Education Endowment Foundation

1stClass@Number

Evaluation report and executive summary

July 2018

Independent evaluators:

Professor Terezinha Nunes, Mrs Rossana Barros, Dr Maria Evangelou, Professor Steve Strand, Mrs Sandra Mathers, and Mr David Sanders-Ellis

UNIVERSITY OF OXFORD

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.

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

1stClass@Number was independently evaluated by a team from the University of Oxford: Prof Terezinha Nunes, Mrs Rossana Barros, Dr Maria Evangelou, Prof Steve Strand, Mrs Sandra Mathers, and Mr David Sanders-Ellis.

The lead evaluator was Terezinha Nunes.

Contact details:
Terezinha Nunes
University of Oxford Department of Education
15 Norham Gardens
Oxford
OX2 6PY

terezinha.nunes@education.ox.ac.uk

Contents

List of Figures

List of Tables

Executive summary

The project

1stClass@Number was developed by the Every Child Counts (ECC) team at Edge Hill University to support pupils who are struggling with mathematics. The ECC team trained teaching assistants (TAs) to deliver highly scripted lessons to small groups of up to four children. The programme is normally implemented outside of mathematics lessons in other lesson time. It covers five basic mathematics topics: the number system, place value, addition, subtraction, and multiplication. The lessons include the teaching of procedures (for example, counting on, counting up to, counting backwards, and counting in twos and in fives), mathematical signs (+, -, =) and mathematical language and concepts (such as more, less, and equal). Schools are advised to deliver 30 lessons of approximately half an hour, usually three times a week for ten weeks, and are encouraged to deliver additional sessions for pupils who need them. A classroom teacher colleague (the "Link Teacher") is expected to meet the TA once a week to help them review and plan upcoming lessons, and provide feedback.

1stClass@Number was evaluated using a randomised controlled trial. 133 schools in South and West Yorkshire each nominated four children in Year 2 to participate in the project. The schools were then randomly assigned either to receive the intervention or to continue with their normal teaching in Year 2 and receive the opportunity to implement another ECC intervention with a different year group. The primary outcome for the trial was the Quantitative Reasoning Test, which focuses on number knowledge and mathematical problem solving, and the secondary outcome was performance in end-of-KS1 maths tests. A process evaluation collected additional data through observations, questionnaires, and phone interviews. The recruitment of schools began in September 2015 and the research completed in July 2017. The implementation of the intervention took place, approximately, between September 2016 and February 2017.

Key conclusions	
1.	Pupils who received 1stClass@Number made two months' additional progress in maths, on average, compared to pupils in the control group. This result has a high security rating.
2.	The primary result was not statistically significant. This means that, in this trial, even if the intervention had not had an impact, the probability that just by chance we would have observed an effect size as large as the one found is greater than 5%.
3.	Pupils who received 1stClass@Number did not perform better in the end-of-KS1 maths test, on average, than pupils in the control group. This could be because the headline maths measure used in the trial was more sensitive than the simple five point scale available for the end-of-KS1 maths test, or because it tests specifically those skills taught in 1stClass@Number.
4.	Among pupils eligible for free schools meals, those who received the intervention did not make any additional progress in maths compared to pupils in the control group. This result has lower security than the overall result because of the smaller number of pupils.
5.	The intervention was implemented as intended by the developer: most TAs and Link Teachers attended most training sessions, and most of the TAs observed during the evaluation followed the written lesson plans closely.

EEF security rating

These findings have a high security rating. This was an effectiveness trial, which tested whether the intervention worked under everyday conditions in a large number of schools. It was a well-designed randomised controlled trial; relatively few pupils who started the trial were not included the final analysis, and the pupils in the 1stClass@Number schools were similar to those in the comparison schools in terms of prior attainment.

Additional findings

There was no evidence of an impact on KS1 maths. However, there was some evidence that the intervention widened the gap in KS1 maths results between pupils who are eligible for free school meals and their peers. An analysis of impact which included only those pupils who attended all the 30 lessons showed a higher impact of the intervention on the Quantitative Reasoning Test, equivalent to an extra three months of progress, which was not statistically significant by the traditional standards. There was some evidence that pupils with lower prior attainment made more progress on the Quantitative Reasoning Test than pupils with higher prior attainment. This might suggest potential benefits from targeting the intervention at the lower range of prior attainment.

All measures indicated that the intervention was delivered as intended by the developer. Trainers were faithful to the training guidelines, and there was a good level of attendance at the training. Schools were able to provide a designated space for the intervention which was free of interruptions. Observations suggested that the TAs had the resources at hand during the lessons, engaged well with the children, and tended to follow the written script. However, some TAs felt that they did not always have sufficient time to prepare for the sessions.

TAs were confident that the training they received prepared them to implement the intervention, they enjoyed doing so, and they had positive comments about the materials and the lesson plans. Most TAs expressed the view that pupils' confidence in maths increased considerably; this was noted more often than a perceived increase in pupils' maths ability.

The description of 'business as usual' in control schools showed that control and intervention schools had similar policies for pupils struggling with maths. Pupils in intervention schools participated in activities for pupils requiring extra support as well as in the 1stClass@Number intervention. Less than half of pupils in control schools received extra support. It is therefore unlikely that the effects of the intervention were diluted due to the use of alternative interventions in control schools, and it remains possible that any impact was caused simply by additional time being spent on maths rather than any particular aspects of the 1stClass@Number programme.

