through Maths Hubs, working with the NCETM. Eight hubs were identified to participate in the trial, with each hub aiming to recruit 20 schools. The participating hubs were as follows:

- Archimedes (North East, Durham and Tees Valley region)
- Central (Birmingham)
- GLOW (Gloucestershire, Oxfordshire, Worcestershire)
- Kent and Medway
- London South East
- North West Three
- Salop and Herefordshire
- Sussex

² Please see the discussion under randomisation for further information on the definition used for school-level KS1 results.

All English state primary and infant schools within the eight hubs were eligible to participate in the trial. This included schools with one class per year group and multiple classes per year group, as well as schools with classes with mixed year groups. Within each school, all Year 2 pupils and teachers were eligible to participate.

The recruitment process was led by the NCETM working with the participating Maths Hubs. Hubs advertised the project and the opportunity for schools to take part through their newsletters, websites, and mailing lists (using a common set of words produced by the NCETM). Prior to this there was also some informal contacting of schools by hubs to alert them to the project. Interested schools were invited to contact their Maths Hub and were sent an information sheet and expression of interest form to complete and return. The NCETM website also advertised the project, listed participating Hubs, and provided further details and the expression of interest form. The opportunity was also mentioned in other NCETM communication channels including the NCETM newsletter.

Based on the expression of interest forms completed by interested schools, each Maths Hub selected around 20 schools. The expression of interest form included a space for schools to say why they were interested in being part of the study. Once expressions of interest forms were received, Maths Hubs reviewed the forms and had informal conversations with schools, checking that they were eligible and understood what committing to the study involved. Having confirmed eligibility, where more schools expressed interest than could participate (after making some allowance for over-recruitment), Maths Hubs were instructed to select those schools which appeared most committed to the trial.

The selected schools were then asked to sign a Memorandum of Understanding, distribute consent letters to parents (along with an accompanying letter from the headteacher, at the discretion of the school), and return a completed data sheet asking for school and pupil details. All of these were required to be returned prior to randomisation taking place. The NCETM and Maths Hub administrators worked closely with NIESR during this process so that schools had a single point of contact and clear lines of communication.

Outcome measures

The primary outcome is attainment in maths, as measured by the GL Assessment Progress Test in Maths (PTM). Level 7 of the test was used—the appropriate level for this age group (6–7 years of age). This measure was chosen as the previous efficacy trial had identified a positive and significant impact on attainment in maths as measured by the GL Assessment Progress in Maths (the predecessor to the Progress Test in Maths series). By the time of testing for this trial, the former Progress in Maths test was no longer available for use.

The test assesses pupils' mathematical skills and knowledge. As specified in the evaluation protocol, the overall standardised score—the 'Standard Age Score' as provided by GL Assessment—was chosen as the outcome measure. This is based on the pupil's raw score (equivalent to the sum of the scores for each of the areas mentioned below) and adjusted for age, and allows for comparison with a nationally representative sample of students of the same age in the U.K. (with the average score equal to 100). The standardised scores are based on the results from a standardisation exercise undertaken by GL Assessment, selecting a sample of U.K. schools through stratified random sampling. For Level 7 of the test the standardisation involved around 4,000 pupils. Further details on the standardisation process are available at <https://www.gl-assessment.co.uk/media/1346/ptm-technical-information.pdf>

The standardised score—chosen over the raw score as the primary analysis—does not control for pupil age. In the efficacy trial, Worth et al. (2015) used the raw scores but also controlled for age. As an additional check, we also test the sensitivity of our results to using the raw score instead and find that this made little difference to the results. More information about the test can be found on the GL Assessment website: <https://www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/>

The assessments were administered by NatCen in June and July 2017. The paper version of the test was used, as although a digital version also exists, it was felt that a paper version would be more appropriate for the age group and would help avoid practical issues of implementing digital tests in schools. A team of assessors were trained in carrying out the assessments who were blind to whether schools were in the treatment or control groups. The test typically takes between 45–60 minutes to complete; tests were generally administered to whole classes, although in a few schools tests were carried out in smaller groups. Teachers were present in the classroom to provide reassurance for the pupils. The tests were scored by GL Assessment, which is the standard procedure for marking this test; GL Assessment were also blind to the treatment status of the schools.

No secondary outcomes were specified in the protocol. However, in this report we do provide some exploratory analysis of the separate subscores provided as part of the PTM results in order to investigate whether impacts were apparent for the different aspects of mathematical knowledge. As noted above, subscores are provided for performance in four areas: fluency in facts and procedures, fluency in conceptual understanding, problem solving, and mathematical reasoning, which are intended to link to national curriculum areas. These subscores are based on the total raw score obtained on the questions that relate to each of the four areas. These subscores are not standardised for age. Although the possibility of using KS1 assessments as a secondary outcome was considered at the design stage, ultimately this was not included in the protocol as it was felt there was potential for this to introduce bias as the KS1 data available through the NPD is based on teacher assessment.

Sample size

Information from the efficacy trial (Worth et al., 2015) was used to inform the sample size calculations for this trial. Worth et al. reported an intra-cluster correlation, before controlling for covariates, of 0.12 and that 57% of both school-level and individual-level variance was controlled for by covariates. This information was used to estimate the minimum detectable effect size (MDES). The effect size in the efficacy trial was 0.2; given the possibility of some dilution as a result of the train-the-trainers model, the sample size was chosen with this in mind. The proposed sample size was therefore 160 schools (across eight hubs), assuming 45 pupils per school. On this basis, the MDES was estimated at 0.11 at the time of preparing the protocol (see Table 3).

It was assumed that 15% of pupils in the sample were FSM-eligible (Department for Education statistics indicate the percentage of pupils aged 5–10 eligible for free school meals stood at 15.3% in January 2016; Department for Education, 2016), and assuming that all other parameters remain the same, the MDES for this subgroup stood at 0.14. In our analysis, we identify children eligible for FSM on the basis of whether pupils have been recorded as eligible for FSM at any time in the last six years in any termly or annual School Census (as indicated by the variable `everfsm_6_p_spr17` in the NPD).

The later section on participant flow provides further details on the eventual sample size achieved, and reports the implications for the MDES.

Randomisation

Schools recruited by the project team were randomly assigned by the evaluation team.

Schools were randomised within blocks defined on the basis of hubs, proportion of children eligible for FSM and prior attainment at KS1 (that is, school-level KS1 attainment in the academic year 2014/2015). Within each of the eight hubs, two FSM groups were determined: ‘high’ and ‘low’—with schools ranked within hubs by the proportion of pupils eligible for FSM, with thresholds for the ‘high’ and ‘low’ groups then chosen so that half of all schools fall into each group. Within each of these Hub/FSM groups, schools were then allocated into two KS1 groups (again ranking schools on the basis of their KS1 attainment and allocating schools to ‘high’ and ‘low’ groups, so that half of schools fall into each category). Schools were allocated to ‘high’ and ‘low’ groups for KS1 attainment based on the proportion

of pupils in the school attaining level 2 or above in maths at KS1, then the proportion attaining level 3 and then the proportion attaining level 4. This multiway sorting is used to break ties, with outstanding ties broken by random ordering.

With eight hubs, two FSM groups, and two KS1 groups, this resulted in 32 blocks.

The purpose of this blocking was to improve cross-arm comparability of schools and also to increase the precision of estimates. The regression models used to estimate impacts include block identifier variables to reflect this aspect of the randomisation design. Note that the analysis was not undertaken blinded to randomisation.

Randomisation of schools, to achieve a 50:50 allocation, was performed as follows:

- each school was assigned a randomly generated number;
- schools were sorted by hub/FSM/KS1 block and randomly generated number;
- the first school was randomised to treatment or control; and
- each subsequent school was assigned to have the opposite outcome of the previous school.

This continued until all schools had been assigned.

The computer code used to carry out the randomisation is reported in Appendix D.

Statistical analysis

The impact of the intervention is estimated using linear regression models. Outcomes were regressed on an indicator of whether the school was in the treatment group, block indicators, and a measure of prior attainment. Our analysis is conducted on an intention-to-treat basis. A specific Statistical Analysis Plan was not published for this evaluation as this was prior to the EEF requirement to do so. Our plans for analysis were set out in the trial protocol, however, and the sections relating to randomisation and analysis were reviewed by an independent reviewer.

As specified in the trial protocol, our measure of prior attainment was lagged school-level attainment at KS1. Administering the Progress Test in Maths assessment prior to the intervention would have added considerably to the burdens on participating schools and to the cost of the trial, and thus it was decided to use a measure of prior attainment available from the NPD. The decision to use a measure of school-level attainment was also motivated in part by the findings of Bloom et al. (2007), who show, based on analysis of cluster trials for U.S. schools, that school-level pre-tests can achieve power levels comparable to using individual-level pre-tests.

More precisely, the measure of prior attainment used is KS1 attainment in 2016 (thus relating to the cohort of pupils who were in Year 2 in the year prior to those pupils participating in the trial).³ It is measured as the percentage of pupils in a school who are identified as working at the expected standard or above in maths.⁴ As a robustness test, we also checked the sensitivity of the results to replacing this school-level measure of attainment with the actual Early Years Foundation Stage Profile (EYFSP) scores of the pupils participating in the trial (using their total EYFSP score).⁵ While EYFSP scores have

³ We used 2016 KS1 attainment in the analysis, but the block randomisation was based on 2015 KS1 scores. This should reduce any concerns about variation across years.

⁴ We constructed these school-level averages from pupil-level data based on information from the NPD.

⁵ An advantage of the school-level measure of prior attainment is that there are no missing values. Consequently, the primary analysis is on a complete case basis (all pupils providing outcome data). This is not the case when using the EYFSP scores. We explore the sensitivity to missing values for estimates using this measure of prior attainment by also producing impacts with multiply-imputed prior attainment measures.

the advantage of being available at pupil-level, arguably this measure lacks sufficient granularity and so this was not chosen as the pre-test for the primary analysis.

As this is a school-level RCT, inference was based on standard errors adjusted for school-level clustering using Stata's 'cluster' option. Clustering standard errors in this way is reasonable given the large number of schools involved. Furthermore, it is an attractive approach relative to the leading alternative of a multilevel model since it avoids the need to assume that school-level effects are uncorrelated with other regressors and the biases that can result when this assumption is not met (Ebbes et al., 2004).

The regression results capture the effect of intention to treat. Estimates are presented as effect sizes, calculated using the Hedges' g formula. Formally, the effect sizes are calculated as follows:

$$g^* = \frac{\Gamma((n_T + n_C - 2)/2)}{\sqrt{(n_T + n_C - 2)/2} \cdot \Gamma((n_T + n_C - 3)/2)} \cdot \frac{\beta_T}{\sqrt{\frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2}{n_T + n_C - 2}}}$$

where n_T is the number of treatment group observations, n_C is the number of control group observations, $\Gamma()$ is the gamma function, β_T is the regression coefficient on the dummy variable indicating membership of the treatment group, s_T^2 is the variance of the outcome variable among the treated group and s_C^2 is the variance of the outcome variable among the control group.

Impacts were also estimated for the subgroup of FSM pupils using the same approach as above. This subgroup was identified using the indicator of whether pupils had been recorded as eligible for FSM at any time in the last six years in any termly or annual Census (as indicated by the variable `everfsm_6_p_spr17` in the NPD).

The protocol states that the evaluation will also consider how the data collected on fidelity through the survey could also be used in quantitative analysis to investigate any potential moderation of treatment effects. With this in mind, the survey of treatment schools included three questions that could potentially be used to consider implementation of the programme. Responses to these are summarised later in the report, although the fact that we do not have responses to the survey from all schools and that there is limited variation in the responses to these questions among those schools that do respond, limits the usefulness of that analysis.

In addition, we undertake an exploratory analysis not specified in the trial protocol: investigating whether there are impacts on the subscores of the Progress Test in Maths using the same model as for the primary outcome as specified above. We do so in order to explore whether the intervention may have affected particular aspects of mathematical skills; it is conceivable that such effects are masked when considering the overall score.

Implementation and process evaluation

The overarching purpose of this process evaluation was to examine fidelity to the implementation of the Mathematical Reasoning programme by treatment schools. It explores how the intervention was implemented, whether this differed from the intended treatment model, and the factors that informed this in order to identify potential influences on the programme's impact. It also explores the perceived impact of the intervention from the perspective of treatment schools as well as monitoring the activity of the control group to establish what was done in the absence of the intervention.

This process evaluation drew upon a wide range of activities and sources. Key elements include:

- NIESR researchers attending initial training for Work Group Leads (WGLs), hosted by the NCETM in February 2016;
- NIESR researchers attending follow-up training for WGLs, hosted by the NCETM in June/July 2016;
- NIESR researchers attending teacher training days in three different Maths Hubs, hosted by WGLs; these took place between September and October 2016;
- visits to eight schools across three Maths Hubs to interview teachers implementing the intervention, and to observe sessions; these took place in February and March 2017;
- attendance at the final training day for WGLs, hosted by the NCETM in July 2017; and
- online surveys of both treatment and control schools; this was undertaken in June/July 2017.

Attendance at all training events was supplemented by evaluation forms completed by attendees. These evaluation forms were produced and collated by the NCETM or WGLs respectively. The process evaluation team was also provided with access to the online NCETM forum, scanned copies of WGL's school visit forms, and with data collected by the project team at the University of Oxford regarding treatment schools' use of the programme's online games.

The process evaluation team attended both WGL and teacher training days to understand experiences of the intervention as well as expectations and any concerns. Training content and resources were reviewed to understand both how WGLs are prepared to train other teachers and how all teachers are prepared to deliver the intervention. Findings were triangulated with training day evaluation surveys undertaken and shared by the NCETM, as well as our own interviews and survey of treatment schools.

In addition to obtaining experiences of the training, school visits were undertaken to interview teachers on their experiences of using the project resources and delivering the intervention, including views on the appropriateness and value of the resources, any adaptations they made, and pupil response. These were supplemented with classroom observations (using a proforma designed for this purpose) which sought to assess pupil engagement and evidence relating to their understanding of the principles taught in the sessions, as well as observe pupils using online games.

Interviews were digitally recorded, with the agreement of teachers, and transcribed. Data was analysed using a social research 'framework' approach, drawing themes and messages from an analysis of interview transcripts, observations of training and of lessons, and other materials collected by evaluation and project teams.

The online survey of all intervention schools gathered data in a consistent way on implementation and perceived outcomes and on factors that could affect fidelity. It covered experiences of using the lesson plans and guidance materials, including preparation and delivery time. This included questions on teachers' and schools' usual approaches to teaching maths, in particular whether this includes a focus on the principles underlying the Mathematical Reasoning programme and whole class teaching. The end-of-project survey was completed by 50 schools, equivalent to 62.5% of all treatment schools that finished the programme (N = 80).

The survey of control schools sought to obtain information regarding usual approaches to teaching maths to Year 2 in order to contribute to broad understanding of implementation, feasibility, and impact. The control group survey was completed by 59 schools, equivalent to 73.8% of control schools that remained in the trial (N = 80).

Steps were taken to ensure that the eight case study schools included a variety of delivery contexts. This included variation by Maths Hub, Ofsted rating, proportion of pupils receiving free school meals, and geographical location. However, due to the relatively small proportion of the treatment group as a whole, the findings therefore may not necessarily reflect the views of the wider population of treatment and control schools. Nevertheless, we believe the qualitative data collected through the visits provide useful insights into the range and diversity of views, and the experience of participants, in the

Mathematical Reasoning intervention. The findings of the process evaluation should be considered with these strengths and limitations in mind. The sections discussing the results from the process evaluation later in this report summarise the key findings from the combined process evaluation data-gathering exercises.

Costs

Information on costs was obtained primarily from the NCETM and the participating Maths Hubs. Some information relating to costs was also provided by the University of Oxford team. This information was then used by the evaluation team to produce an estimate of the cost per pupil per year, based on the direct, marginal costs of the programme and following EEF guidance. In addition, issues around costs incurred were discussed on the visits to schools that formed part of the process evaluation; and respondents to the survey in treatment schools were asked about any additional costs incurred in implementing the programme.

Timeline

Table 2: Timeline

| Date | Activity |
|---------------------------|---|
| October–December 2015 | Recruitment of Maths Hubs and further development of the programme |
| January–May 2016 | Recruitment of schools (Maths Hubs/NCETM) and training of WGLs (NCETM/Oxford) |
| May–June 2016 | Collection of consent and pupil data (NCETM/Maths Hubs/NIESR) |
| July 2016 | Randomisation |
| September–October 2016 | Training of teachers (NCETM/Maths Hubs) |
| September–April 2017 | Delivery of programme (NCETM/Maths Hubs) |
| Feb–March 2017 | School visits (NIESR) |
| June–July 2017 | Survey of intervention and control schools (NIESR) |
| June–July 2017 | Post-tests administered (NatCen) |
| September–December 2017 | Training of teachers from control schools (NCETM/Maths Hubs) |
| September 2017–April 2018 | Analysis and reporting (NIESR) |

Impact evaluation

Participant flow including losses and exclusions

Eight Maths Hubs were identified to take part in the trial by the NCETM. The aim was to recruit 20 schools per hub so that in total 160 schools would be participating in the trial. The opportunity for schools to participate in the trial was widely advertised by the NCETM and the Maths Hubs (see the earlier section explaining the recruitment strategy), with interested schools initially asked to complete an expression of interest (EOI) form. It is not possible to identify precisely how many schools were approached (as it is not possible to say with accuracy how many schools may have seen the trial advertised). However, a total of 335 EOI forms were returned, based on records provided by the Maths Hubs.⁶

Within each Maths Hub, around 20 schools were then invited to participate such that in total 175 schools were approached.⁷ This followed agreement with the EEF for an allowance to over-recruit by up to 15%. In order to select the schools that would participate, Hubs reviewed the completed EOI forms (which included a space for schools to explain why they wanted to participate), followed up by conversations with schools in order to ascertain which schools had fully understood the requirements of participating in the evaluation and were the most committed to the trial.