Cost

1st Class@Number costs approximate £77 per pupil per year when averaged over three years. The cost per child was estimated using the assumption that the intervention is delivered to two groups of four pupils each year over a three-year period, because there is sufficient time in a school year to implement the intervention twice. 1stClass@Number costs £1,612 per school in the first year. Because training does not have to take place again and some of the materials are used again, the cost is reduced in subsequent years.

Table 1: Summary of impact on primary outcome

Outcome	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No. pupils	P value	EEF cost rating
Quantitative Reasoning	0.18 (-0.08 to 0.43)	2	🔒🔒🔒🔒🔓	491	0.09	£ £ £ £ £
Quantitative Reasoning (FSM)	- 0.03 (-0.30 to 0.36)	0	N/A	149	0.92	£ £ £ £ £

Introduction

Intervention

This evaluation aims to test the impact of the programme '1stClass@Number 1' (see below) on student outcomes in mathematics. The intervention, designed by Edge Hill University, is described as a "light touch" intervention for pupils with moderate difficulties in mathematics, which "aims to help them to make faster progress and catch up with their peers**"** (Edge Hill University, 2016). 1stClass@Number also aims to help pupils "to become more confident and interested in learning mathematics, so that they will continue to learn successfully after the end of the intervention"(Edge Hill University, 2016). There are three stand-alone versions of the programme:

- Becoming 1stClass@Number - for pupils with difficulties mainly in Year 1;
- 1stClass@Number 1 for pupils with difficulties mainly in Year 2; and
- 1stClass@Number 2 for pupils with difficulties mainly in Year 3.

This project assessed the impact of 1stClass@Number 1 in schools in South and West Yorkshire.

Who participates in 1stClass@Number?

Pupils are selected for participation in the 1stClass@Number intervention because they are struggling with the curriculum. Under usual conditions, schools follow guidance given on the training course to select pupils to participate who need support at the level of the Year 1 mathematics curriculum. The 1stClass@Number team suggests that schools use a standardised test before and after the intervention and recommends the Sandwell Early Numeracy Test (Revised) because it is designed to measure progress of pupils performing below age expectations. Pupils selected typically are able to count to ten forwards and backwards, can read and write some numbers, and show some knowledge of number facts (for more details, please refer to the description of pupils selected for this project).

Who is involved in the intervention delivery

The intervention is normally delivered by specially trained TAs to small groups of up to four children, but it also can be delivered by a teacher and to one, two, or three pupils. TAs are expected to receive the support of a Link Teacher, who should meet with them once a week to help them review the topic, to plan for delivering the sessions, and to provide feedback. TAs receive six sessions of professional development (discussed below), a handbook with plans for how to organise their time during the weeks, detailed lesson plans and most of the required materials; materials that are not included are common in classrooms (such as blocks and coins).

The 1stClass@Number team advises schools on the selection of TAs to implement the intervention. The advice is that TAs should:

- have successful experience of supporting children's mathematics;
- be able to engage fully in training sessions that include the programme's procedures and the mathematics curriculum; and
- be able to make independent decisions while planning for sessions and teaching children based on an understanding of their needs and with the support of a Link Teacher**.**

The intervention team requires the school to nominate a Link Teacher to support the TAs. Edge Hill University recommends that the Link Teacher should be a senior member of staff who would:

- provide the strategic leadership for the programme;
- be suitably experienced to give support with the teaching of mathematics;

- have time to support the teaching assistant and liaise with the class teacher, senior managers, SENCo (Special Educational Needs Coordinator), and other professionals as necessary; and
- attend two half-day training sessions with the teaching assistant.

How are the TAs and Link Teachers trained?

TAs and teachers are trained by ECC (Every Child Counts) trainers who are accredited annually by Edge Hill University. Evidence that trainers meet the standards for accreditation includes the "Trainer's attendance at and contributions to training and CPD events, field visits, including observation of and discussion with the Trainer, working and professional development files compiled by the Trainer, records and reports submitted by the Trainer to the Edge Hill University, data entered and monitored by the Trainer on the ECC data system, and other evidence as necessary, requested by the National Adviser" (https://everychildcounts.edgehill.ac.uk/accreditation/accreditation-for-trainers/; last accessed on 5th March 2018).

Trainers are provided with a set of slides and guidance notes for the six training sessions, which are run with small groups (between about 9 and 13 at a time) of TAs and teachers in face to face delivery. The first five training sessions are roughly matched with the expected pace of delivery of the topics in the intervention (that is, about four weeks apart); the final training session is an opportunity for TAs to reflect on the work done during programme delivery. Trainers can also visit the school to offer further support to the TA, if asked by the school (this is optional and, under usual implementations, involves an additional fee). In this project, all schools received a visit by the advisor. There is further online support for the maintenance of pupils' records at a website for schools registered with 1stClass@Number. The records help TAs to keep track of the pupils' progress. In the present project, the on-line records were not used, but TAs were asked to keep paper records, which were used for planning and as part of the fidelity assessment.

What are the principles, duration, and contents of the intervention?

The 1stClass@Number Handbook describes the programme as offering rich and varied experiences to children, which are designed to be fun. Counting is viewed as "the basis for the conceptual development of number and arithmetic" (Edge Hill University, 2016) and is used throughout the different topics, but in different ways that are in tune with children's development. For example, TAs are expected to promote children's use of "count on" strategies, and not just count all, and to use counting in different ways to solve subtraction problems (such as counting backwards or counting up from the subtrahend). When multiplication is introduced, children count in twos and in fives.