These 175 schools were asked to sign the MOU, distribute opt-out consent letters, and return the completed data collection form. The majority of the selected schools did so such that a total of 169 schools were eligible for the trial at the point of randomisation.⁸

After excluding those pupils who opted-out of participating in the trial (78 pupils), across the 169 participating schools, a total of 7,419 pupils were participating at the point of randomisation.

Of these 169 schools, 84 were allocated to receive the intervention in the school year 2016/2017 (comprising 3,793 pupils), while the remaining 85 schools were allocated to the control group, receiving the intervention in the school year 2017/2018 (comprising 3,626 pupils).

One hundred and sixty of the 169 participating schools took part in the post-test in summer 2017 (80 treatment and 80 control schools). The five control schools that withdrew from the testing (and thus future receipt of the programme) all cited staff changes as a key reason for their withdrawal. In addition, two of these schools also mentioned that they were planning to introduce an alternative maths programme in the following school year instead. The four treatment schools that did not participate in the testing had all withdrawn from delivering the programme (citing varying reasons including ICT issues, staff changes, and issues with the actual programme itself). While often the schools had indicated they would still be willing to participate in the testing, despite withdrawing from the programme, in practice this did not prove to be the case for these four schools.

In addition, three further treatment schools, although they took part in the post-test, stopped delivering the programme. In addition, in one school the programme was delivered in one class, but not within a second mixed year group class. Following the intention-to-treat principle these schools (and class) are

⁶ This comprised: London and South East (34 schools); Salop and Herefordshire (36 schools); Central (48 schools); GLOW (42 schools); NW3 (44 schools); Archimedes (45 schools); Kent and Medway (50 schools); and Sussex (36 schools).

⁷ The numbers selected by Hub were: London and South East (20 schools); Salop and Herefordshire (22 schools); Central (23 schools); GLOW (22 schools); NW3 (23 schools); Archimedes (23 schools); Kent and Medway (22 schools); and Sussex (20 schools).

⁸ One school in the Archimedes hub and two schools in the NW3 hub withdrew or did not provide the necessary information to participate. This applied for five of the original schools selected to participate in the Sussex hub; three additional schools were approached to participate instead, two of which did so (resulting in a total of 17 participating schools in the Sussex hub).

included in the primary analysis, but we also present the results of sensitivity analysis excluding these three schools, plus the class where the pupils did not receive the programme.

As noted above, 7,419 pupils were participating at the point of randomisation. The four treatment schools that did not participate in testing accounted for 198 pupils, and the five control schools for 175 pupils. A further 357 pupils in treatment schools did not participate in the post-test. The most common reason, applying to 221 pupils, was having moved school. In addition, 62 pupils did not take part in the test due to having SEN or a disability, 56 pupils were absent on the day of the test, 11 pupils were long-term absent (sometimes due to illness), and a further seven were not eligible to take the test for other reasons. Among control schools, 209 pupils did not participate in the post-test as they had moved school, 57 were absent on the day of the test, 35 had SEN or a disability, nine were long-term ill or absent, and 22 were not eligible for other reasons. In addition, consent was withdrawn for four pupils in control schools after the point of randomisation.

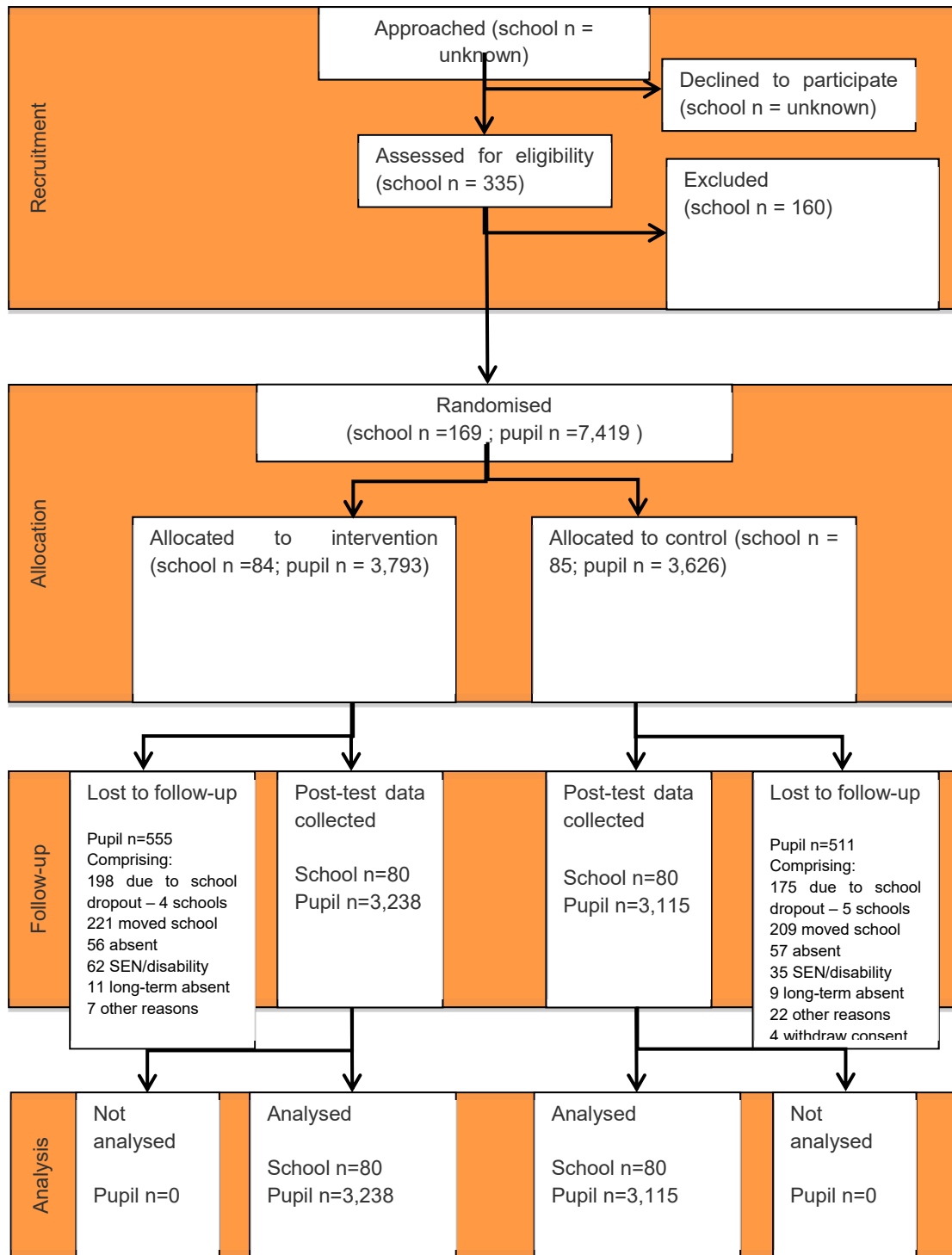
As the initial data collection to obtain details of the participating schools and pupils took place in the summer term before the intervention was to be delivered, some pupils moved school before joining Year 2. In light of this, a second round of data collection was undertaken in Autumn 2016, once pupils were in Year 2, with the aim of identifying any new pupils who had joined the schools since the time of the original data collection, and any who had left. Details of any new pupils who had joined the schools were collected at this point (providing that consent was obtained). In practice, around half of schools responded to this second round of data collection. Our primary analysis focuses, therefore, on those pupils in the study at the point of randomisation. However, as discussed later in this report, we do also conduct some sensitivity analysis by altering the analysis sample to additionally include those pupils identified in the second data collection and removing those who had left the school by this point.⁹

Almost all of those pupils who undertook the post-test could be matched to the NPD (with the exception of one pupil). Our primary analysis does not rely on NPD linkage at pupil-level (as we are using a school-level measure of prior attainment), and so the pupil for whom it is not possible to match to NPD at pupil-level can still be included within our primary analysis. Thus our final sample for analysis comprises 3,238 pupils in 80 treatment schools, and 3,115 pupils in 80 control schools.

As reported in Table 3, at the point of preparing the trial protocol, the MDES stood at 0.11 (and at 0.14 for analysis of the subgroup of pupils eligible for FSM). At randomisation, the MDES stood at 0.12 (and at 0.16 for the FSM subgroup). Based on the final sample available for analysis, the MDES stands at 0.14 (and at 0.20 for the FSM subgroup).

⁹ This makes no substantive difference to the results.

Figure 2: Participant flow diagram¹⁰



¹⁰As noted above, for some, but not all schools we were able to obtain updated information on Year 2 classes in the Autumn term. In the analysis section we check the robustness of our results based on this additional information (which results in an additional 91 pupils being included in the analysis); this makes no substantive difference to the findings.

Table 3: Minimum detectable effect size at different stages

| | | Protocol | | Randomisation | | Analysis | |
|---|----------------------------------|----------|------|---------------|------|----------|------|
| | | Overall | FSM | Overall | FSM | Overall | FSM |
| MDES | | 0.11 | 0.14 | 0.12 | 0.16 | 0.14 | 0.20 |
| Pre-test/ post-test correlations | level 1 (pupil) ^a | 0.75 | 0.60 | - | - | - | - |
| | level 2 (school) ^b | 0.75 | 0.60 | 0.75 | 0.60 | 0.58 | 0.44 |
| Intracluster correlations (ICCs)^c | level 2 (class) | 0.12 | 0.12 | 0.12 | 0.12 | 0.11 | 0.13 |
| 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? | | 2 | 2 | 2 | 2 | 2 | 2 |
| Average cluster size | | 45 | 6.8 | 43.9 | 9.9 | 39.7 | 8.9 |
| Number of schools | intervention | 80 | 80 | 84 | 84 | 80 | 76 |
| | control | 80 | 80 | 85 | 85 | 80 | 73 |
| | total | 160 | 160 | 169 | 169 | 160 | 149 |
| Number of pupils | intervention | 3600 | 540 | 3793 | 830 | 3238 | 649 |
| | control | 3600 | 540 | 3626 | 835 | 3115 | 674 |
| | total | 7200 | 1080 | 7419 | 1665 | 6353 | 1323 |

- Note there is no level 1 pre-test/post-test correlation since we do not have a level 1 pre-test. This was, however, included in the calculations at randomisation.
- The pre-post correlations under the 'Analysis' heading are the square root of the R-squareds from regressions of the outcome variable on the school-level baseline measure, also including blocking variables. These, rather than the raw correlations, are used in the power calculations.
- We use an assumed ICC of 0.12 based on Worth et al. (2015) for MDES estimation at protocol and randomisation, but the actual ICC at analysis stage.

Attrition

As 7,419 pupils were in the sample at the point of randomisation, and 6,353 of these pupils remained in the sample at the point of analysis, this represents an attrition rate of 14%. The reasons for attrition are discussed above in the section on participant flow; the two key contributors were school dropout and pupils who moved school between the point of randomisation and delivery of the post-test.

Pupil and school characteristics

Table 4 presents school- and pupil-level characteristics at the point of randomisation for both the intervention and control group. Overall, this indicates that the sample was balanced at baseline, at least in terms of the characteristics we are able to observe.

There were no significant differences by treatment arm in the percentage of schools located in urban areas, or the percentage of schools with a religious affiliation. There were also no significant differences by treatment arm in the distribution of schools by type, or by Ofsted overall effectiveness rating. In addition, there were no significant differences by treatment arm in school size, or in the composition of

pupils (that is, the proportions eligible or ever eligible for free school meals, for whom English is an additional language, or those with special educational needs).

School-level KS1 attainment in maths in 2015 was one of the blocking variables used in randomisation and therefore we would anticipate that the sample would be balanced in this respect. Although mean school-level KS1 attainment in 2016 appears slightly higher in the control arm (74.7% of pupils working at the expected standard or above) compared with the treatment group (72%), this difference in means is not statistically significant (the effect size here is 0.23).

Although there is no individual-level pre-test in the primary analysis, we can use the EYFSP scores to gain an insight into the degree to which attainment is balanced at baseline. There was no difference in the mean EYFSP score between pupils in the treatment and control groups, and the distribution of scores was similar in the two groups, as shown by the histograms in Appendix E. The balance on this measure is also demonstrated by the effect size of 0.00. There were no statistically significant differences by treatment arm on any of the other pupil characteristics considered (gender, age, and proportion eligible for free school meals).

Table 4: Baseline comparison

| School-level (categorical) | Intervention group | | Control group | |
|--|--------------------|------------|---------------|------------|
| | n/N (missing) | Count (%) | n/N (missing) | Count (%) |
| Ofsted overall effectiveness:¹ | | | | |
| Outstanding | 21/83 (1) | 21 (25.3%) | 12/80 (5) | 12 (15.0%) |
| Good | 53/83 (1) | 53 (63.9%) | 62/80 (5) | 62 (77.5%) |
| Requires improvement | 8/83 (1) | 8 (9.6%) | 6/80 (5) | 6 (7.5%) |
| Inadequate | 1/83 (1) | 1 (1.2%) | 0/80 (5) | 0 (0.0%) |
| | | | | |
| School type:² | | | | |
| Academy converter | 14/84 (0) | 14 (16.7%) | 16/85 (0) | 16 (18.8%) |
| Academy sponsor led | 3/84 (0) | 3 (3.6%) | 6/85 (0) | 6 (7.1%) |
| Community school | 39/84 (0) | 39 (46.4%) | 38/85 (0) | 38 (44.7%) |
| Foundation school | 3/84 (0) | 3 (3.6%) | 3/85 (0) | 3 (3.5%) |
| Voluntary aided school | 18/84 (0) | 18 (21.4%) | 13/85 (0) | 13 (15.3%) |
| Voluntary controlled school | 7/84 (0) | 7 (8.3%) | 9/85 (0) | 9 (10.6%) |
| | | | | |
| In urban area³ | 69/84 (0) | 69 (82.1%) | 62/85 (0) | 62 (72.9%) |
| Religious affiliation⁴ | 34/84 (0) | 34 (40.5%) | 29/85 (0) | 29 (34.1%) |
| | | | | |

| School-level (continuous) ⁵ | n (missing) | Mean (SD) | n (missing) | Mean (SD) | |
|---|----------------|---------------|----------------|---------------|-------------|
| Number of pupils on roll | 83 (1) | 317.9 (159.6) | 82 (3) | 289.2 (156.4) | |
| % pupils eligible for FSM | 84 (0) | 16.8 (12.7) | 85 (0) | 16.5 (12.3) | |
| % pupils ever eligible for FSM | 84 (0) | 26.2 (16.4) | 85 (0) | 26.3 (17.1) | |
| % pupils with EAL | 83 (1) | 11.3 (16.9) | 82 (3) | 12.5 (17.7) | |
| % pupils with SEN support | 83 (1) | 14.1 (6.5) | 82 (3) | 14.4 (6.8) | |
| | | | | | |
| KS1: % pupils working at expected standard or above, 2016 | 84 (0) | 72.0% (12.7) | 85 (0) | 74.7% (11.1) | |
| KS1: % pupils working at level 2 or above, 2015 | 84 (0) | 93.5% (5.4) | 85 (0) | 94.0% (5.0) | |
| Pupil-level (categorical) | n/N (missing) | Count (%) | n/N (missing) | Count (%) | |
| Eligible for FSM | 830/3758 (35) | 830 (22.1%) | 835/3596 (30) | 835 (23.2%) | |
| Female | 1879/3793 (0) | 1879 (49.5%) | 1756/3622 (4) | 1756 (48.5%) | |
| EYFSP – achieved good level of development | 2427/3712 (81) | 2427 (65.4%) | 2344/3527 (99) | 2344 (66.5%) | |
| Pupil-level (continuous) | n (missing) | Mean (SD) | n (missing) | Mean (SD) | Effect Size |
| EYFSP total points score | 3712 (81) | 34.3 (7.7) | 3526 (100) | 34.3 (7.3) | 0.00 |
| Age at September 2016 (months) | 3758 (35) | 78.5 (3.5) | 3596 (30) | 78.5 (3.5) | |
| Age at September 2016 (years) | 3758 (35) | 6.1 (0.3) | 3596 (30) | 6.1 (0.3) | |

Notes and sources:

1. Ofsted inspection ratings as at 31 August 2016.
2. As reported in Edubase, extract downloaded 6 July 2016
3. As reported in Edubase, schools located in 'Urban city and town' or 'Urban major conurbation'.
4. As reported in Edubase, includes Church of England and Roman Catholic schools.
5. School-level characteristics (FSM, SEN, EAL, number of pupils), as reported in DfE Performance Tables, 2015.

All other characteristics are as reported in the NPD extracts supplied for this project.

Table 5 presents pupil- and school-level characteristics for the analysis sample. We may be concerned that attrition may have led to an imbalance in the sample. However, the sample remained balanced by treatment arm in terms of all the observed characteristics as reported in the table below.

Histograms of the EYFSP scores, based on the analysis sample, are provided in Appendix E. Again, there was no statistically significant difference in the mean school-level KS1 attainment in maths, or in the mean total EYFSP score. The effect size for the difference in total EYFSP scores stood at 0.02 based on the analysis sample.