The intervention is based on the idea that children "construct their mathematical understanding through: (1) Concrete experiences – real, physical objects, such as cubes, counters, fingers, dice and board games; (2) Language – both formal, abstract mathematical language (for example, 'subtract' and 'equals') and natural language that describe an experience ('take away' and 'how many are left?'); (3) Pictures - for example number tracks, number lines and arrays; and (4) Symbols – for example those used for numbers (0, 1, 2, 3….), operations (+, −, ×, ÷) and equality (=)." TAs are encouraged to use these different sources of experience and promote connections between them through the activities designed for the programme.

The programme covers five topics each comprising six half-hour lessons (a total of 30 half-hour sessions), normally three times a week over approximately ten weeks. These sessions are in addition to usual, daily mathematics lessons. The topics in the intervention are: (1) all about number; (2) exploring place value; (3) addition and subtraction 1; (4) addition and subtraction 2; (5) towards multiplication and division. They are basic topics in the domain of number and arithmetic in the National Curriculum.

Background evidence

1stClass@Number is widely used: according to the intervention's website, over 55,000 children in Years 1 to 11 have been supported by different levels of the 1stClass@Number intervention in 4,000 schools. It is one of the programmes created under the umbrella of Edge Hill University's Every Child Counts (ECC) project. Although 1stClass@Number has not been evaluated yet in a rigorous randomised controlled trial (RCT), the programme runs its own evaluation. 1stClass@Number schools are encouraged to report pupils' pre-test and post-test outcomes to ECC online, receiving a detailed analysis in return. According to information provided by the intervention team, between 2012 and 2014, 520 schools reported on 3,702 pupils, 35% of whom were eligible for the Pupil Premium and 42% of whom had a special educational need. After 3.4 months of 1stClass@Number support, they made an average gain of 11.2 months on the Sandwell Early Numeracy Test and their standardised scores rose from 87 to 97. Class teachers reported that 93% of them showed a more positive attitude towards learning mathematics after completing 1stClass@Number. Over the course of the school year, pupils in Year 2 (69% of participants) made an average gain of 4.9 National Curriculum points in teacher assessments for mathematics and 76% of them achieved at least Level 2 at the end of Key Stage 1. This impressive evidence, however, is not based on a rigorous design because the assessment is not entirely independent from the intervention team and the assessors were not blinded to the pupils' participation in the programme.

Further evidence relating to a similar intervention, Numbers Count, which was also developed by Edge Hill University, is available in a rigorous RCT carried out by a team of independent evaluators (Torgerson et al., 2010). Numbers Count was developed on the basis of the same theoretical background as 1stClass@Number, but it is aimed at a different group of pupils. It targets the lowest achieving 5% of pupils, and thus aims to support pupils with more severe mathematical difficulties than 1stClass@Number. Numbers Count is delivered by specifically trained teachers (rather than TAs) and is more intensive: it is delivered in one-to-one daily half-hour sessions for 12 weeks (42 hours in total).

The study by Torgerson et al. involved 44 schools and 409 children. Two measures were used to evaluate the impact of Numbers Count: the primary measure was Progress in Maths 6 (PiM), which is a measure designed independently of the intervention team and was implemented by testers blinded to the children's group membership in the trial; the secondary measure was the Sandwell Early Numeracy Test, which, as already mentioned, is a measure used by the Numbers Count teachers to evaluate the children's progress and was administered by testers who were not blind to the children's group membership. The children who participated in the Numbers Count intervention attained a significantly higher mean in the PiM than the children who did not receive the intervention; the effect size was 0.33, which is equivalent to seven additional weeks improvement in the National Curriculum, in comparison to the control group. The Sandwell Test also indicated a significant difference between the intervention and control groups, with a considerably larger effect size (greater than 1), but Torgerson et al. comment on the threats to the reliability of this result, as the Sandwell Test was not designed independently of the intervention team and was administered by testers who knew the pupils' group membership in the trial. Torgerson et al. further report a small and only marginally significant effect on a measure of attitudes towards maths, but caution is needed about the reliability of this finding, which was not replicated in other measures in the same survey.

In summary, 1stClass@Number runs its own evaluation, which has provided evidence for its efficacy. Although this evidence does not meet the standards of a RCT, it does suggest that pupils make progress in a variety of measures. Numbers Count, a more intensive intervention based on the same principles and delivered individually by specially trained teachers to pupils with measurably larger difficulties, was found effective in improving their performance in a rigorous RCT. This seems, therefore, to justify the evaluation of a similar evaluation of 1stClass@Number, which is a less intensive intervention delivered by specially trained TAs to small groups of up to four children.

Evaluation objectives

This project aimed to carry out a rigorous evaluation of the impact of 1stClass@Number on pupils' numeracy outcomes using an RCT design and an independently designed assessment, the Quantitative Reasoning Test. This test assesses topics that are taught in the 1stClass@Number intervention, such as additive composition of number (that is, understanding that any number can be seen as the sum of two other numbers), addition, subtraction and multiplication in simple contexts. The test has good reliability and is a predictor of KS1 Maths outcomes even after controlling for cognitive skills and working memory. The impact information was supplemented by a carefully designed process evaluation in order to describe the conditions necessary for successful implementation and the challenges faced by schools and TAs using the intervention. The project also analysed whether the impact of the intervention was the same for the subgroups defined by eligibility for free school meals (FSM) or by gender. Finally, the project estimates the cost of the intervention when it is delivered to two small groups of four children in the school each year over a period of three years.