Table 5: Comparison—analysis sample

| School-level (categorical) | Intervention group | | Control group | |
|--|--------------------|------------------|--------------------|------------------|
| | n/N (missing) | Count (%) | n/N (missing) | Count (%) |
| Ofsted overall effectiveness:¹ | | | | |
| Outstanding | 21/79 (1) | 21 (26.6%) | 12/75 (5) | 12 (16.0%) |
| Good | 49/79 (1) | 49 (62.0%) | 58/75 (5) | 58 (77.3%) |
| Requires improvement | 8/79 (1) | 8 (10.1%) | 5/75 (5) | 5 (6.7%) |
| Inadequate | 1/79 (1) | 1 (1.3%) | 0/75 (5) | 0 (0.0%) |
| School type:² | | | | |
| Academy converter | 13/80 (0) | 13 (16.3%) | 14/80 (0) | 14 (17.5%) |
| Academy sponsored | 3/80 (0) | 3 (3.8%) | 6/80 (0) | 6 (7.5%) |
| Community school | 38/80 (0) | 38 (47.5%) | 36/80 (0) | 36 (45.0%) |
| Foundation school | 1/80 (0) | 1 (1.3%) | 3/80 (0) | 3 (3.8%) |
| Voluntary aided school | 18/80 (0) | 18 (22.5%) | 13/80 (0) | 13 (16.3%) |
| Voluntary controlled school | 7/80 (0) | 7 (8.8%) | 8/80 (0) | 8 (10.0%) |
| | | | | |
| In urban area³ | 65/80 (0) | 65 (81.3%) | 60/80 (0) | 60 (75.0%) |
| Religious affiliation⁴ | 34/80 (0) | 34 (42.5%) | 28/80 (0) | 28 (35.0%) |
| School-level (continuous) | n (missing) | Mean (SD) | n (missing) | Mean (SD) |
| Number of pupils on roll | 79 (1) | 313.0 (161.1) | 77 (3) | 292.2 (158.3) |
| % pupils eligible for FSM | 80 (0) | 16.0 (11.8) | 80 (0) | 16.2 (11.7) |
| % pupils ever eligible for FSM | 80 (0) | 25.3 (15.8) | 80 (0) | 25.9 (16.5) |
| % pupils with EAL | 79 (1) | 10.4 (15.1) | 77 (3) | 12.7 (17.8) |
| % pupils with SEN support | 79 (1) | 13.6 (6.2) | 77 (3) | 14.2 (6.6) |
| | | | | |
| KS1: % pupils working at expected standard or above, 2016⁵ | 80 (0) | 72.2 (12.4) | 80 (0) | 74.8 (11.0) |
| KS1: % pupils working at level 2 or above, 2015 | 80 (0) | 93.8 (5.3) | 80 (0) | 94.1 (4.9) |

| Pupil-level (categorical) | n/N (missing) | Count (%) | n/N (missing) | Count (%) |
|---|----------------|--------------|----------------|--------------|
| Eligible for FSM | 649/3237(0) | 649 (20.1%) | 674/3115 (0) | 674 (21.6%) |
| Female | 1647/3237(0) | 1647 (50.9%) | 1535/3115 (0) | 1535 (49.3%) |
| EYFSP—achieved good level of development | 2170/3201 (36) | 2170 (67.8%) | 2092/3064 (51) | 2092 (68.3%) |

| Pupil-level (continuous) | n (missing) | Mean (SD) | n (missing) | Mean (SD) | Effect size |
|---------------------------------------|-------------|------------|-------------|------------|-------------|
| EYFSP total points score | 3201 (36) | 34.8 (7.5) | 3064 (51) | 34.7 (7.1) | 0.02 |
| Age at September 2016 (months) | 3237 (0) | 78.5 (3.5) | 3115 (0) | 78.5 (3.5) | 0.00 |
| Age at September 2016 (years) | 3237 (0) | 6.1 (0.3) | 3115 (0) | 6.1 (0.3) | 0.00 |

Notes and sources:

1. Ofsted inspection ratings as at 31 August 2016.
2. As reported in Edubase, extract downloaded 6 July 2016.
3. As reported in Edubase, schools located in 'Urban city and town' or 'Urban major conurbation'.
4. As reported in Edubase, includes Church of England and Roman Catholic schools.
5. Equivalent to an effect size of 0.22.
6. School-level characteristics (FSM, SEN, EAL, number of pupils), as reported in DfE Performance Tables, 2015.

All other characteristics are as reported in the NPD extracts supplied for this project.

We also explored the characteristics of pupils who drop out from the analysis sample (focusing on those in non-dropout schools only and excluding those for whom consent was withdrawn), regressing an indicator of whether or not pupils dropped out on treatment status, pupil characteristics (gender, age, eligibility for FSM, and EYFSP scores) and interactions between treatment status and these pupil characteristics. This showed that pupils who were eligible for free school meals were more likely to drop out than those who were not eligible, as were those with lower prior attainment as measured by their total EYFSP score. This could partly reflect teachers' perceptions of whether pupils were able to cope with the test situation. There were however no statistically significant associations with treatment status. While the above does not suggest differences by treatment arm, this may have consequences for our ability to generalise the results for the population of pupils as a whole.

Outcomes and analysis

Table 6 summarises the results of the primary analysis. Comparison of the mean scores indicates a slightly higher mean score among members of the treatment group (103.6) compared with the control group (102.2). Histograms show the distribution of scores to be similar across treatment and control groups (Figure 3).

The results indicate a small positive, but non-significant impact of the programme on the primary outcome measure. The effect size, using the values from Table 7 to calculate Hedges g , is estimated to be 0.08, which is equivalent to one additional month's progress. Since this is below the MDES, the result should be interpreted with caution; the trial was not powered to securely detect effects below 0.14.

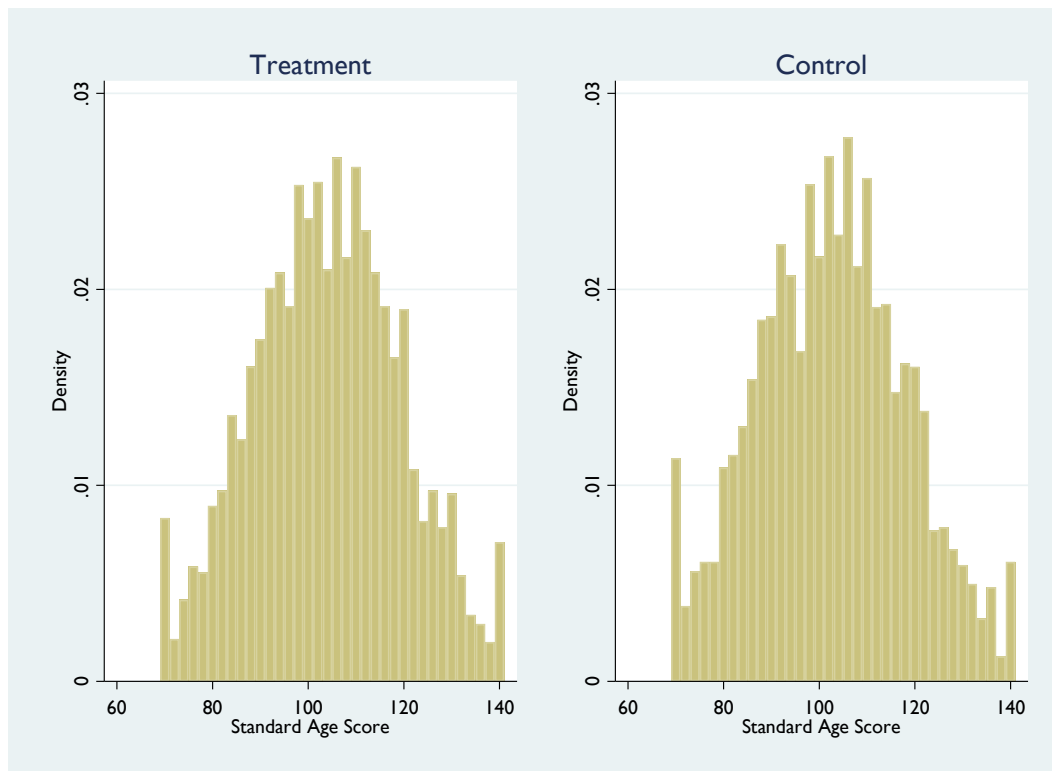
Table 6: Primary analysis

| Outcome | Raw means | | | | Effect size | | |
|---|--------------------|-------------------------|----------------|----------------------------|--|-----------------------|-------------|
| | Intervention group | | Control group | | n in model (intervention; control) | Hedges g (95% CI) | p- value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| PTM – standard age score | 3238 | 103.6 (103.0, 104.1) | 3115 | 102.2 (101.6, 102.7) | 6353 (3238; 3115) | 0.08 (-0.03, 0.18) | 0.156 |

Table 7: Effect size estimation

| Outcome | Unadjusted differences in means | Adjusted differences in means | Intervention group n (missing) | Intervention group Variance of outcome | Control group n (missing) | Control group Variance of outcome | Pooled variance | Population variance (if available) |
|---|---------------------------------------|-------------------------------------|--------------------------------------|---|---------------------------------|--|--------------------|---|
| PTM – standard age score | 1.414 | 1.181 | 3238 | 238.97 | 3115 | 246.77 | 242.80 | - |

Figure 3: Histograms of PTM standard age score, by trial arm



Subgroup analysis—pupils eligible for FSM

As specified in the trial protocol, we also estimate the model solely for those pupils eligible for free school meals. The results are presented in Table 8 below, with the components underlying the estimation of the effect size presented in Table 9. The estimated effect size is comparable to that for all pupils. Again, there is no significant impact of the intervention on our primary outcome measure for this subgroup of pupils.

Table 8: Impact on primary outcome, pupils eligible for FSM

| Outcome | Raw means | | | | Effect size | | |
|---|--------------------|-------------------|----------------|-------------------------|--|-----------------------|-------------|
| | Intervention group | | Control group | | n in model (intervention; control) | Hedges g (95% CI) | p- value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| PTM – standard age score | 649 | 96.4 (95.3, 97.6) | 674 | 95.7 (94.6, 96.8) | 1323 (649; 674) | 0.09 (-0.07, 0.25) | 0.288 |

Table 9: Effect size estimation, pupils eligible for FSM

| Outcome | Unadjusted differences in means | Adjusted differences in means | Intervention group | | Control group | | Pooled variance | Population variance (if available) |
|---|---------------------------------------|-------------------------------------|--------------------|---------------------------|----------------|---------------------------|--------------------|---|
| | | | n (missing) | Variance of outcome | n (missing) | Variance of outcome | | |
| PTM – standard age score | 0.74 | 1.31 | 649 | 221.29 | 674 | 213.91 | 217.53 | - |

Robustness checks

We also undertook a number of robustness checks of our analysis.

Firstly, we tested sensitivity to altering our pre-test measure. Instead of using a lagged school-level measure of KS1 attainment, we replaced this with the individual EYFSP scores obtained by the children participating in the trial. We use the total EYFSP score obtained by the pupils. As noted earlier in the report, although this measure has the advantage of being available at pupil level, it lacks granularity (Dockrell et al., 2017). Nevertheless, it is still of interest to consider, as a robustness check, how the use of this measure affects (or does not affect) the results. The results are reported in Table 10 below; this makes no substantive difference to the results—we still observe a positive but non-significant impact of the programme on the primary outcome. Note this model is estimated on a slightly smaller sample as information on EYFSP scores was not available for all pupils. We also repeated these models using just the maths component of the total EYFSP score, but again the results are effectively unchanged. We also explored including both the total EYFSP score and the lagged school-level measure of KS1 attainment, with these measures entered simultaneously as separate variables in the model (second row of Table 10). The final row of Table 10 reports results after imputing missing values for the total EYFSP score. Missing values are multiply-imputed based on gender, age in months at the start of Year 2, whether ever eligible for FSM, school level KS1 attainment in 2016, treatment arm, and blocking dummies. This was implemented using Stata's mi command, with ten imputations.

Perhaps the most striking finding to note from these various checks is that the effect size appears robust to these different analytical choices.

Table 10: Impact on primary outcome, EYFSP scores used as pre-test

| Outcome | Raw means | | | | Effect size | | |
|--|--------------------|----------------------|----------------|----------------------|--|----------------------|-------------|
| | Intervention group | | Control group | | n in model (intervention; control) | Hedges g (95% CI) | p- value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| PTM – standard age score (EYFSP score used as pre-test) | 3201 (37) | 103.6 (103.1, 104.1) | 3064 (51) | 102.2 (101.7, 102.8) | 6265 (3201; 3064) | 0.08 (-0.03, 0.19) | 0.139 |
| PTM – standard age score (including EYFSP score and lagged KS1 attainment) | 3201 (37) | 103.6 (103.1, 104.1) | 3064 (51) | 102.2 (101.7, 102.8) | 6265 (3201; 3064) | 0.09 (-0.02, 0.20) | 0.095 |
| PTM – standard age score (including EYFSP score and lagged KS1 attainment, with MI) | 3237 (1) | 103.6 (103.0, 104.1) | 3115 | 102.2 (101.6, 102.7) | 6352 (3237; 3115) | 0.09 (-0.01, 0.20) | 0.089 |

Secondly, while our primary analysis is estimated on an intention-to-treat basis, we also checked the robustness of our results to removing those schools that we are aware stopped delivering the programme. We also exclude some Year 2 pupils within a mixed year class in one participating school that did not receive the programme. Again, the results are effectively unchanged (Table 11).

Table 11: Impact on primary outcome, excluding schools that did not deliver programme

| Outcome | Raw means | | | | Effect size | | |
|---------------------------------|--------------------|----------------------|----------------|----------------------|--|----------------------|-------------|
| | Intervention group | | Control group | | n in model (intervention; control) | Hedges g (95% CI) | p- value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| PTM – standard age score | 3144 | 103.6 (103.1, 104.1) | 3115 | 102.2 (101.6, 102.7) | 6259 (3144; 3115) | 0.08 (-0.03, 0.18) | 0.144 |

Thirdly, we tested the sensitivity of our results to an alternative definition of the sample (that is, including those pupils identified in the second round of baseline data collection). Again, this has no substantive effect on the results, with the effect size remaining around the same magnitude and not statistically significant (Table 12).

Table 12: Impact on primary outcome, including pupils identified at second data collection

| Outcome | Raw means | | | | Effect size | | |
|---------------------------------|--------------------|----------------------|----------------|----------------------|--|----------------------|-------------|
| | Intervention group | | Control group | | n in model (intervention; control) | Hedges g (95% CI) | p- value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| PTM – standard age score | 3282 | 103.5 (103.0, 104.1) | 3162 | 102.1 (101.5, 102.6) | 6444 (3282; 3162) | 0.08 (-0.02, 0.19) | 0.131 |

Finally, we check the sensitivity of our results to using the total raw score (which ranged from a minimum score of zero to a maximum of 43) as the outcome measure instead of the standard age score (Table 13). Histograms showing the distribution of the raw score are presented in Appendix G. It should be noted that some moderate degree of negative skew is apparent here, although this is not evident in the standardised scores which we use in the primary analysis. The negative skew may suggest that the assessment was more sensitive to differences at the lower end of the distribution and less sensitive to differences at the higher end of the distribution. However, when using the raw score, the results are effectively unchanged. Note that the results presented in Table 13 do not control for pupil age, however, including this does not have any substantive impact on the results.

Table 13: Impact measured using raw score

| Outcome | Raw means | | | | Effect size | | |
|------------------------|--------------------|----------------------|----------------|----------------------|--|----------------------|-------------|
| | Intervention group | | Control group | | n in model (intervention; control) | Hedges g (95% CI) | p- value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| PTM – raw score | 3238 | 26.81 (26.51, 27.11) | 3115 | 25.98 (25.66, 26.29) | 6353 (3238; 3115) | 0.08 (-0.02, 0.18) | 0.123 |

Exploratory analysis

No secondary outcomes were specified in the trial protocol. However, as an exploratory analysis, we also investigate the four components of the overall PTM score (presented in Table 14 below). Note that since the analyses presented in this section were not specified in the protocol, the results presented below can only be interpreted as suggestive. We hold all other assumptions the same as for our main model—using the same sample and the same pre-test measure. It is important to note that each of the subscores is on a different scale, as these are effectively a score based on the number of questions answered correctly for each of the four areas.¹¹ Histograms for each of the four measures are provided in Appendix G. These may suggest some concerns over ceiling effects for the measures relating to fluency in facts and procedures and fluency in conceptual understanding, however, the same does not apply for the mathematical reasoning and problem solving scores. The results in Table 14 provide some indication that there may be a positive impact of the programme on the mathematical reasoning component of the PTM score.

¹¹ The scales were as follows (minimum–maximum): mathematical reasoning, 0–20; problem solving, 0–4; fluency in facts and procedures: 0–6; fluency in conceptual understanding, 0–13.

Table 14: Impact on components of PTM score, process categories, exploratory analysis

| Outcome | Raw means | | | | Effect size | | |
|--|--------------------|----------------------|----------------|----------------------|--|----------------------|-------------|
| | Intervention group | | Control group | | n in model (intervention; control) | Hedges g (95% CI) | p- value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| Mathematical reasoning | 3238 | 11.95 (11.79, 12.11) | 3115 | 11.48 (11.32, 11.65) | 6353 (3238; 3115) | 0.10 (-0.00, 0.20) | 0.059 |
| Problem solving | 3238 | 1.54 (1.50, 1.58) | 3115 | 1.46 (1.42, 1.50) | 6353 (3238; 3115) | 0.07 (-0.02, 0.16) | 0.127 |
| Fluency in facts and procedures | 3238 | 4.56 (4.51, 4.60) | 3115 | 4.52 (4.47, 4.56) | 6353 (3238; 3115) | 0.00 (-0.10, 0.10) | 0.991 |
| Fluency in conceptual understanding | 3238 | 8.76 (8.65, 8.86) | 3115 | 8.52 (8.41, 8.63) | 6353 (3238; 3115) | 0.06 (-0.04, 0.15) | 0.228 |

We also explored, for the overall PTM score, whether there was evidence of differential impacts according to ability (as proxied by prior levels of attainment). Using EYFSP scores, we classified pupils as having low, medium or high prior attainment. ‘Low’ was defined as being below the expected level on either of the early learning goals relating to maths, ‘medium’ was defined as achieving the expected level on both, and ‘high’ was defined as being at least the expected level on both and above expected on at least one. We then estimated models interacting this measure of prior attainment with the treatment dummy; the results are presented in Appendix H. The results indicate no significant variation in the impact of the programme by prior attainment—as indicated by tests of the equality of the interacted coefficients $F(2, 159) = 1.30$, $p\text{-value} = 0.275$ —although the estimate for pupils in the low prior attainment group is statistically significant at conventional levels. These findings should be treated as suggestive only as this exploratory analysis was not specified in the protocol. It may also be the case, given the negative skew in the distribution of the raw scores, that the assessment was less sensitive to differences in the performance of pupils at the higher end of the distribution.

Exploring the use of responses to the survey

In the protocol we stated that we would explore how the data collected on fidelity through the survey could also be used as part of the impact analysis to investigate any potential moderation of treatment effects.

With this in mind, questions were included in the survey for treatment schools about implementation. However, in practice there are issues with incorporating this information into the analysis.

First of all, while the survey achieved a reasonable response rate of 62.5% among treatment schools, this still leaves a sizeable proportion of schools for whom we have no responses to the survey. As discussed in the IPE section of this report, a total of 50 treatment schools responded to the survey (with 52 responses in total from these schools).

Furthermore, there was very little variation among the schools in the responses given to the question, ‘To what extent did you follow the Mathematical Reasoning programme as outlined by the handbook?’ Respondents were able to respond on a four-point scale; not at all; very little; somewhat; to a great extent. In practice, 47 of the 52 respondents indicated that they followed the programme ‘to a great extent’. Respondents were also asked whether they delivered all units of the programme, and whether this was done within the expected time frame. However, only three respondents stated that they did not deliver all the units.

Respondents were also asked about any changes they had made to elements of the programme. Here there is potentially more scope for analysis as some more variation is apparent. Respondents were asked, 'To what extent did you make changes to the following elements of the programme, as outlined in the handbook?', in relation to six different aspects of the programme: programme structure, use of teacher-led presentations, use of pupil answer sheets, use of extension worksheets, class differentiation, and use of online games. For each aspect, respondents were again asked to respond on the same four-point scale (see Table 24 in the IPE section of this report for the distribution of responses to these questions). We used this information to construct three indicators of change: whether the school made changes to any of the six aspects, a 'total change score', summing together the responses on the four point scale for each of the six aspects, and an 'average change score' based on the average (mean) of the responses given for each of the six aspects. We regress the PTM scores on these measures of change (running three separate models for each change measure), additionally controlling for lagged school-level attainment and the blocking dummies (Table 15 reports the coefficient on the measures of change). The results of this analysis suggest that changes to the programme may be associated with lower overall test scores. However, this relationship is weak. Furthermore, we emphasise that there is no basis for asserting a causal relationship; this particular analysis is only exploratory and can only give an indication of an association.

Table 15: Linear regression of PTM score on measures of change to the programme

| | PTM score | PTM score | PTM score |
|-----------------------------|------------------|------------------|------------------|
| Any change | -1.56 (-0.98) | | |
| Total change score | | -0.12 (-0.44) | |
| Average change score | | | -0.76 (-0.46) |
| N | 2050 | 2050 | 2050 |

Note: t statistics based on school-level clustered standard errors reported in parentheses. Models also control for lagged school-level KS1 attainment and blocking dummies. Statistical significance indicated as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

Cost

We first estimate the cost per pupil per year of participating in the programme. This is based on the direct, marginal costs that are incurred. These include:

- the costs of training (including both the initial training of WGLs and the training of teachers);
- costs of school visit/support from WGLs;
- travel/subsistence for teachers attending training and WGL visits; and
- costs of photocopying.

We present estimates for each of these costs in Table 16. These estimates are generally based on information provided by four of the eight participating Maths Hubs that supplied information on costs of training teachers, school visits and WGL support, and travel/subsistence costs. To obtain an estimate of cost per pupil for these items, we take the average cost per pupil across these four hubs based on the number of schools and pupils in the treatment arm of the trial within these four hubs. In addition, information on the costs of the training of WGLs was provided by the University of Oxford team (covering all WGLs and thus all eight hubs).

Training costs comprised an initial training day for WGLs, a repeat of this day for WGLs unable to attend the initial day, plus two follow-up days. It is worth noting that as the training was provided at Oxford, no costs associated with hiring a venue were incurred. The total cost of £5,634 includes costs of catering (£1,350) and the cost of materials provided to WGLs, including the programme handbook (£228). The remaining £4,056 is an estimate of staff costs incurred. We calculate a cost per school by dividing the total cost of training by the number of schools in the treatment arm of the trial. Similarly, we estimate a cost per pupil by dividing the total cost of training by the number of pupils in the treatment arm of the trial.

Table 16: Cost of delivering Mathematical Reasoning programme

| Item | Type of cost | Cost | Total cost (per school) over 3 years | Total cost per pupil per year over 3 years |
|--|--------------------------|--------|--------------------------------------|--|
| WGL training (and provision of materials) | Start-up cost | £5,634 | £67 | £0.50 |
| One-off teacher training | Start-up cost per school | £302 | £302 | £2.46 |
| School visit by WGL/WGL support | Start-up cost per school | £293 | £293 | £2.07 |
| Travel/subsistence for WGLs and teachers (training and visits) | Start-up cost per school | £41 | £41 | £0.30 |
| Photocopying worksheets | Running cost per pupil | £2.40 | £316.80 | £2.40 |
| IT support costs | Start up cost per school | £28 | £53 | £0.40 |
| Total | | | £1,073 | £8 |

In the previous efficacy trial of the mathematical reasoning programme, Worth et al. (2015) noted that the costs of photocopying associated with the programme were a significant factor mentioned by schools (Worth et al. estimated this cost at £125 per class). Photocopying costs were commonly also reported in this trial as an additional cost incurred (42 of the 52 teachers responding to the survey of treatment schools reported that they had incurred additional costs for photocopying, with six reporting that they had not and the remaining four stating that they did not know or did not answer). WGLs were allocated an allowance of £100 for photocopying when implementing the programme in their schools, and information collected from two of the WGLs indicated that this allowance had been about right for the requirements of the programme. Based on the number of worksheets required, which is estimated to be an average of 40 per pupil, we estimate that the average cost of photocopying per pupil would be £2.40.¹² Given the average size of the year group in this trial was 44 pupils (at the point of randomisation), this would imply a total cost per school per year of around £105.

Finally, the programme also involves costs in terms of IT support, including setting up users with logins for the games, responding to queries, and data maintenance. A total cost for this as delivered in the effectiveness trial was provided by the University of Oxford team, which we divide by the number of schools in the treatment arm to obtain the per school cost presented in Table 16. Part of this cost related to initial setup, but some costs would be incurred each year in which the programme was delivered as a new set of users would need to be set up to play the games each year, and there would be some time required for associated queries and data cleaning.

Based on the information outlined above, we estimate the total cost per pupil per year of implementing the programme to be £19 (Table 17). The running costs of the programme are low (effectively the costs of photocopying and ongoing IT support in relation to the computer games), once the initial set-up costs

¹² Assuming an average cost of 6 pence per page for colour photocopying.

are incurred, such that when considered over a three year period, the estimated cost per pupil per year is around £8. In the previous efficacy trial, Worth et al. (2015) estimated the cost of the intervention at £21 per pupil. This included the cost of training and the school visit, as well as costs associated with implementation. The cost of implementing the programme over three years was estimated at £10 per pupil per year by Worth et al. Given there is inevitably some degree of imprecision in these estimates, this indicates a fairly similar cost of implementation of the programme across both trials.

Table 17: Cumulative costs of Mathematical Reasoning Programme, per pupil

| | Year 1 | Year 2 | Year 3 |
|---|--------|--------|--------|
| Mathematical reasoning programme | £19 | £22 | £25 |

Note these figures have been rounded to the nearest £.

The trial also included the provision of an online forum to provide additional support for participating teachers; the running of such a forum clearly also incurs an additional cost. However, it is unclear whether such a forum would exist if the programme were delivered outside of the evaluation,¹³ and we have therefore not included this within our cost per pupil estimate. While it would not be a cost incurred directly by schools, ultimately the cost of running such a forum, if it formed part of the intervention, would need to be factored into the cost of delivering the programme.

It is also important to consider the additional resources required for implementation of the programme. As noted earlier in the report, an integral part of the programme is the accompanying computer games. Access to computers or tablets is therefore an important pre-requisite for implementing the programme. While many schools already have access to such facilities—meaning this is considered as a prerequisite rather than a marginal cost—providing the necessary access did prove an issue in some of the schools participating in the trial, as discussed in further detail in the implementation and process evaluation section of this report. The ability to have the necessary IT resources in place is therefore an important factor for schools to consider when thinking about implementing the programme. Having the necessary support from teaching assistants is also an important prerequisite—TAs potentially helped children with the practicalities of logging on to the online games, and often assisted with other aspects such as photocopying worksheets. Finally, the lessons also required schools to make use of resources such as dice and counters. It is assumed that most schools would already have access to such equipment and therefore we consider this as a prerequisite rather than a marginal cost, but a small handful of schools did note in responses to the survey that they did not have sufficient supplies.

Costs in terms of staff time are also an important consideration. In line with EEF guidance, we report this here in terms of staff time incurred, rather than converting this to a financial cost, as schools may choose to resource the time required in different ways.

The programme required teachers to attend one day of training. Schools typically sent all Year 2 teachers to attend the training (they were encouraged to do so), and some schools also sent teaching assistants. In addition, teachers were visited by their WGL during the delivery of the programme. While this visit covered the period of a lesson, the delivery team estimated that visits may have required an additional one to one and a half hours for discussion prior to and/or following the lesson. Both the training and potentially the visit would have required schools to provide supply cover, depending on how schools opted to resource this time.

In terms of costs relating to the time taken to deliver the intervention, the programme is delivered as part of the normal school day, during pupils' usual mathematics lessons, and consists of ten teaching units that are anticipated to take approximately 12 to 15 weeks to implement. However, it is important

¹³ Advice from the delivery team indicated that it would depend on how the programme were to be run in future, so we do not treat this as a core component of the programme and hence exclude it from the cost estimate.

to note that around 40% of respondents to the survey indicated that it had taken them longer to deliver all ten units (see also the later discussion in the implementation and process evaluation section; 21 of 52 respondents stated they had delivered all ten units but over a longer time frame, three respondents had not delivered all ten units, and 28 respondents had delivered all units within the expected time frame). The programme may also have created additional time for teachers in terms of lesson preparation. We do not have robust information on how much additional time this would have involved, although estimates of the time required were provided by two of the WGLs, which suggested that preparation was likely to have involved an average of one to two hours per teaching unit. It is not clear how this might have compared with the time teachers would have otherwise have spent on lesson preparation in the absence of participating in the programme. It could also be expected that preparation time would take longer the first time the programme is delivered, but would reduce as teachers become more familiar with delivering the programme over time.

Another important additional resource consideration is the time taken for administration and management within Maths Hubs and the central co-ordination by the NCETM, to facilitate the programme. Estimates suggest that this required around 26 days of support from the project lead at the NCETM with estimates of administrator time in Maths Hubs varying from 5 to 15 days per hub.

Finally, it is important to bear in mind that the costs reported above are the costs incurred based on the effectiveness trial. By way of comparison, in a recent case where the Oxford team trained a group of teachers directly, the total cost of training was £3,168 (this was equivalent to a cost per school of around £211). In this version of the programme, schools were given the opportunity to purchase a box of materials (to cover 32 children), including not just the handbook but also a set of children's workbooks, card games, and log-in information (thus avoiding the need to photocopy worksheets). The cost of this decreases with the number of schools recruited; on the basis of ten schools, the cost of the materials per pupil is around £26; on the basis of 20 schools, the cost per pupil is around £16. Combining the cost of training, materials (based on 20 schools), and IT support would result in a total cost per pupil of around £24. On a three-year basis, this would be equivalent to a cost per pupil of around £20. This is slightly higher than the cost estimated based on the effectiveness trial, but still remains a relatively low cost intervention.

Implementation and process evaluation

The purpose of this implementation and process evaluation was to establish fidelity to the Mathematical Reasoning programme, and to identify factors that may have influenced its impact in treatment schools. More specifically, this implementation and process evaluation used a range of methods and data sources to explore WGLs' and teachers' experiences of the training, to identify experiences of implementation (including any barriers), and to explore potential impacts of the programme through the actions of both treatment and control schools.

Implementation

Training of Work Group Leads

All attendees provided feedback on the training via evaluation forms supplied by the NCETM on the day. Feedback from training reports was positive overall, with the majority of respondents stating that they felt confident using the materials within their own classrooms (Table 18). In response to the question 'What session did you find particularly interesting or useful?', WGLs commonly identified learning about the theoretical underpinnings and research behind the programme as the most useful.

Table 18: How confident do you feel about using these materials with your children?

| Number of responses | |
|--------------------------|----|
| Very confident (1) | 12 |
| (2) | 7 |
| (3) | 1 |
| Not at all confident (4) | 0 |

Number of respondents (N) = 20. Attendees asked to respond on a 4-point scale where 1 is 'very confident' and 4 is 'not at all confident'.

Source: WGL Initial Training Evaluation Survey, February 2016.

Table 19: How well have we supported your understanding of the four strands of the programme?

| Number of responses | |
|---------------------|----|
| Very well (1) | 18 |
| (2) | 2 |
| (3) | 0 |
| Not at all well (4) | 0 |

N = 20. Attendees asked to respond on a 4-point scale where 1 is 'very well' and 4 is 'not at all well'.

Source: WGL Initial Training Evaluation Survey, February 2016.

During this initial training event, conversations with WGLs indicated widespread agreement that the training was of high quality and appropriately made the transition between its theoretical underpinnings and practical application. Participants expressed a high level of engagement both with the trainers and each other. In particular, participants frequently discussed how the programme could be best implemented within their own school environments and with the children within their class. During conversations at the training event, WGLs expressed confidence they could express the concepts behind the programme as well as the more practical application when delivering the training to trial

schools. At this early stage, some WGLs expressed concerns at being able to access the resources required for the project. Having the necessary ICT facilities and support staff were particular concerns.

Feedback on the follow-up on the July training event from WGLs was positive overall. On a scale from 1 to 4 (where 1 is 'very well' and 4 is 'not at all well'), all responding WGLs rated the training at a '1' or '2' in its ability to support their understanding of the four strands of the project. Furthermore, when responding to the questions 'How confident do you feel about presenting the sessions at your training day?' and 'How confident do you feel about carrying out school visits?', all respondents scored their confidence at 1 or 2 (on a scale of 1–4 where 1 is 'very confident' and 4 'not at all confident'). Open-ended responses identified that the majority of WGLs strongly valued the opportunity to work together to go through and annotate the presentations in order to prepare for their own training days.

Training of treatment schools

WGL training days for treatment schools were held throughout September and October 2016. Treatment schools' experiences of training were explored through field notes taken during attendance at three Maths Hubs' training days, interviews with eight teachers implementing the programme, and through the subsequent survey of all treatment schools.

Field notes taken during three training sessions identified a high level of consistency between Maths Hubs; each day followed a programme provided by the NCETM that mirrored WGLs' own training. Each day began with an introduction to the intervention, which included background theory and evidence, followed by a description of the specific concepts addressed in the programme and an explanation of the structure and detail of the teaching units. This was followed by a session to explore the use of online materials (including the games), as well as some discussion regarding the nature of the intervention, including arrangements for visits and testing. During all three training sessions, WGLs stressed the importance of implementing schools following the exact chronology and composition of the programme, and the benefits of engaging with the online learning community. Where teachers were unable to attend, WGLs arranged to visit the schools to deliver the training. No assessment was administered in order to pass the training.

Evaluation forms of the training days showed the majority of teachers considered the training day as very useful (Table 20). Of the 142 respondents who provided a written clarification to this question, 86 respondents identified the practical elements of the training as the most valuable. This included going through the handbook, outlining the structure of the programme, and observing practical demonstrations.

Table 20: How would you rate the usefulness of the day?

| Number of responses | |
|---------------------|----|
| Very useful (5) | 92 |
| (4) | 48 |
| (3) | 9 |
| (2) | 1 |
| Not useful (1) | 1 |

N = 151. Attendees asked to respond on a 5-point scale where 1 is 'not useful' and 5 is 'very useful'. Source: Treatment Schools' Training Evaluation Survey, September/October 2016.

The survey of teachers in treatment schools, administered at the end of the project, also explored experiences of training. Of the 51 respondents who attended training days, 48 stated that the session adequately prepared them to deliver the programme; 26 out of the 51 respondents, however, thought

the training could be improved. Open-ended responses to the survey regularly identified that the inclusion of more practical elements would have improved the training. Examples included more demonstration of the activities and better instruction around the use of the games.

'More time could have been spent looking at the materials in the book and considering how the sessions would be delivered.'

Treatment School [School 13]

Interviews with implementing teachers reinforced this finding. While some teachers said the training had provided them with clear guidance as to the expectations of delivery, others felt that the training day had focused too much on the research and theory underpinning the programme, rather than the practicalities of implementing it in schools.