Research questions

Primary research questions:

- Do the children identified by their teachers as struggling with mathematics at the start of Year 2 who participate in the 1stClass@Number intervention show greater gains in the Test of Quantitative Reasoning than children identified by their teachers as struggling with mathematics at the start of Year 2 who do not participate in the intervention?
- Do these children show better outcomes in the national KS1 mathematics tests if they participate in the 1stClass@Number intervention than the comparable group that does not participate in the intervention?

Secondary research questions:

- Do children entitled to FSM benefit to the same extent as other children from the 1stClass@Number intervention as assessed by the Test of Quantitative Reasoning?
- Do children entitled to FSM benefit to the same extent as other children from the 1stClass@Number intervention as assessed by their performance in KS1 Maths?
- Is the 1stClass@Number intervention equally effective for girls and boys as assessed by the Quantitative Reasoning Test?
- Is the 1stClass@Number intervention equally effective for girls and boys as assessed by their performance in KS1 Maths?

Ethical review

The project was approved by the Central University of Oxford Research Ethics Committee on 24th February 2016. Head teachers signed a Memorandum of Understanding (see Appendix 1) agreeing to the conditions set out by the intervention and the evaluation team. Schools agreed that they would: (a) seek parental permission for participation using an opt-out form; (b) nominate four children for participation in the project; and (c) provide all the conditions required for the implementation and assessment of the intervention.

The nominations of the children for participation and surveys answered by the school staff were collected electronically using a secure website; this procedure was approved by the Ethics Committee. When entering the children's nomination, schools confirmed by answering yes, that the school had permission for the children's participation in the trial. Parents were informed that the data would be shared with the EEF, the EEF's data contractor FFT Education, the Department for Education, and the UK data archive for research purposes. They had the option of allowing their children to participate in the project, but not giving consent to access to their children's UPN (unique pupil number); 26 parents refused to release the children's UPNs.

Pre- and post-tests were collected by testers who were trained by the evaluation team to follow a protocol for the testing of young children and for safe keeping of data sets. Data management followed the Data Protection Act. Children were identified in the database only by an identification number created for the project, which was also used to obtain data from the National Pupil Database (NPD). The data were stored in a secure site in the Department of Education, University of Oxford, which is password protected and only accessible to the evaluation team. Once all the information was included in the data set, the data were anonymised and no one is able to identify individual children.

Project team

The intervention team

The grant holder, Nick Dowrick, and his team were extremely cooperative and worked with the evaluation team throughout the project. The Every Child Counts (ECC) mathematics team at Edge Hill University created 1stClass@Number and supervises all uses of ECC programmes. The team members were:

- Linda Lavagna-Slater, ECC National Adviser
- Rebecca Lloyd-Lewis, ECC National Adviser
- Louise Matthews, Head of ECC Mathematics
- Susie Nicholson, ECC National Adviser
- Kathy Secular, ECC National Adviser
- Andy Tynemouth, ECC National Adviser

The Every Child Counts administrative team at Edge Hill University was responsible for recruitment and implementation of the project.

- Sarah Brand, ECC Co-ordinator
- Angela Dixon, ECC Co-ordinator
- Cathy Hayton, ECC Co-ordinator
- Yvonne Panteli, ECC Project Manager

The evaluation team is based at the University of Oxford and consisted of six members:

- Prof Terezinha Nunes, PI
- Mrs Rossana Barros, PI
- Dr Maria Evangelou, PI
- Prof Steve Strand, PI
- Mrs Sandra Mathers, PI
- Mr David Sanders-Ellis, Researcher

Trial registration

The trial was registered with ISRCTN, Protocol serial number - 37482473.1, ISRCTN15932880 DOI 10.1186.

Methods

Trial design

This was a two-arm randomised controlled trial, with an intervention and an unseen[1], business-as-usual control group. The unit of randomisation was the school in order to avoid contamination between groups. For the purposes of the trial, the number of intervention groups in each school was restricted to one, which means that each school could only nominate four pupils, because it was anticipated that some schools might not identify eight children with the appropriate profile in order to include two groups per school in the trial. It was also considered that, if schools attempted to implement the intervention twice for the purposes of the trial, this might increase the demands on the school resources, which could ultimately influence the evaluation of the intervention programme. However, schools were free to implement the intervention with further groups, if their resources permitted, and some schools in fact did so.

The control schools were asked to take a business as usual approach to supporting pupils who struggle with mathematics. As an incentive to participate, control schools were offered the option of implementing another ECC intervention designed by Edge Hill University, either on numeracy or literacy, with a different year group. Schools agreed that, if they did not comply with the trial requirements, they would incur the costs of the intervention. Schools agreed that, if they were allocated to the control group, they would implement with the nominated children the interventions that they would normally use with children struggling with the mathematics curriculum at the beginning of Year 2.

The design is intention to treat, which provides a rigorous assessment of the intervention effects and of the barriers to compliance. The evaluation team is independent from the intervention team and the primary outcome measure was developed independently from the intervention team

The intervention was implemented outside the time dedicated to regular mathematics lessons in the school, and thus includes extra time dedicated to mathematics learning. Sessions were scheduled at different times on different days to avoid pupils often missing the same lessons.

It was not necessary to make any changes to the protocol during implementation. Only one intervention school was unable to implement the intervention to the end because the nominated TA became chronically ill, and it was not possible to identify another TA in the school to continue implementing the intervention. However, the intervention school did not withdraw from the trial and the pupils were post-tested. Control schools were offered the training and the use of another intervention developed by the ECC (Every Child Counts) team appropriate for a different year group.

Participant selection

Schools

The project was part of the EEF strategy for evaluating interventions delivered by TAs in South and West Yorkshire; the EEF emailed schools and invited schools to participate through its website. The ECC team also displayed information about the project on their website with an invitation to contact them. The ECC team has strong relations with the schools in this region and distributed invitations to about 200 schools that attended an earlier project supported by the EEF (Making the Best Use of TAs)

[1] Unseen control group is used in psychology. When the control group is active, the researcher implements activities with the control group considered irrelevant to the aim of the intervention and thus should not affect the outcome measure. Unseen control groups have no activities directed by the researchers. Psychologists see active control groups as a means of controlling for the Hawthorn effect (that is, an effect that appears simply because some change was made to the context).

in which the team was involved. Further invitations were sent to 330 schools in the region. The ECC team was contacted by 233 schools, 225 of which were in South and West Yorkshire. Schools that had already used the 1stClass@Number intervention, junior schools and pupil referral units were not considered eligible; this led to the exclusion of 25 schools. Agreement forms were sent to 208 schools; 141 signed and returned the agreement. The first 138 schools that returned the agreement were invited to a briefing meeting and three were placed on a waiting list and were eventually included in the trial because schools that did not send a representative to the briefing meeting were excluded (n=6); one school dropped out after the briefing meeting but before the pre-test; one school dropped out after the pre-tests but before randomisation; 133 schools were randomised, of which 132 were in South and West Yorkshire (see Figure 3 for a detailed description).

Pupils

At the briefing meeting, the intervention team explained to the schools the eligibility criteria for pupil nomination to participate in the trial, which were also provided in writing, as follows:

"1stClass@Number 1 is for pupils who have fallen behind their peers and need additional support around the level of the Year 1 curriculum. They may typically:

- be able to count forwards and backwards in ones from 10
- have some knowledge of number facts
- be able to perform simple addition and subtraction calculations using counting all and counting out approaches
- be able to read and write numbers but lack secure understanding of their magnitude and quantity
- struggle with mental calculation strategies and mathematical vocabulary
- lack confidence in mathematics and be reluctant to talk about their mathematical learning." (source: Edge Hill University document sent to Head Teachers in preparation for the EEF trial of 2016/17).

Teachers were also advised to nominate for the same group children who can 'gel' and learn together and were not participating in another intervention, in order to avoid overload.

Every Child Counts recommends the use of the Sandwell Early Numeracy Test (Revised) as a tool to assess the children before and after they have received the intervention. In this trial, the intervention team gave guidance to schools on how they might take into account the pupils' scores in the pre-test as part of the information to be considered when they nominated pupils for participation in the project.

After the Head teachers had signed an agreement to participate in the trial, schools sought permission from all parents of children at the end of Year 1 for participation using an opt-out consent form (see Appendix 2 for consent forms).

All schools were asked to nominate four pupils using the eligibility criteria above. At the end of Year 1, the teachers administered the Quantitative Reasoning Test, the baseline measure for the project, in the Summer term before the children moved on to Year 2. The pre-test results were made available in September to teachers when the children started in Year 2. Pre-test results were presented as percentile ranks with children in the class being ranked against a regional sample of over 5,000 children. Because percentile ranks are familiar to teachers, it was decided to present the information in this format. The use of a large sample in the region as a reference group could help teachers make informed decisions about the true extent of pupils' need for additional support (i.e. not just making decisions based on class averages).

Nomination of children and of the staff in charge of the intervention (TA and LT) was sent by schools to the evaluation team at the start of the school year, when the children were in Year 2, prior to

randomisation. Only children who had participated in the pre-test were eligible for nomination. Nomination took place at the start of Year 2 and prior to randomisation.

TAs and Link Teachers

Prior to randomisation, schools nominated one TA and one Link Teacher (LT) who would be trained, if the school were allocated to the intervention group. Edge Hill University's recommendation for the identification of a suitable TA and LT were described previously (under the heading "who is involved in delivery"). It is recommended that the Link Teacher should be a senior member of staff. If the school prefers to nominate a class teacher who will liaise closely with the teaching assistant, then a senior manager can provide the leadership and additional support that will maximise the impact of the intervention. Schools were made aware of these criteria in the briefing meeting; TAs and Link Teachers were nominated by all schools to the evaluation team prior to randomisation at the same time as the pupils were nominated.

Outcome measures

Primary outcome

The primary outcome measure used at pre-test and post-test for the trial was the Quantitative Reasoning Test (Nunes et al., 2007: http://onlinelibrary.wiley.com/doi/10.1348/026151006X153127/abstract; 2015: https://doi.org/10.1080/10986065.2015.1016815). The measure was chosen for its relevance to the intervention programme and its psychometric properties.

The measure assesses arithmetic in simple contexts, knowledge of additive composition (which is a precursor of understanding place value, a topic taught in the intervention), and solving problems that involve additive and multiplicative reasoning. It covers the basic aspects of number and problem solving that children start to understand at the start of primary school, but not spatial reasoning and geometry. The topics are similar to those covered in 1stClass@Number, which does not cover geometry. Multiplication was included in the 1stClass@Number programme only recently, after this topic became part of the curriculum for Year 1 pupils, and receives considerably less attention than additive reasoning (one out of five sets of six lessons).

The Quantitative Reasoning Test has been found to be a strong predictor of KS1 attainment (Nunes et al., 2007): when the two assessments were separated by 14 months, their correlation was equal to .7. It remained a significant predictor of KS1 even after controlling for the children's age, number skills (measured by the British Abilities Scale), and their Working Memory (measured by Counting recall, a sub-test of the Working Memory Battery for Children; Nunes et al., 2015).