'They talked to us about the vocabulary, some of the games, and the background behind the project and everything. But it would have been helpful had we had printouts of the actual units so we could have, I don't know, gone to the first one or something, demonstrated that.'

Year 2 Teacher [School 1]

Interviews with teachers, conducted throughout February and March 2017, identified some criticism of the training. Two out of the eight treatment schools interviewed suggested WGLs lacked specialist maths knowledge, resulting in more technical enquiries raised during their training days to be left unanswered. These interviewees' wider responses to the training, however, indicated the experience to still be positive.

'I think what they did was fine but I think maybe there could have been a maths specialist there to be able to answer the maths specialist things.'

Year 2 Teacher [School 2]

Interviews with treatment schools raised some criticism regarding the timing of training days. Held in September and October, some teachers emphasised the lack of time to prepare for implementation of the programme when faced with other commitments of the Autumn term and Christmas holidays. In line with this, responses from the treatment survey showed schools to implement the programme at different times, some up to three months after their initial training (Table 21). Interviews with treatment schools also showed that some took a break part-way through implementing the programme.

'I think we started in October, November and I think we got up to I think number three, unit three or unit four, but then we had to stop because we had nativity rehearsals and things like that ... so then we start it up back again in January.'

Year 2 Teacher [School 6]

Table 21: When did you begin teaching the Mathematical Reasoning programme in your school?

| Number of responses | |
|---------------------|----|
| October 2016 | 13 |
| November 2016 | 21 |
| December 2016 | 2 |
| January 2017 | 15 |
| February 2017 | 1 |

N = 51. Source: Treatment Schools' Survey, July 2017.

Additional support—school visits and use of the online forum

In addition to the training, WGLs provided support to treatment schools through school visits. School visits were recommended by the NCETM to take place at the early stages of the project, approximately at the time of implementing Unit 3. Due to schools implementing the programme at different times, these visits took place between November 2016 and March 2017. Responses to the treatment schools' survey show the majority of treatment schools to rate the support they received during these visits as 'good' (Table 22).

Table 22: How would you rate the support you received from your Work Group Lead during the visit to your school?

| Number of responses | |
|---------------------|----|
| Excellent | 8 |
| Good | 31 |
| Average | 10 |
| Poor | 1 |

N = 50. Source: Treatment Schools' Survey, July 2017.

Findings from the treatment survey suggest that for the majority of schools (38 out of 51), the training day and school visits were the only direct support they received from WGLs. Interviews with implementing teachers indicate this to be largely at their request on an as-needed basis. All teachers interviewed emphasised that there had been open lines of communication with WGLs, and that they knew where to go if they required additional support. Findings from the survey of treatment schools suggest that the majority of the 14 treatment schools that sought additional support from WGLs felt their problems and questions were sufficiently addressed by the WGLs (Table 23).

Table 23: To what extent did the WGL solve your problems / answer your questions?

| Number of responses | |
|---------------------|---|
| To a great extent | 7 |
| Somewhat | 6 |
| Very little | 1 |
| Not at all | 0 |

N = 14. Source: Treatment Schools' Survey, July 2017.

Interviews with treatment schools identified some incidences of non-response from WGLs resulting in uncertainty in some of the fundamental aspects of programme implementation and testing. A lack of clarifying information from WGLs was identified as undermining fidelity to the programme (see section on Fidelity below).

'We tried emailing and stuff and didn't really get anything back. We thought [the WGL] would have come in at the beginning or been like "when are you starting?", but we didn't get anything so ... [enquiry] was just what happens if we don't have a TA or what happens if we don't have computer access and they said on the day that they would do this assessment on [the pupils] and we thought, "well—surely that happens before we start?"'

Year 2 Teacher [School 5]

Interviews with WGLs identified significant effort was made to encourage treatment schools to maintain regular contact and make use of the NCETM forum. Observation of the NCETM forum, however, shows little ongoing activity across the majority of Maths Hubs. Interviews with treatment schools showed awareness of the forum, but little incentive to engage. Reasons included the relatively short nature of the programme, and a preference to discuss and resolve issues within schools internally.

'I know the leaders of this programme have been saying "go on and put an update or a blog"; I must admit, I haven't done that. It's just something I haven't done, or asked any questions particularly. We've just sort of tried to resolve things in school.'

Year 2 Teacher [School 1]

'I vaguely remember them saying something like "if anyone has any problem or anyone wants to share any ideas or anything like that you can go on there" ... I think if it was a project where we were teaching this every single day for say three months then maybe there might be a point where you might want to speak to other schools and say how's it going there.'

Year 2 Teacher [School 5]

Experiences of delivering the intervention

Overall, teachers described the programme design—lesson plans provided for one hour per week to replace the normal mathematics lesson—as straightforward and a positive experience to implement. Interviews with teachers identified the project handbook and pre-prepared PowerPoints as particularly positive elements of the programme.

'The handbook's really, really good. It's good because there's a little set of instructions next to each slide, so if you've had a look through the morning before you teach it and you forget, for example, you've got the book in front of you so you can read out the questions.'

Year 2 Teacher [School 7]

Teachers were also generally positive about the structure of the unit plans—beginning with a teacher-led activity or presentation with answer sheets followed by differentiated worksheets to challenge and scaffold learning where appropriate.

'I mean [the structure has] been nice in a way because we're used to having carpet time, then we send the children off to do their sheets independently and then I'll float around the work table.'

Year 2 Teacher [School 2]

WGLs, interviewed teachers, and respondents to the treatment survey also commented positively about the content of the programme and the concepts it was teaching. Teachers frequently likened the structure of the programme units to a mastery approach already being implemented in their schools that was thought to be helping pupils to develop a secure knowledge and understanding of mathematics; this likely increased initial buy-in from treatment schools.

'We use the mastery approach in our teaching anyway, so everybody does the same and then we have an extension or challenge for those who grasp that really quickly and then a scaffold for those children who need the scaffold. So we very much do that anyway, so it sort of followed that pattern.'

Year 2 Teacher [School 8]

Some treatment schools raised concerns regarding the L1/L2 differentiation, suggesting it had a potentially dis-incentivising impact on the children who were not able to use the computer games. Observation visits identified that pupils consistently considered the games as the more enjoyable task. Pupils, however, showed no obvious signs of disengagement when not selected.

'[They] see the rest of the class line up and go to the computer suite and then they sit in there and they feel like "I don't get that", and even though we explain it to them—"this is not you missing out on the reward, you'll get to go to the computer suite next week"—they still feel like "I've got to stay behind".'

Year 2 Teacher [School 6]

A common criticism identified during interviews and from the treatment survey was the level of preparation required to deliver the programme. In particular, the number of worksheets and answer sheets required—on average 40 sheets per pupil throughout the ten units—was thought to be excessive, putting additional stress on the staff and, in some instances, de-motivating pupils.

'The amount of photocopying is ridiculous and actually disengages pupils as they are unmotivated with worksheet after worksheet.'

Treatment School [School 12]

The need for manipulatives and other practical resources was also identified to be problematic for teachers, requiring a great deal of preparation for sometimes little use.

'There was one unit when every single child had to have twenty counters, so you imagine—twenty times sixty. So [a colleague] and I spent about an hour on a Friday night after PPA just setting out counters and then you put them all out ... the children use them for about two minutes for the first sheet and then you collect them all in again.'

Year 2 Teacher [School 7]

Barriers to implementation

Out of 51 responding treatment schools, 27 stated they had experienced barriers to implementing the Mathematical Reasoning programme. Table 24 identifies the most common barriers to be insufficient time and a lack of IT resources.

Table 24: What barriers have you experienced?

| Number of responses | |
|--------------------------------|----|
| Insufficient time | 17 |
| Lacked access to IT resources | 16 |
| Other teaching priorities | 11 |
| Lack of engagement from pupils | 7 |
| Lack of support staff | 6 |
| Insufficient training | 1 |

N = 25. Source: Treatment Schools' Survey, July 2017

Insufficient time and resources

Interviews with treatment schools consistently identified the opinion that there was too much content to cover within each unit. In response to the treatment survey, 24 out of 52 schools stated that they did not implement the programme in the expected time period (12 to 15 weeks); three of these schools did not implement all ten units by the time of testing. Teachers generally thought each unit took around an hour and a half rather than the expected hour, which was commonly thought to be too long for the age group. As a result, teachers commonly reported splitting the unit across two separate lessons or not undertaking particular tasks.

'Some units required much, much longer than the suggested one hour. We ended up having to miss some of the independent tasks in order to fit the content in before the test!'

Treatment School [School 23]

Making use of the games was identified as a key and consistent barrier to implementation of the project, for a number of reasons. Firstly, not all treatment schools had access to IT resources. Of the 27 treatment schools who stated they experienced barriers to implementing the programme, 16 identified a lack of access to IT resources. Interviews with teachers also identified this to be a problem.

'There's been a couple of weeks where we haven't had a chance to do, to complete it all. There's been maybe one or two weeks when we haven't been able to go onto the computers either because the ICT suite wasn't free or just simply because we ran out of time. There's a lot to fit into the hour.'

Year 2 Teacher [School 3]

In each of the three observed training days a number of treatment schools raised concerns regarding having access to IT facilities. In response to these concerns, WGLs referred treatment schools to the MOU, which committed schools to providing teachers' access to the resources needed to fully implement the project. In instances when this commitment was not expected to be upheld, teachers were also offered the support of WGLs and the NCETM to discuss the matter with senior leadership within their schools.

Use of online games

In addition to lacking access to ICT facilities, treatment schools regularly identified problems regarding more general access to the games. These included issues regarding log-in, identifying the correct

website, and the pupils being able to complete them independently. Lesson observations confirmed some of these difficulties.

'It was difficult to get computer access. The password signing-on and giving the children the web address was difficult. There was confusion with another website called numeracygames.com. Some of the games are difficult to manipulate with a mouse on a laptop because of their fine motor skills, therefore making the games that were timed hard.'

Treatment School [School 21]

Due to difficulties accessing and using the games, a number of schools, either through interview or the treatment survey, identified either partial or complete non-use. As a result, a number of pupils in treatment schools did not have the opportunity to reinforce their learning in this way. Furthermore, there were large inconsistencies between treatment schools as to whether they allowed pupils' access to the online games at home. These inconsistencies, at least in part, occurred as a result of a lack of clarity regarding the monitoring of games use data.

'When we had training they said you could send them home, but then they also talked about monitoring it and keeping a check on which games they're playing and when they're playing them ... I couldn't figure out how I was possibly going to monitor that if it was at home. So I decided—well, we decided in school—that we weren't going to send the passwords home.'

Year 2 Teacher [School 1]

Non-use of games occurred not only due to problems with access. Some responses from the treatment survey identified some negative opinions of the games' aesthetic, and their ability to benefit pupils' learning. These types of responses were frequently made in comparison to other online maths programmes currently being used by treatment schools.

'We felt these were very dated and the children were not inspired to use them. We stopped using them after the first sessions as we didn't feel they added any benefit to the children's learning.'

Treatment School [School 2]

'I think the idea of it, the theory behind it, is fantastic. I think the games themselves could probably be a little bit more updated, a little bit more modern, but that would be my only negative.'

Year 2 Teacher [School 3]

These difficulties and opinions, however, were not universal, with some teachers stating they had no problems accessing and using the games.

'[The games] were quite a nice part of the session, actually, because we can give half the class the laptops and they just get on with it, they know their passwords, they've got little cards. And it's about organisation really, isn't it? We've got all that sorted and that's fine.'

Year 2 Teacher [School 5]

Access to support staff

Interviews, responses from the treatment survey, and lesson observations identified it to be necessary to have support staff to implement the Mathematical Reasoning programme. Teaching assistants, when present, were involved in a lot of preparatory photocopying and resourcing, and were essential in enabling class differentiation and to help pupils as they worked through the activities and the computer

games. A lack of support staff in schools had a direct impact on teachers being able to implement the programme in the way intended.

'[The games] were really tricky as I often lacked a TA and the children were being left to complete these independently, therefore some children found it really tricky to read questions and use the system properly, so at times, we didn't use the online games.'

Treatment School [School 28]

Treatment schools raised the possibility of not having support staff during their training days. WGLs throughout emphasised the important of having support staff in order to best implement the programme. This however was not always possible.

'They hadn't started using the computer games yet as they had IT issues, as well as the headteacher saying that they were not going to be able to give them a TA during the sessions. I did say that it was important that they had a TA there, however, there was still no TA to be supplied.'

[WGL—statement made during review data observation]

Overall, treatment schools emphasised the high—and frequently considered unrealistic—expectations of the programme in terms of resourcing and time-commitment.

'I think the people running it just assumed that we had a TA, that we had access to computers. And we've had a lot of issues with our computer suite, so we haven't done everything we should have done, to be honest, and TA—we both didn't have a TA at the time ... So I think, in terms of what they expected us to have, that probably should have been looked at a bit more.'

Year 2 Teacher [School 6]

Many of the issues identified by this process evaluation regarding barriers to implementation are continuing concerns of the Mathematical Reasoning programme. The findings of the previous efficacy trial, for example, also indicated that the main barriers to implementing the programme as intended were excessive content (both in terms of resourcing and delivery), and issues regarding access and use of online games (Worth et al. 2015). Access to support staff was not identified as a primary barrier to implementation in the previous efficacy trial. Support staff were, however, identified as important in the preparatory work required of the programme.

Fidelity

Overall, the findings of the process evaluation suggest mixed results as to whether the intervention was delivered as intended to all in the treatment group. While the evidence suggests a high degree of fidelity in regard to the implementation of the Unit Plans, other elements of the programme were considered to be either adaptable or unrealistic to implement by treatment schools. The online survey of teachers in treatment schools, administered at the end of the academic year, explored to what extent schools had followed the programme and what adaptations they had made. Table 25 outlines the extent to which teachers made changes to particular elements of the programme.

Table 25: To what extent did you make changes to the following elements of the programme as outlined in the handbook?

| | Not at all | Very little | Somewhat | To a great extent |
|----------------------------------|------------|-------------|----------|-------------------|
| Programme structure (10 units) | 38 | 7 | 5 | 1 |
| Use of teacher-led presentations | 38 | 10 | 2 | 1 |
| Use of pupil answer sheets | 41 | 8 | 1 | 1 |
| Use of extension worksheets | 32 | 10 | 4 | 5 |
| L1 / L2 group split | 27 | 9 | 10 | 5 |
| Use of online games | 28 | 7 | 8 | 7 |

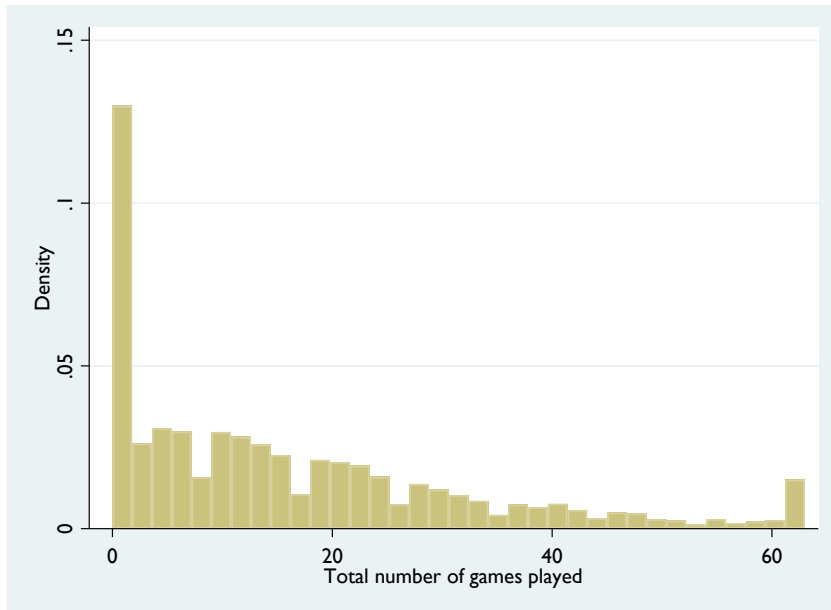
N = 51. Source: Treatment Schools' Survey, July 2017.

Overall, the evidence suggests a high degree of fidelity to the implementation of the core Unit Plans. In response to the treatment survey, 49 out of 52 responding schools stated they had implemented all ten units. Furthermore, in response to the question, 'To what extent did you follow the Mathematical Reasoning programme as outlined in the handbook?', 47 out of 52 responded, 'To a great extent.' Interviews with treatment schools identified that units had replaced one of their normal mathematics lessons, as the design of the trial intended. A common finding across all elements of the process evaluation, however, was that teachers felt it was not possible to deliver units of one hour per week. Interviews with teachers identified that the majority began each session where they had previously left off in the previous lesson, as advised by WGLs. There was, however, also some evidence of teachers not completing units and omitting certain elements of the programme. Variation of this kind was observed during lesson observations, including not using games or extension worksheets, and no differentiation.

The evidence suggests opportunities to play the computer games varied considerably between schools. Key reasons for this variation include lack of time, lack of access, and lack of support staff to facilitate their use.

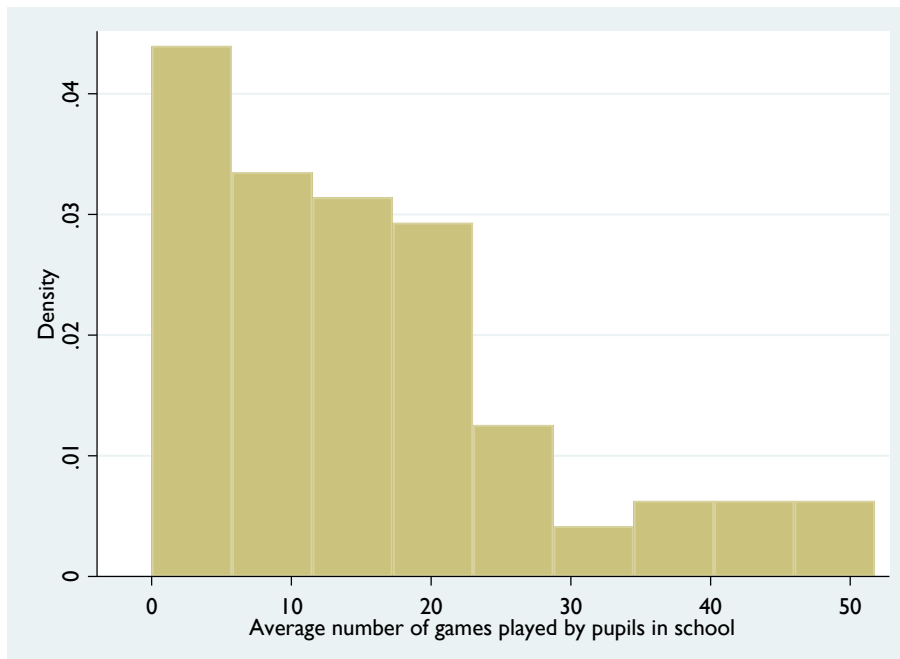
Analysis of game-use data collected by the project team at the University of Oxford shows significant variation in usage. Data was available for 3,755 pupils within 83 treatment schools. Of these pupils, around 21% did not play any games. The mean number of games played was 16, while the median number of games played was 12. Figure 4 presents a histogram of the number of games played per pupil, identifying considerable variation, with some pupils playing more than 60 games.

Figure 4: Total number of games played per pupil



Data on the number of games played per school also identifies considerable variation. In ten of the treatment schools, no games were recorded as being played. In a further three schools, less than ten games were played. At the other extreme, in a few schools more than 2000 games were played. Differences in the total number of games played by school will partly reflect school size; looking at the average number of games played per pupil at school level (Figure 5), we find the mean stands at 15 per pupil (and the median at 13).

Figure 5: Average number of games played by pupils in school



Also, during the process evaluation a number of different approaches to class differentiation were identified other than those intended by treatment model. A number of schools did not split their classes

into L1/L2 groups. This often occurred as a result of a lack of support staff, a lack of time, or due to issues preventing use of the online games.

'It was just easier to keep them all on the same ... because some of them were like playing a game and things when they differentiated it. And we didn't play our games straight after the maths, you see, which is probably where the differentiating came, we didn't have the facility to be able to do that because our ICT slot was a Monday afternoon straight after lunch, so it didn't work. So we did our maths on a different day to the actual games and we had to rethink how we were differentiating the whole lesson.'

Year 2 Teacher [School 5]

Other techniques used included differentiating into higher and lower ability groups that remained static throughout the implementation of the project and differentiating to reflect pupils' changing ability or to ensure fair use of the online games.

Furthermore, when asked to what extent they had made changes to the use of the extension worksheets, 5 of the 51 respondents stated 'to a great extent'. Open-ended responses identified reasons for this to include either running out of time, or a belief that they were too difficult for the children to complete without support. While some omitted their use altogether, others used them as whole-class activities, or as a task for mixed-ability pairs.

Outcomes

The process evaluation explored the perceived outcomes of the intervention among Year 2 teachers implementing the programme. It should be noted that the fieldwork visits to case study schools took place before implementation had finished at around the half-way point of the intervention, while the survey was answered at the end of the project.

Perceived impact on pupil learning and engagement

Overall, teachers generally considered the programme to promote pupils' reasoning and problem-solving skills and helped them to improve their ability to demonstrate their thinking (Table 26). These findings were reinforced by interviews with teachers who frequently suggested an increase in pupils' confidence to participate in class, and improvement in their reasoning.

'I think they've got much better at reasoning, I do because they were linking things together and linking their learning and just getting better at explaining and more confident.'

Year 2 Teacher [School 6]

Table 26: To what extent do you consider the programme has promoted pupils' mathematical reasoning and problem-solving skills?

| | Not at all | Very little | Somewhat | To a great extent |
|---|------------|-------------|----------|-------------------|
| Understanding of numbers and number sense | 2 | 9 | 31 | 9 |
| Reasoning and problem solving skills | 1 | 7 | 31 | 12 |
| Flexible use of language | 1 | 10 | 31 | 9 |
| Ability to demonstrate their thinking | 1 | 3 | 28 | 19 |

N = 51. Source: Treatment Schools' Survey, July 2017.

Interviews with teachers and responses to the treatment survey identified the opinion that the programme best supported lower ability children. Some teachers felt that the programme was too easy for the higher ability children who already had developed strong reasoning skills. This was often associated to other programmes having been, or currently being, implemented. In the treatment survey, 41 schools described implementing a mastery type curriculum in the same year as implementing the Mathematical Reasoning programme; 19 specifically identified White Rose Maths. In line with this, teachers commonly identified a lack of engagement amongst their highest ability pupils who could easily explain their answer during activities if they did not find the content challenging.

'Their reasoning, I think, was already quite good because of the White Rose and that's why I think this EEF project would have been fantastic in maybe other schools that I taught in or other classes, but I think this school was already in a quite strong place because of the White Rose and we were already doing a lot of reasoning.'

Year 2 Teacher [School 1]

'I think the White Rose is just ... it's so good and it's so tight and it is just that little bit more challenging and it covers everything in the curriculum.'

Year 2 Teacher [School 8]

Teachers commonly considered the programme to best support their lower ability pupils in helping them develop their thinking and give them the opportunity to learn from the verbal reasoning of their peers. Some teachers, however, thought the programme was completely inaccessible to their lowest ability pupils, or those with certain special educational needs, which resulted in particular adaptation of the programme materials.

'They couldn't access it. ... Well, I'm not letting four children sit there and not access anything for an hour and a half. They'd get bored, they become disruptive and so on. And they're not learning. So I'd far rather they take the part of it that they could do and could benefit from and just do that than not do anything at all.'

Year 2 Teacher [School 2]

During lesson observations the vast majority of pupils could be seen to be attentive and engaged and contributing their own ideas and rationalisations to group discussions. During one observation one pupil was observed to wholly disengaged with the lesson. Discussion with the teacher and support staff identified the pupil to have special educational needs.

As shown in Table 27, the majority of treatment schools thought the Mathematical Reasoning programme would likely offer some improvement to pupil attainment. During the case study visits, most interviewees were hesitant in concluding that the programme would lead to improvements in pupil attainment within the relatively short time. Some also reasoned it would be hard to disentangle from other initiatives, such as White Rose Maths, or the general upward trajectory of the school.

Table 27: To what extent do you think the Mathematical Reasoning programme is likely to improve pupil attainment?

| | Number of responses |
|-------------------|---------------------|
| To a great extent | 9 |
| Somewhat | 31 |
| Very little | 10 |
| Not at all | 1 |

N = 51. Source: Treatment Schools' Survey, July 2017.

Impact of classroom practice

Teachers offered mixed opinion as to whether they thought the programme had improved their own practice (Table 28). A common response in interviews and the treatment survey was that the programme worked against usual school and classroom practice, particularly regarding the heavy use of worksheets and the amount of time dedicated to teacher-led demonstration and instruction.

Table 28: To what extent do you think the programme has improved your own practice?

| | Number of responses |
|-------------------|---------------------|
| To a great extent | 3 |
| Somewhat | 32 |
| Very little | 15 |
| Not at all | 3 |

N = 51. Source: Treatment Schools' Survey, July 2017.

Nonetheless, a number of teachers thought the programme enabled them to develop their questioning skills, and to elicit more meaningful responses from the pupils to monitor their learning.

'It has made me develop my questioning skills, allowing the children to think more deeply about their explanations, showing me more of an insight into their understanding. The explanations I am receiving now, from some children, are much developed and refined.'

Treatment School [School 9]

When asked if they would continue to use the Mathematical Reasoning programme in the future, the majority of teachers said 'yes', but with some adaptations (Table 29). Adaptations were commonly thought of to minimise the content, and to make use of particular activities that were thought the most useful in supporting pupils' reasoning skills. A common response from case study visits included:

'I would be reluctant to use it as it is next year, but certainly to use the resources, now we know what's there, and to use them within what we do. Because I think that the skills they get from it in terms of reasoning is really good. So I think we still need to be teaching that.'

Year 2 Teacher [School 2]

Table 29: Do you think you will continue to use the Mathematical Reasoning programme in the future?

| Number of responses | |
|--|----|
| Yes—implementing the programme as outlined in the handbook | 9 |
| Yes—but with some adaptations to the programme as outlined in the handbook | 33 |
| No | 10 |

N = 52. Source: Treatment Schools' Survey, July 2017

The process evaluation did not identify any perceived unintended consequences or negative impacts.

Formative findings

This process evaluation identifies a number of ways the programme could be improved which, in large part, reflect those identified in the earlier efficacy trial (Worth et al., 2015). Generally, while teachers who took part in this trial were positive about the content of the units, the quantity of the materials proved problematic. This resulted in most teachers being unable to complete the units in the time provided. We therefore reinforce the finding of Worth et al. (2015) to suggest that teachers should be allowed to deliver the units in whatever way is most suitable for their own particular situations—that is, one unit per week, without the time restriction.

Furthermore, the programme should consider the implications of implementing the 'train the trainers' model and take steps to actively demonstrate the expertise of the trainers. Although the 'train the trainers' model worked well in ensuring teachers were familiar with the broad structure and the delivery of the ten core units of the programme, better emphasising the importance of the other elements (such as differentiation and use of online games) in terms of challenging and scaffolding learning would seem beneficial. It would also seem beneficial to introduce more practical demonstrations into the training, to allow teachers the opportunity to consider any potential practical barriers to implementation, and think of ways to overcome them. Finally, increasing opportunities for discussion in addition to the online forum would also seem helpful to ensure ongoing communication between WGLs and schools beyond the training day and single school visit, and to allow further opportunity to address variation in implementation.

Other suggestions include:

- reducing the number of worksheets and the amount of photocopying;
- reducing the length of whole-class activities to maintain engagement across ability groups;
- updating the format of the PowerPoint presentations and games to make them more appealing to pupils;
- resolving issues regarding access to the online games; and

- ensuring schools are committed to providing access to IT facilities and providing support staff.

Control group activity

An online survey was administered to schools allocated to the control group, coinciding with the end of the programme in treatment schools. Its purpose was to gather information on what activity had been undertaken in relation to the concepts underlying the Mathematical Reasoning programme. The control group survey was completed by 59 schools, equivalent to 73.8% of control schools that remained in the trial (N = 80 schools).

Control schools were asked whether they had used any materials or resources that support children's reasoning in mathematics during the time of the trial. These could include a published scheme, resources from other sources, an intervention programme or a range of different approaches including following a 'teaching for mastery' curriculum model. 56 out of 57 schools stated that they had; 18 of these schools referred to White Rose Maths. A number of other programmes were described, including Maths No Problem, Abacus, and other more general mastery approaches. As shown by the activity of treatment schools through this process evaluation, the implementation of such programmes is commonplace within Year 2 in primary schools in the study and therefore can be considered business as usual. Yet in the efficacy trial, Worth et al. report that around 20% of schools in the control group had introduced a new scheme or change that could affect numeracy since the point of the pre-test. This is a somewhat different measure (it does not capture whether schools had such a scheme that they had implemented prior to the pre-test), however, this difference could potentially provide some explanation as to why a greater impact of the programme was observed in the efficacy trial.

Finally, control schools were asked whether they had participated in any interventions specifically aimed at improving maths reasoning. Four stated that they had, identifying a programme called 1stClass@Number from around September 2016 onwards. This programme is aimed at pupils who are having difficulties with mathematics and involves training teaching assistants to deliver specific, intervention-style sessions to small groups of children, outside usual mathematics lessons.

Conclusion

Key conclusions

1. Pupils who received Mathematical Reasoning made the equivalent of one additional month's progress in maths, on average, compared to children who did not. This result has high security.
2. Among pupils eligible for free school meals, those who received Mathematical Reasoning made an average of one additional month's progress compared to those who did not. This result may have lower security than the overall finding because of the smaller number of pupils.
3. There was some evidence that the programme also had a positive impact on mathematical reasoning.
4. The intervention was generally well received by schools. Teachers reported positive experiences with the training and materials, and were positive about the programme's focus on fundamental mathematical principles.
5. The process evaluation found that there was some variation in how schools implemented aspects of the programme, particularly in relation to the use of the online games.

Interpretation

This effectiveness trial set out to identify the effect of the Mathematical Reasoning programme on pupils' attainment in maths at the end of Year 2. The results of the impact evaluation indicate a small positive, but non-significant impact of the programme on attainment in maths as measured by the standard age score on the Progress Test in Maths. The estimated effect size is 0.08. Since this is below the MDES, the result should be interpreted with caution as the trial was not powered to securely detect effects below 0.14. This result is robust to a number of sensitivity checks based on alternative model specifications.

The findings for the subgroup of children eligible for FSM were similar to the results for all children. The effect size was of a similar magnitude, and again, as this was below the MDES, it should not be concluded that there is zero effect for children eligible for FSM but rather that the effect was not large enough to be detected by the trial.

Exploratory analyses of the different components of the overall Progress Test in Maths score suggest the programme may have had a positive impact on some aspects of mathematical skills. The results are indicative of a positive effect on the mathematical reasoning subscore. Given the findings of the process evaluation regarding teachers' views that the programme was less suitable for the highest and lowest ability learners, we also conducted exploratory analyses to investigate whether there were differential impacts by prior attainment. This suggested no significant variation in the impact of the programme by prior attainment, although the estimate for pupils in the low prior attainment group is statistically significant at conventional levels. It is perhaps also worth noting that we also found some indication that pupils of lower prior attainment were more likely to drop out from the analysis sample and thus we should be cautious in drawing definitive conclusions about impacts for this group. Furthermore, given the negative skew apparent in the distribution of the raw scores, it may be the case that the assessment was more sensitive to differences in performance at the lower end of the distribution.

The previous efficacy trial did identify a positive and larger effect of the programme. This may raise questions about the extent to which the 'train the trainer' model may have potentially diluted the impact of the programme. However, not all other factors were unchanged between the efficacy and the effectiveness trial. In this effectiveness trial, the primary outcome of attainment in maths was measured through an updated version of the Progress Test in Maths, and rather than using the same measure as

a pre-test, instead used measures taken from the NPD (a pre-test that uses the same outcome measure as at post-test is most likely to be successful in capturing variance). The characteristics of participating schools may also differ between the two trials; for example, the schools in this trial had, on average, higher proportions of pupils eligible for free school meals compared with the efficacy trial (around 25% compared with around 10% for the intervention group in Worth et al.), although the average proportion of pupils with SEN was lower. It is also worth noting that in the efficacy trial there was no school-level drop out, whereas in this effectiveness trial, 7 of the 84 intervention schools withdrew from delivering the programme. However, our results are effectively unchanged when we exclude these schools from our analysis. In addition, it is also important to consider activity, or business as usual, among control group schools. The process evaluation suggested that many control schools had used other materials or resources to support children's reasoning in maths. While a precise comparison with the efficacy trial is not possible, the findings of Worth et al. (2015) suggest that such behaviour was less common among control schools in the efficacy trial. This difference could also potentially provide some explanation as to why a greater impact of the programme was observed in the efficacy trial.

The findings from the process evaluation have also demonstrated the variation across schools in the extent to which the programme was delivered as intended in some elements (notably the games and differentiation). It may well be the case that positive impacts were confined to those schools where the programme was implemented more faithfully. Exploratory investigations have provided no strong evidence that changes made to the programme were associated with lower test scores.

Limitations

One potential source of imprecision and bias arises due to attrition. While the trial was successful in recruiting, and therefore the eventual sample size is in line with expectations at the design stage, attrition stood at around 14%. However, attrition was fairly evenly split among treatment and control groups. Furthermore, comparison of school and pupil characteristics by treatment arm indicates that the groups remained balanced, at least in terms of observable characteristics.

The protocol for the trial clearly identifies our primary specification and thus reduces any concerns about multiplicity of analysis. Although we undertake a number of sensitivity analyses, none of these have a substantive impact on the results. As discussed earlier in this report, using a school-level measure of prior attainment has both advantages and disadvantages; nevertheless, our results are effectively unchanged when we replace this with a pupil-level measure of prior attainment, although we acknowledge that neither is a perfect measure. We have also discussed the fact that some degree of negative skew is apparent in the raw Progress Test in Maths scores; however, our primary analysis makes use of the standardised score, for which the distribution is not skewed.

It should be noted that some concerns were raised during the testing for the evaluation, focusing particularly on the appropriateness of such testing for children of a relatively young age. The assessment is designed to be used with the age group in question and is routinely used in schools. The concerns were investigated by all parties involved and all schools were contacted to provide an opportunity to air any issues and to ensure lessons were learned for future evaluations. In terms of interpreting the results regarding the impact of the programme, we have no reason to believe that the prevalence of such issues differed systematically between the treatment and control groups.

As this is an effectiveness trial, the evaluation involved a relatively large number of schools, spread across a range of geographical areas. Nevertheless, we cannot assume that the schools participating in the trial are representative of all schools, or all pupils, especially given the indications that certain groups of pupils were more likely to be lost from the eventual analysis sample. In interpreting the results of the process evaluation, it is also important to bear in mind that while schools are selected for case study visits with the intention of reflecting a range of different circumstances, their experiences will not necessarily be representative of all schools. In addition, while reasonable response rates were achieved

for the online surveys, we cannot rule out the possibility of a bias in terms of which schools respond (although it is not clear *a priori* in which direction any such bias may operate).