Each child is given an answer book where the items are presented, two on each page. Each item is presented with the support of a figure that provides information to reduce the memory demands. For example, the story problem "In a toy shop there are 9 dogs and 5 cats. How many more dogs than cats are in the shop?" is illustrated by a figure that contains drawings of 9 dogs and 5 cats; the children do not need to recall the numbers because the drawing preserves this information. No item requires reading text as all the information is presented orally by the tester. During testing, the figure that illustrates each item is presented on a screen in front of the class as well as in the child's booklet; this procedure makes it easier for the children to check whether they are looking at the correct item when the tester reads the information.

Pre-tests were carried out prior to randomisation by teachers who received detailed instructions for implementing the assessment from evaluators at Oxford University. The post-test was delivered to whole classes according to a standardised procedure by testers blinded to the condition to which the school was assigned. The pre-test was administered to the whole class in order to support the nomination of pupils for participation; at post-test, this procedure aimed to avoid stigmatising the targeted pupils. After the whole class testing, testers checked whether any of the nominated children was absent. If any of the nominated children were absent, testers scheduled a second visit to the school

and, whenever possible, the child was tested in a small group with other children who had been absent. The testing procedure in follow-up cases was as similar as possible to the whole class testing: the items were presented on a computer screen and the tester pointed to the screen rather than to the children's answer book when reading the item.

The form used for this trial as the outcome measure includes a total of 20 items, 12 of which are identical to items used in the pre-test and eight which are different and more difficult than those items which they replaced. Teachers did not keep copies of the pre-test and were not aware that some items would be the same at post-test. This combination of identical items and new items allows for measuring progress from pre- to post-test while at the same time increasing the level of difficulty of the assessment at post-test.

Marking of the tests was carried out by assessors trained by the evaluation team; machine marking is not possible because young children's handwriting cannot be machine read. Assessors trained by the evaluation team scored all the tests. Pre-tests were scored before randomisation as the information about the pupils' performance had to be provided to the teachers. Post-tests were scored by assessors blinded to the condition and to the children's nomination for participation in the project; all tests from all classes were scored. Assessors noted any doubts in scoring; blind double-marking was used for all participants in order to identify discrepancies in scoring. Discrepancies and doubts were discussed and resolved by the evaluation team. Once marking was complete, the data files were machine generated and were subsequently double-checked for accuracy.

Raw scores and rank percentiles can be obtained for this test. In this project, only raw scores were used in the analyses. There are two forms of the test for young children, which were used in this project as pre- and post-test, and two further forms, one for 7 to 9 year-olds and one for 10 to 12 year-olds. All these forms have been administered to large samples of pupils in the UK and validated by inspection of their correlations with KS1, KS2, and KS3 tests (Nunes, Bryant, Barros and Sylva, 2011);

http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8279.2011.02033.x/abstract).

Secondary Outcome

Key Stage 1 (KS1) maths was used as a secondary outcome measure. This is administered by the pupils' own teachers; thus the testers could not be blinded to pupils' condition and would know whether the pupils had participated in the 1stClass@Number intervention. KS tests are familiar to schools and can have a significant impact on pupils' lives, as teachers use the results for tracking pupils and often for setting by ability levels. KS1 maths results were obtained through the NPD using the variable KS1_MATH_OUTCOME.

The test has two parts, reasoning and arithmetic, but schools are not required to report the number of marks in the two tests nor the total marks. The NPD records contain only an ordinal scale of measurement with five points:

1. BLW (Below the Foundation stage);
2. PKF (Pre-Key stage; foundations for the expected standard);
3. WTS (Working towards the expected standard);
4. EXS (Working at the expected standard); and
5. GDS (Working at a greater depth within the expected standard).

For the purposes of this analysis, these levels were converted into numbers, as indicated above. The measure was chosen due to its high ecological validity, as Key Stage tests are the only official assessments in primary school used in the country. However, it is noted that the measure has a

restricted range of scores, and this was taken into consideration in the analyses. Nevertheless, because of the restricted range of scores, it is not a sensitive measure.

Sample size

PowerUp software (Dong & Maynard, 2013) was used for the power calculations to determine the sample size. The following assumptions were made: (i) pupil outcomes measured at pre-test and at post-test have a correlation of r=0.7[2] (Nunes, Bryant, Evans & Barros, 2015); (ii) a within school sample of four pupils per school; (iii) an intra-class correlation coefficient of 0.15; (iv), power of 0.80, alpha of 0.05 and a two tailed significance test. A sample of 60 schools per trial arm (120 schools in total) would yield 480 children in total giving 80% power to observe an effect size of 0.21 SD[3]. It was considered advisable to recruit 130 schools to allow for loss of schools after randomisation.

Randomisation

Prior to randomisation, children were assessed at the end of Year 1 and teachers in Year 2 completed the nomination process. The nomination of the TAs who would be implementing the programme was also completed prior to randomisation; one school asked to wait until later in the year because the nominated TA became chronically ill but no substitute was nominated later; this school was subsequently assigned to the control group and the alternative intervention was not delivered. Randomisation was implemented by the evaluation team in September 2016, when the children were in Year 2; the intervention team informed the schools of the outcome.