Finally, it is also important to note that information on costs was not available from all Maths Hubs. The information on cost should be considered as an estimate; however, even allowing for a degree of variation, the overall cost per pupil is likely to remain very low.

Future research and publications

While the trial was not able to identify a significant impact on attainment in maths as measured by the Progress Test in Maths, it would be valuable for future research to explore whether the programme had an impact on pupils' attainment in maths as captured by KS1 assessments, and in time, it would also be valuable to investigate whether any impact on performance in maths at KS2 is apparent.

While a number of factors may have contributed to the difference in findings between the efficacy and effectiveness trial, this nevertheless raises questions about how the programme can be delivered at scale, if its effectiveness relies on proximity to the original developers. This may warrant further research into how the programme may be delivered in future in order to retain positive benefits for pupils while being practical to deliver on a larger scale. Further research could also include exploration of whether particular aspects of the programme (for example, the use of the online games) are key to determining its impact on maths attainment.

It is the intention of the project and evaluation teams to seek to publish the findings from the evaluation.

References

- Department for Education (2016) 'Schools, pupils and their characteristics: January 2016', SFR20/2016, Department for Education: <https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2016>
- Dockrell, J., Llauro, A., Hurry, J., Cowan, R., Flouri, E. and Dawson, A. (2017) 'Review of assessment measures in the early years. Language and literacy, numeracy, and social emotional development and mental health', London: EEF.
- Ebbes, P., Bockenholt, U. and Wedel, M. (2004) 'Regressor and random-effects dependencies in multilevel models', *Statistica Neerlandica*, 58 (2), pp.161–178.
- 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, pp.147–166. DOI: 10.1348/026151006X153127
- Nunes, T., Bryant, P., Barros, R. and Sylva, K. (2011) 'The relative importance of two different mathematical abilities to mathematical achievement', *British Journal of Educational Psychology*, 82, pp. 136–156.
- Nunes, T. and Bryant, P. (2015) 'Mathematical Reasoning: a Programme for Year 2 Pupil', University of Oxford.
- Worth, J., Sizmur, J., Ager, R. and Styles, B. (2015) 'Improving numeracy and literacy', Evaluation Report, London: EEF.

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

| Cost rating | Description |
|-------------|---|
| £ £ £ £ £ | <i>Very low:</i> less than £80 per pupil per year. |
| £ £ £ £ £ | <i>Low:</i> up to about £200 per pupil per year. |
| £ £ £ £ £ | <i>Moderate:</i> up to about £700 per pupil per year. |
| £ £ £ £ £ | <i>High:</i> up to £1,200 per pupil per year. |
| £ £ £ £ £ | <i>Very high:</i> over £1,200 per pupil per year. |

Appendix B: Security classification of trial findings

| Rating | Criteria for rating | | | Initial score | Adjust | Final score |
|--------|---|------------|------------|---------------|--|-------------|
| | Design | Power | Attrition* | | | |
| 5 | Well conducted experimental design with appropriate analysis | MDES < 0.2 | 0-10% | | | |
| 4 | Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity | MDES < 0.3 | 11-20% | 4 | Adjustment for Balance [] | 4 |
| 3 | Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity | MDES < 0.4 | 21-30% | | | |
| 2 | Weakly matched comparison or experimental design with major flaws | MDES < 0.5 | 31-40% | | Adjustment for threats to internal validity [] | |
| 1 | Comparison group with poor or no matching (E.g. volunteer versus others) | MDES < 0.6 | 41-50% | | | |
| 0 | No comparator | MDES > 0.6 | >50% | | | |

- **Initial padlock score:** This was a well-conducted trial with adequate power to detect an effect size of 0.12 at randomisation. However, attrition was recorded at 14.3% which does reduce the security of the findings = 4 padlocks
- **Reason for adjustment for balance** (if made): none made
- **Reason for adjustment for threats to validity** (if made): none made
- **Final padlock score:** initial score adjusted for balance and internal validity = 4 padlocks

Appendix C: Recruitment materials

C.1 School information sheet and memorandum of understanding

UNIVERSITY OF OXFORD
DEPARTMENT OF EDUCATION

Professor Terezinha Nunes
15 Norham Gardens, Oxford OX2 6PY
Tel: +44(0)1865 284893
www.education.ox.ac.uk



Mathematical Reasoning: A programme for Year 2 children

Information sheet (for schools)

The Education Endowment Foundation (EEF) is supporting a project to roll out a mathematics reasoning programme designed for Year 2 pupils. This programme was found effective in improving children's mathematics achievement in a previous randomised control trial led by a team from Oxford University. It is now being extended in a larger scale study by Oxford University in partnership with the NCETM (National Centre for Excellence in Teaching Mathematics) and the NIESR (National Institute of Economic and Social Research). The programme is a 12 week intervention and is implemented by teachers in their classrooms.

Because this is a research project, schools will be randomly assigned to use the Mathematics Reasoning programme either in 2016 or in 2017. The research team from Oxford University will train Work Group Leads (WGL) from the Maths Hubs who will, with support from the NCETM, train teachers in each participating school to use the programme.

Participating schools will receive training free of charge and a Teacher Handbook to support programme implementation. Schools will need to have use of computers or tablets for the pupils and access to the internet.

In line with much of the work supported by the EEF, the evaluation of the programme will be carried out by randomly assigning schools to participate in the Mathematics Reasoning intervention in 2016 or to a waiting-list group. The waiting-list group will be able to participate in equivalent training in 2017.

The project involves a partnership between the University of Oxford, the NCETM and the EEF, which is providing the resources for the project. The team from NIESR are responsible for the independent evaluation of the programme. The project has been approved by the University of Oxford Ethics Committee. If you require clarification of the ethical approval or have any concerns during the course of the research, please contact the Chair of Department of Education Research Ethics Committee, Dr Liam Gearon (liam.gearon@education.ox.ac.uk).

Please keep this information sheet for your records.



Memorandum of Understanding regarding the project

Mathematical Reasoning: A programme for Year 2 children

This agreement is between the School named below, the University of Oxford, Department of Education, and the National Centre for Excellence in Teaching Mathematics, about a randomised control trial of an intervention to improve children's mathematical reasoning, funded by the Education Endowment Foundation and evaluated by a team from the National Institute of Economic and Social Research.

| | |
|----------------|----------|
| Name of School | |
| Address | Postcode |
| Head Teacher | |
| Telephone | e-mail |

The School

- The School understands that it will be randomly allocated either to the group that participates in the mathematical reasoning programme in 2016 or to a waiting list group that will receive the programme in 2017; it commits to full participation in either group.
- All Year 2 pupils will be eligible to participate in the project.
- The School will seek permission from the pupil's parents for sharing the unique pupil number (UPN) and for the data to be shared with the evaluation team (NIESR); the School will provide the data of consenting pupils to NIESR when requested.
- The School will provide access to the researchers from NatCen Social Research (NatCen) to administer an assessment to the participating pupils on behalf of NIESR.
- If the School is allocated to the mathematical reasoning programme in 2016, the school will nominate at least one Year 2 teacher to attend the training session and to manage and deliver the intervention in accordance with guidance from WGLs and the NCETM.
- The School will ensure that children have access to a computer/tablet with internet access in order to play the computer games that are part of the programme.
- The School will communicate fully and promptly with the research and the evaluation teams, share appropriate data and ensure that questionnaires and surveys are completed and returned.
- The School will facilitate visits to the school by WGLs to support the implementation of the intervention, and by the evaluation team to observe and interview staff during 2016-17 and to administer a post-test to participating pupils in 2017.
- The School understands that it will receive intervention training, resources and support free of charge in return for complete participation in the trial as set out in this agreement.

Oxford University and NCETM

- OU will provide training and an intervention training programme for the Work Group Leads (WGL) from the Maths Hubs to deliver to teachers:
 - if the evaluation team allocates the School to the Mathematical Reasoning Programme in 2016, training will be in September/October 2016;
 - if the evaluation team allocates the School to 2017 group, training will be in September/October 2017.
- The NCETM through Maths Hubs will inform the School of the venue and dates of its training programme when the evaluation team has made the allocation.
- OU will provide a free copy of the Teacher Handbook to each participating teacher to support programme implementation and access for the children to play the programme’s computer games.
- OU and NCETM will provide guidance to the WGLs required for the delivery of the mathematical reasoning intervention; WGLs will provide training for the teachers.
- NCETM will provide an online community for schools during the life of the project, supported by WGLs; OU will provide telephone and email support for when teachers have queries about access to the project website (until March 2018).
- NCETM, in association with the NIESR, will ensure that all visitors to the School, for any purpose in connection with this project, hold a valid Disclosure and Barring Service certificate.

Signed for and on behalf of:

School

Signed

Name

(print)

Position

Date

OXFORD UNIVERSITY

Signed

Name

Terezinha Nunes

Date

***Please return a signed copy of this agreement to:
Name and address of Maths Hub contact***

C.1 Letter to parents and opt-out consent form

UNIVERSITY OF OXFORD DEPARTMENT OF EDUCATION

Professor Terezinha Nunes
15 Norham Gardens, Oxford OX2 6PY
Tel: +44(0)1865 284893
www.education.ox.ac.uk



Dear Parent

I am writing to let you know about an exciting project in which your child's school is participating. Some Year 2 teachers in a number of schools will be using this year a Mathematical Reasoning Programme, designed by a team from the University of Oxford to improve children's attainment in mathematics in Key Stage 1. This programme was found effective in previous studies when the teachers were supported directly by a team from the University of Oxford. After only 12 lessons, participating pupils improved an extra three months in mathematics achievement during the year. The current project will evaluate whether the programme continues to be effective when the teachers receive support from the Maths Hubs, working together with the National Centre for Excellence in the Teaching of Mathematics (NCETM), and the support from the University of Oxford is only indirect. The project is funded by the Education Endowment Foundation (EEF).

In order to evaluate the programme, some of the teachers will use the materials this year and others will use them next year. This will allow for a comparison between the groups that used the materials and those that did not use them this year. In line with the EEF guidelines, all the schools will have the opportunity to use the materials in the end. The assignment to this year or to next year will be done randomly by the independent evaluators, a research team from the National Institute of Economic and Social Research (NIESR).

At the end of the project, the children in all participating schools will answer an assessment that will be used to evaluate the programme. It takes approximately an hour to complete. This assessment does not influence your child's placement in school. It will be used only for the research. The assessments will be collected by NatCen (National Centre for Social research) and the results will only be accessible to the researchers in the NIESR, NatCen and Oxford University teams working in the project. No information about individual children will be made available to anyone outside the research teams. The data will be kept confidential, in accordance with the Data Protection Act. Only group results of the programme evaluation will be published. We will not use your child's name or the name of the school in any report arising from the research.

We are asking for your permission to obtain your child's name, date of birth, and UPN (Unique Pupil Number; please, see attached form), in order to complement the information for the assessment of the Mathematical Reasoning Programme. Once this information is included in the data set, the data will be anonymised and no one will be able to identify individual children. Information provided will be linked with the National Pupil Database (held by the Department for Education) and shared with the project teams, the Department for Education, Education Endowment Foundation (EEF), EEF's data contractor FFT Education and in an anonymised form to the UK Data Archive.

If you would like more information about this project, you can contact the principal researcher, Prof Terezinha Nunes (terezinha.nunes@education.ox.ac.uk). The project has been approved by the University of Oxford Ethics Committee. If you require clarification of the ethical approval or have any concerns during the course of the research, please contact the Chair of Department of Education Research Ethics Committee, Dr Liam Gearon (liam.gearon@education.ox.ac.uk).

Kind regards,

Terezinha Nunes

Mathematical Reasoning: A programme for Year 2 children

If you agree to your child's data and UPN being used in the project, you do not need to do anything.

If you **do not** wish to release your child's Unique Pupil Number (UPN), please tick the box below.

I DO NOT consent for my child's Unique Pupil Number to be released to the research team.

Child's name (BLOCK CAPITALS) Child's teacher.....

Parent name (BLOCK CAPITALS)

Parent signature:

Date

(Please return the completed form to your child's class teacher.)

Appendix D: Randomisation code

```

set seed 3636378

* Identify schools in top half of their hub in respect of FSM proportion
gen double FSMRand=uniform()
sort hub pnumfsmever FSMRand
by hub: gen FSMhiSort=_n>_N/2

* Create hub and FSM blocks
gen hubFSMblock=.
replace hubFSMblock=1 if hub==1 & FSMhiSort==0
replace hubFSMblock=2 if hub==1 & FSMhiSort==1
replace hubFSMblock=3 if hub==2 & FSMhiSort==0
replace hubFSMblock=4 if hub==2 & FSMhiSort==1
replace hubFSMblock=5 if hub==3 & FSMhiSort==0
replace hubFSMblock=6 if hub==3 & FSMhiSort==1
replace hubFSMblock=7 if hub==4 & FSMhiSort==0
replace hubFSMblock=8 if hub==4 & FSMhiSort==1
replace hubFSMblock=9 if hub==5 & FSMhiSort==0
replace hubFSMblock=10 if hub==5 & FSMhiSort==1
replace hubFSMblock=11 if hub==6 & FSMhiSort==0
replace hubFSMblock=12 if hub==6 & FSMhiSort==1
replace hubFSMblock=13 if hub==7 & FSMhiSort==0
replace hubFSMblock=14 if hub==7 & FSMhiSort==1
replace hubFSMblock=15 if hub==8 & FSMhiSort==0
replace hubFSMblock=16 if hub==8 & FSMhiSort==1

* Identify schools in top half of block in respect of KS1 maths
gen double KS1Sort=uniform()
sort hubFSMblock xmath2p1 xmath13 xmath14 KS1Sort
by hubFSMblock: gen KS1hiSort=_n>_N/2

* Create hub, FSM and KS1 blocks
gen hubFSMKS1block=.
replace hubFSMKS1block=1 if hubFSMblock==1 & KS1hiSort==0
replace hubFSMKS1block=2 if hubFSMblock==1 & KS1hiSort==1
replace hubFSMKS1block=3 if hubFSMblock==2 & KS1hiSort==0
replace hubFSMKS1block=4 if hubFSMblock==2 & KS1hiSort==1
replace hubFSMKS1block=5 if hubFSMblock==3 & KS1hiSort==0
replace hubFSMKS1block=6 if hubFSMblock==3 & KS1hiSort==1
replace hubFSMKS1block=7 if hubFSMblock==4 & KS1hiSort==0
replace hubFSMKS1block=8 if hubFSMblock==4 & KS1hiSort==1
replace hubFSMKS1block=9 if hubFSMblock==5 & KS1hiSort==0
replace hubFSMKS1block=10 if hubFSMblock==5 & KS1hiSort==1
replace hubFSMKS1block=11 if hubFSMblock==6 & KS1hiSort==0
replace hubFSMKS1block=12 if hubFSMblock==6 & KS1hiSort==1
replace hubFSMKS1block=13 if hubFSMblock==7 & KS1hiSort==0
replace hubFSMKS1block=14 if hubFSMblock==7 & KS1hiSort==1
replace hubFSMKS1block=15 if hubFSMblock==8 & KS1hiSort==0
replace hubFSMKS1block=16 if hubFSMblock==8 & KS1hiSort==1
replace hubFSMKS1block=17 if hubFSMblock==9 & KS1hiSort==0
replace hubFSMKS1block=18 if hubFSMblock==9 & KS1hiSort==1
replace hubFSMKS1block=19 if hubFSMblock==10 & KS1hiSort==0
replace hubFSMKS1block=20 if hubFSMblock==10 & KS1hiSort==1
replace hubFSMKS1block=21 if hubFSMblock==11 & KS1hiSort==0
replace hubFSMKS1block=22 if hubFSMblock==11 & KS1hiSort==1
replace hubFSMKS1block=23 if hubFSMblock==12 & KS1hiSort==0
replace hubFSMKS1block=24 if hubFSMblock==12 & KS1hiSort==1
replace hubFSMKS1block=25 if hubFSMblock==13 & KS1hiSort==0

```

```
replace hubFSMKS1block=26 if hubFSMblock==13 & KS1hiSort==1
replace hubFSMKS1block=27 if hubFSMblock==14 & KS1hiSort==0
replace hubFSMKS1block=28 if hubFSMblock==14 & KS1hiSort==1
replace hubFSMKS1block=29 if hubFSMblock==15 & KS1hiSort==0
replace hubFSMKS1block=30 if hubFSMblock==15 & KS1hiSort==1
replace hubFSMKS1block=31 if hubFSMblock==16 & KS1hiSort==0
replace hubFSMKS1block=32 if hubFSMblock==16 & KS1hiSort==1
rename hubFSMKS1block block
```

```
* Randomise
gen randSeq=uniform()
sort block randSeq
* Randomise T
gen T=randSeq>.5
* For 2nd school onwards, alternate value of T
replace T=1-T[_n-1] if _n>1
lab def T 0 "Control" 1 "Treated"
lab val T T
lab var T "Treated"
```


Appendix E: Histograms of pre-test scores

The figures below present the distribution of EYFSP scores in the treatment and control groups, both at baseline and for the analysis sample. In both cases, these show a similar distribution of scores across treatment and control groups, giving us further confidence that the two groups are well-balanced in terms of pupils' prior attainment.