Schools were clustered in five geographic areas by the intervention team for delivery of the training to TAs; TAs in one location (Leeds) were split in two groups in order to keep the size of the groups in line with the training process offered by the intervention team. The protocol had identified four geographical areas for randomisation, but at the time not all schools had been recruited yet. The original areas identified during recruitment were North (Leeds/Bradford plus York); West (Kirklees/ Calderdale/ Wakefield); East (Doncaster) and South (Barnsley/ Sheffield/ Rotherham). After recruitment was completed, it was necessary to redefine the clusters of schools (see Table 2) and it turned out that there were too many schools in Leeds for the intervention team to run a single training session. The intervention team decided to run two training sessions for schools in Leeds to keep the size of the training groups within the limit of 18 set by Edge Hill University. In this project, the number of schools per training group varied between 9 and 13. The intervention team allowed TAs to attend a training session in a different group if they had not been able to attend the session in their own group. This extra opportunity aimed to increase fidelity of training and of implementation.

The evaluation team decided to use six blocks for randomisation, based on geographical area and training sessions, to allow for an analysis of differences between training sessions. The number of schools in each block is presented in Table 2.

[2] The paper by Nunes et al., 2015, indicates the test-retest correlation for the 31 items. We ran the analysis for just the 20 items used here for the pre- and post-test. The correlation in the reference given will have a higher value as the 31 items give a correlation of .78. We have a pool of items and analyses the correlation for the specific items used in this project.

[3] This is different to the published Statistical Analysis Plan which has an effect size of 0.22 with the same assumptions. This is because the software PowerUP (Dong & Maynard, 2013) was used in this calculation, whereas the SAP used the software Optimal Design. In actuality, the difference is 0.01 when looking at three decimal places (i.e. 0.208 rounds to 0.21).

Table 2: Number of schools by blocks and by percentage of children eligible for Pupil Premium

Geographical block	Number of schools	Blocks defined by the median number of children eligible for Pupil Premium per school			
		Blocks with lower % of Pupil Premium	Number of schools	Blocks with higher % of Pupil Premium	Number of schools
Doncaster	24	Block 1 (11.28%)	12	Block 2 (38.23%)	12
Huddersfield	17	Block 3 (4.95%)	8	Block 4 (36.63%)	9
Leeds A	24	Block 5 (11.09%)	12	Block 6 (42.00%)	12
Leeds B	24	Block 7 (8.68%)	12	Block 8 (41.44%)	12
Sheffield	18	Block 9 (14.24%)	9	Block 10 (53.13%)	9
Wakefield	26	Block 11 (11.60%)	13	Block 12 (32.61%)	13
Total	133		66		67

Schools in each block were split into two blocks using the median for the number of children eligible for Pupil Premium from the database 2016-2017[4], resulting in 12 blocks. Random numbers were generated for all schools using SPSS. Schools were ordered by these random numbers in ascending order within their block. The number of schools per block is displayed in Table 2. The schools that were allocated the highest random numbers in each block were allocated to the intervention group until half of the schools had an allocation; those with the lowest random numbers were allocated to the control group.

The SPSS syntax used for this randomisation process was:

Syntax:

```
COMPUTE random=RV.UNIFORM(1,2).

EXECUTE.


SORT CASES BY block(A) random(A).
```

As the number of schools was even in five blocks and odd in one block, 67 schools were allocated to the intervention group and 66 to the control group.

[4] Source: (**https://www.gov.uk/government/publications/pupil-premium-conditions-of-grant-2016-to-2017**)

Analysis

Preliminary analyses of the pre- and post-test

"Measurement is the corner-stone of science. Without measurement, there is no science" (Graham, 2009, p. 564). In order to analyse whether the measure used in this study was sensitive enough to measure progress in the sample of children who participated in the project, preliminary analyses were carried out. These analyses are also necessary to verify whether the assumptions about the measures made in the use of the statistical methods were violated.

A total of 5,358 pupils participated in the pre-test, which was used as the regional reference group. As indicated earlier on, teachers in all the recruited schools were informed of each pupil percentile rank, and this information could be used by them in the selection of pupils "struggling with maths" to be nominated for participation in the trial. The mean number of correct responses (out of 20 items) for the reference group was 9.39 (SD=4.61).

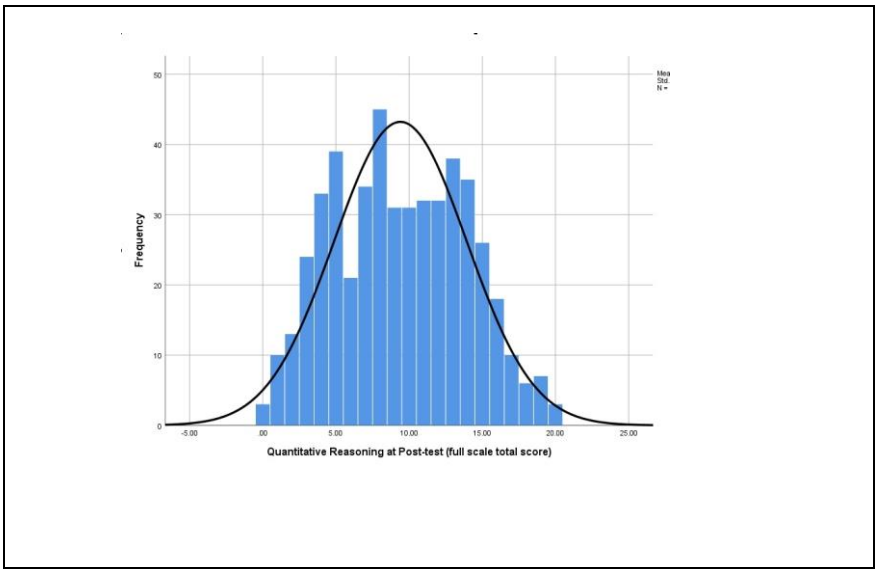
From this larger sample, 532 pupils were nominated by their teachers for the trial, considered by the teachers to be struggling with maths. Teachers could consider in their nomination the pupils' percentile ranks in the test, their own judgements and the pupils' ability to work together as a group. As the nominated pupils were judged by the teachers to be struggling with mathematics, the mean score for this group was expected to be considerably lower than that for the reference group. This was in fact the case: the mean was 4.86 (SD=2.52). This mean was below the 20th percentile rank for the reference group. The median and mode for this group were both equal to 4, a score that corresponded to the 15th percentile rank for the reference group. The distribution of scores met the assumptions of normality.

An analysis of the variation between nominated pupils at pre-test showed that their scores varied between 0 and 17; 31 pupils (5.6%) scored above the mean for the population that took the pre-test; this suggests that teachers took into account information other than the pupils' percentile ranks and that not all pupils could be considered as struggling with maths, as indicated by their scores in the Quantitative Reasoning Test.

The post-test contained 12 items that were identical to those in the pre-test and eight selected from a pool of quantitative reasoning items to increase the difficulty of the test. In the post-test, the mean number of correct responses for the reference group was 13.18 (SD=4.62). For the nominated children, the mean number of correct responses in the post-test was also higher than that for the pre-test: the mean was 9.40 (SD=4.53), which places this group just above the 21st percentile rank for the reference group, and thus still low as a group in comparison to the reference group. The median was equal to nine and the mode to eight. The distribution of scores in the post-test for the pupils participating in the trial is shown in Figure 1; the distribution meets the assumptions of normality.

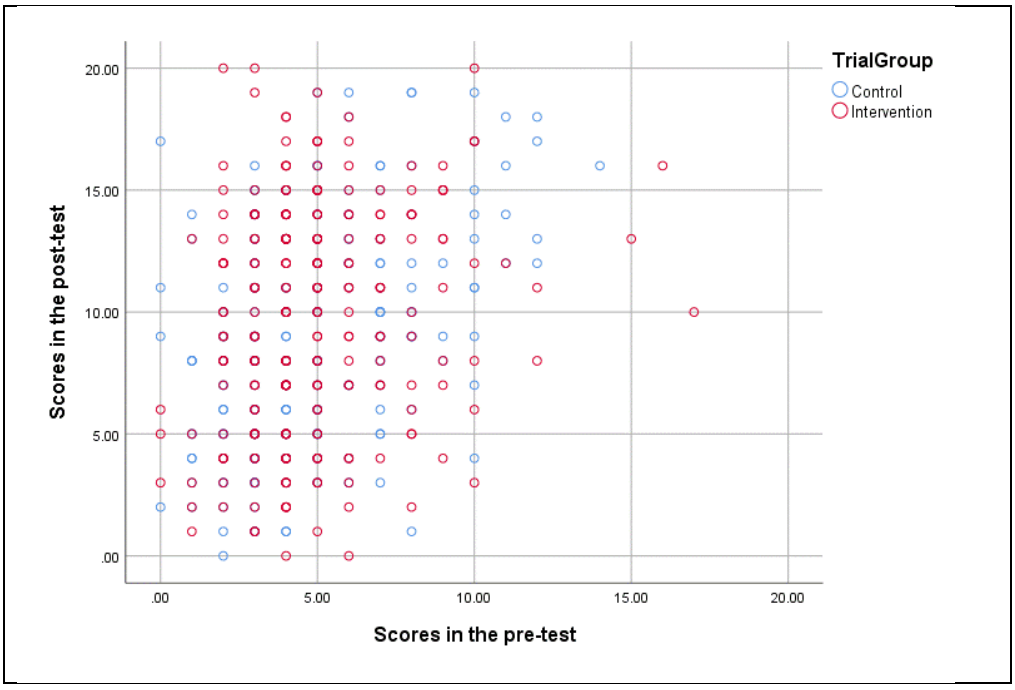
In order to evaluate the sensitivity of the test to growth in this sample of children struggling with maths, we carried out a comparison between the nominated pupils' scores in the pre- and post-test using only the 12 items that were identical in the two occasions of testing. The mean number of correct responses in these 12 items at pre-test was 2.73 (SD=1.77) and in the post-test was 5.79 (SD=2.83). This difference was statistically significant according to a t-test for paired samples; Cohen's d effect size was 1.08. Thus the measure is sensitive to Year 2 children's progress even if they are struggling with maths at the start of the year.

Figure 1: Distribution of scores in the Quantitative Reasoning Test for the pupils nominated for the 1stClass@Number project at post-test



The correlation between the pre- and the post-test for the whole sample was 0.63 (N=4,653), which is significant but lower than the correlation observed in previous studies, which was 0.7. The correlation between the pre- and post-test for the sample of nominated children was considerably more modest: 0.29. This lower value can be attributed to the fact that this is a selected sample with a restricted range of scores, as the children were nominated because they were considered by the teachers to be struggling with maths. Figure 2 shows the scatterplot for the post-test scores as a function of the pre-test scores. Grids have been included to illustrate how the vast majority of scores at pre-test is less than half of the points in the scale because of the selection of the sample for the intervention. At post-test, the spread of scores has clearly changed. The scatterplot also indicates that the best function to describe the relation between pre- and post-test scores would be linear.

Figure 2: Post-test scores on the Quantitative Reasoning Test plotted as a function of pre-test scores on the parallel