Figure E.1: Total EYFSP scores, treatment and control group, baseline comparison

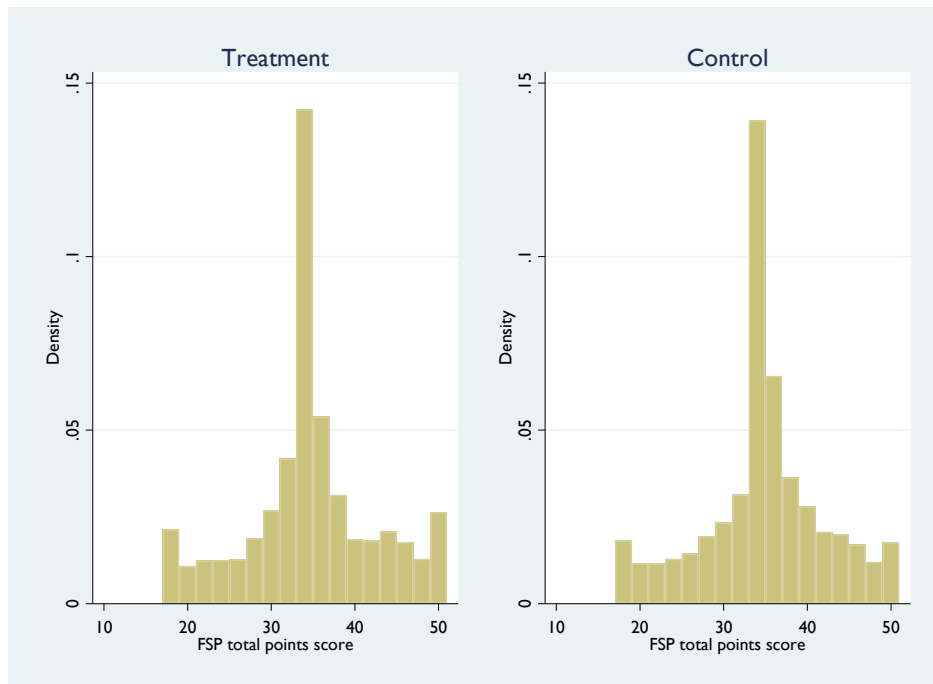
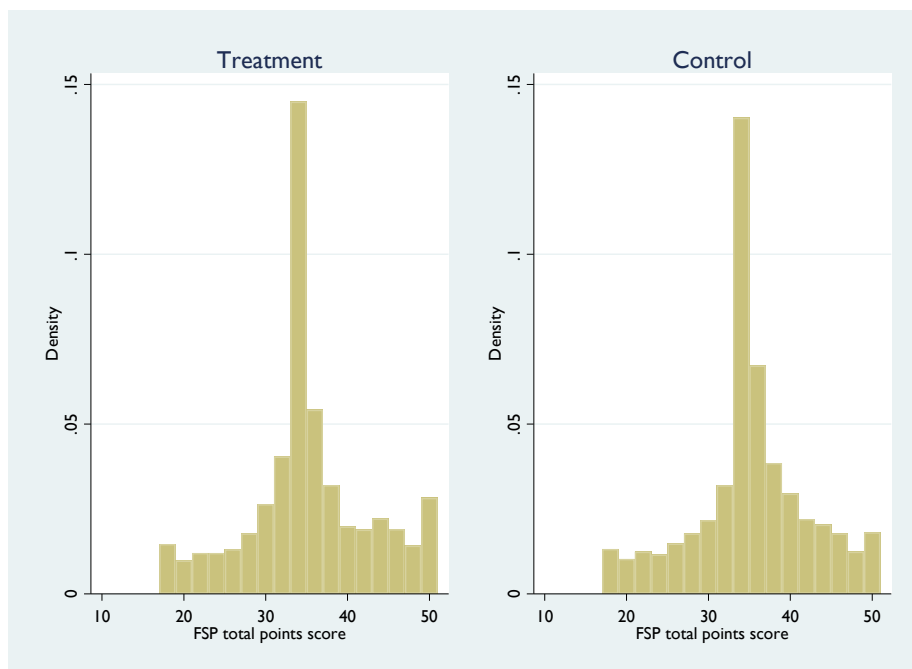


Figure E.2: Total EYFSP scores, treatment and control group, analysis sample



Appendix F: Analysis code

Our primary analysis is run using the following code:

```
regress standardagescore T xnth_expl dblock2-dblock32 if asample==1, cluster(laestab)
```

where

standardagescore is our primary outcome

T indicates whether the pupil is in the treatment or control group

xnth_expl is our measure of lagged school level KS1 attainment

dblock* are the blocking variables used in randomization

The equation we estimate is:

$$Y_{ijt} = \alpha + \beta_1 Treat_j + \beta_2 KS1_{jt-1} + \beta_3 \gamma_j + \varepsilon_{ijt}$$

where i are individuals and j are schools, Y is the standard score on the Progress Test in Maths assessment, $KS1_{jt-1}$ is the measure of lagged school-level KS1 attainment, $Treat$ is our school-level treatment indicator, γ_j being a vector of stratification variables, and ε being an error term. Errors are clustered at school-level (j).

Appendix G: Histograms of PTM raw score and subscores

The charts in this appendix show the distribution of the total raw score as well as each of the four PTM subscores, by treatment and control arm. In each case, the distribution of scores appears similar in both treatment and control groups.

Figure G.1: Total raw score, by treatment arm

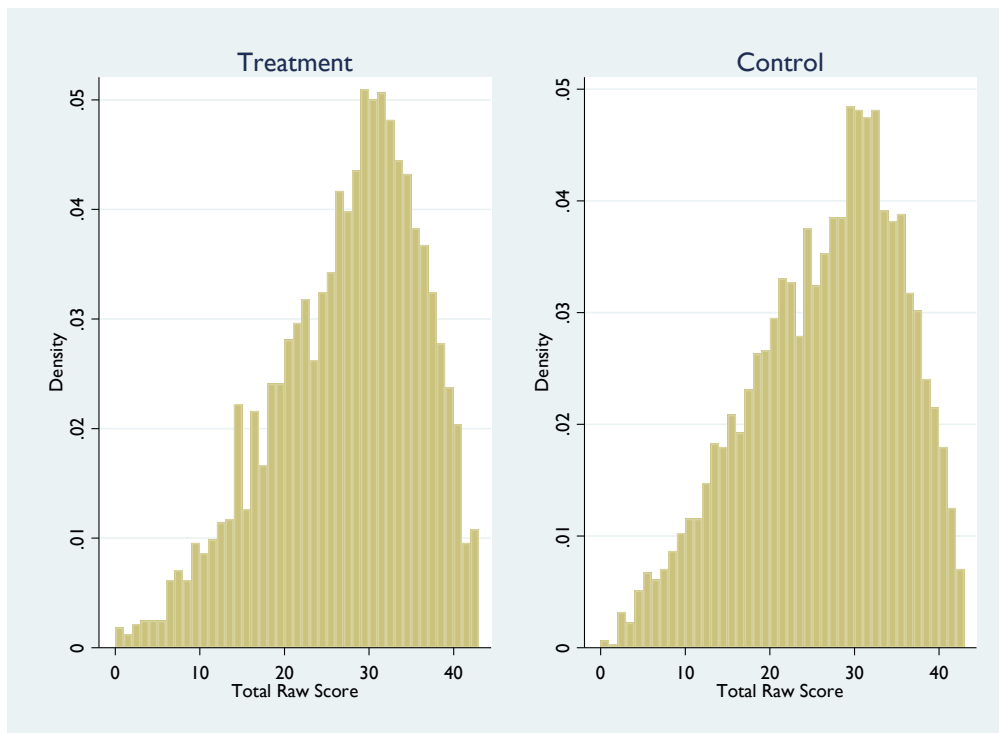


Figure G.2: Mathematical reasoning score, by treatment arm

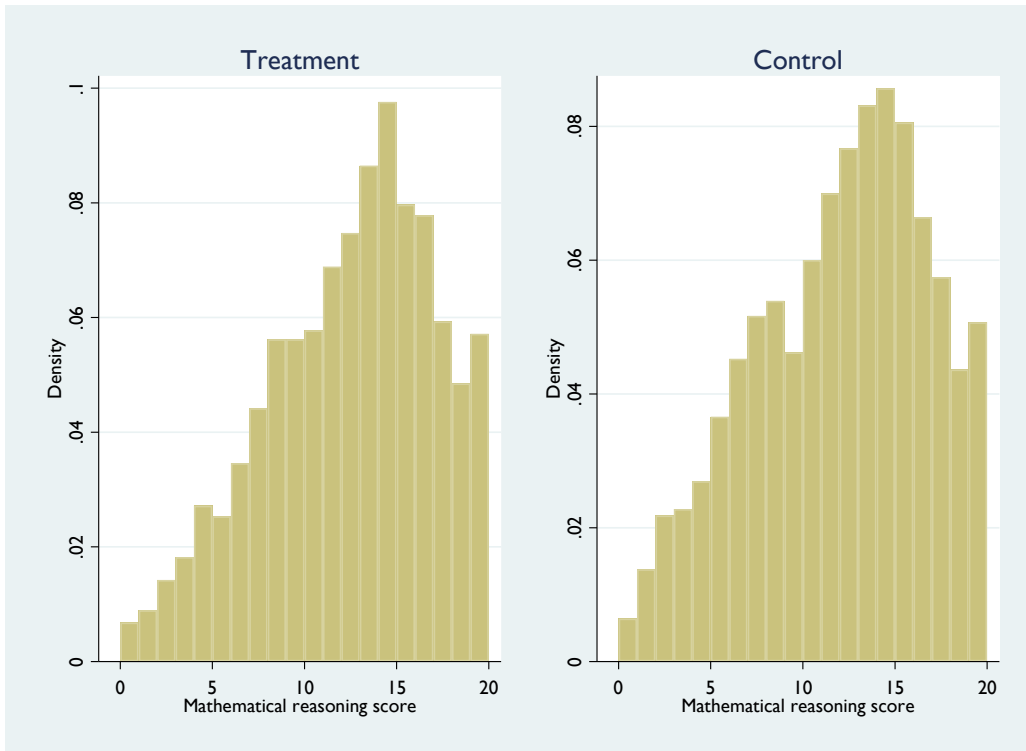


Figure G.3: Fluency in facts and procedures score, by treatment arm

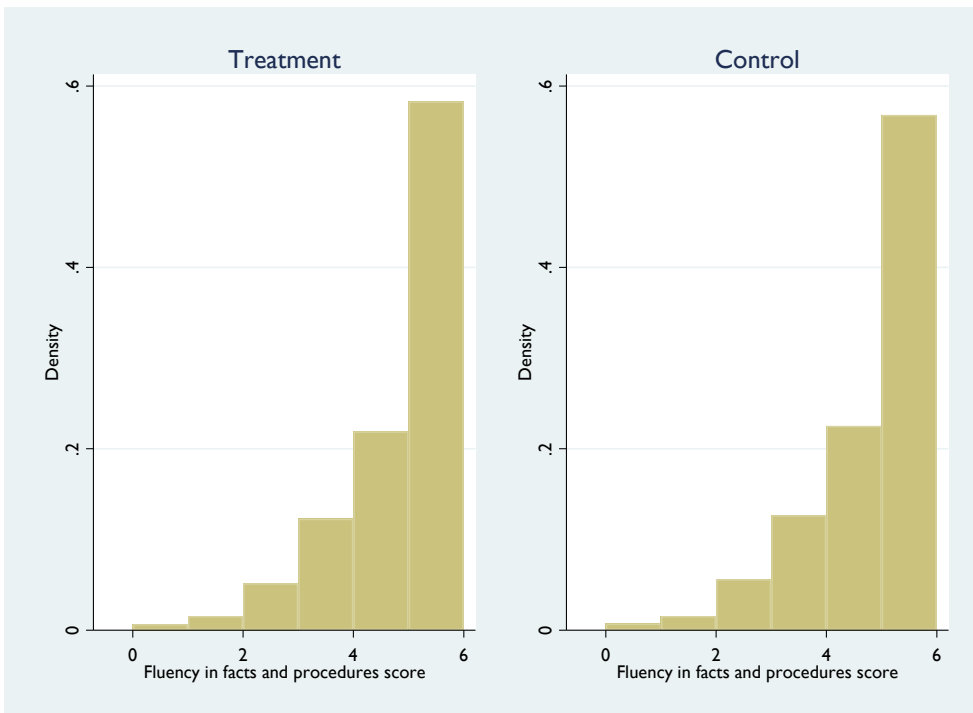


Figure G.4: Fluency in conceptual understanding score, by treatment arm

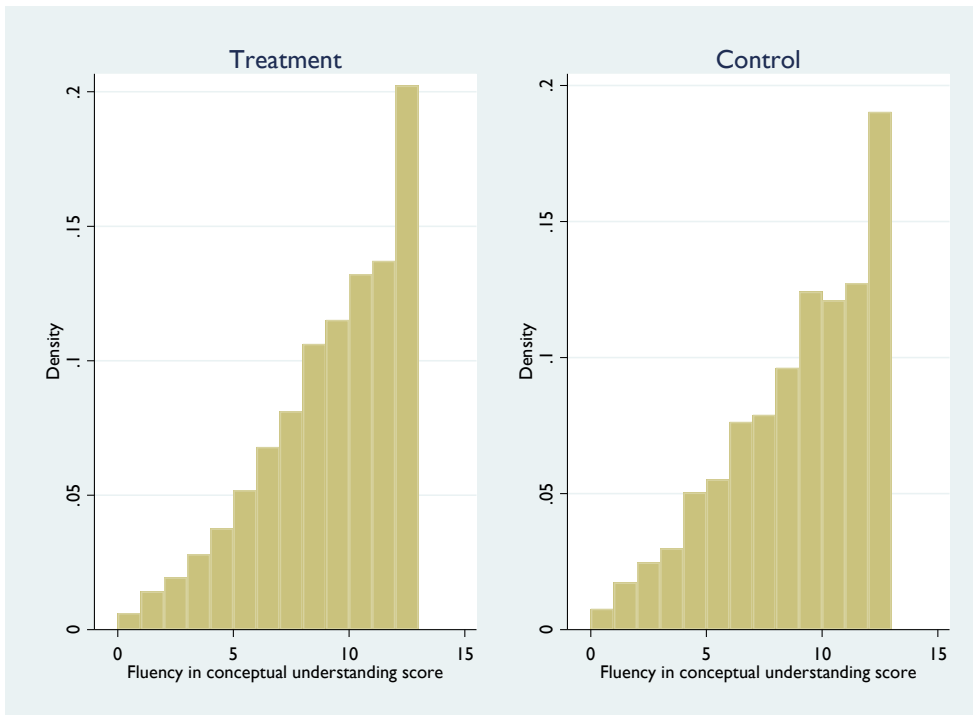
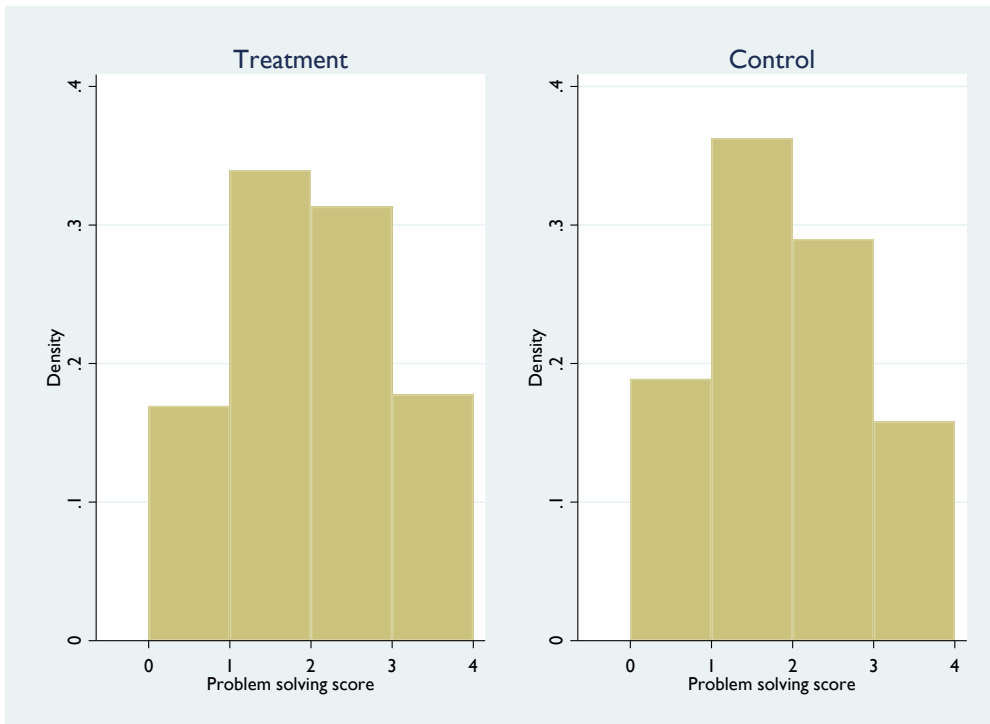


Figure G.5: Problem solving score, by treatment arm



Appendix H: Interacting treatment with prior attainment

Table H.1: Interacting treatment dummy with prior attainment, estimation results

| | PTM score |
|-----------------------------------|-----------------------|
| Low prior attainment | 86.113*** (18.34) |
| Medium prior attainment | 98.932*** (21.11) |
| High prior attainment | 110.921*** (23.47) |
| | |
| Treatment*low prior attainment | 2.156* (2.01) |
| Treatment*medium prior attainment | 1..578 (1.79) |
| Treatment*high prior attainment | -0.082 (-0.07) |
| | |
| N | 6,265 |

Note: t statistics based on school-level clustered standard errors reported in parentheses. Models also control for lagged school-level KS1 attainment and blocking dummies. Statistical significance indicated as follows: * p<0.05; ** p<0.01; *** p<0.001

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

OGL This information is licensed under the Open Government Licence v3.0. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/>

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at www.educationendowmentfoundation.org.uk



Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk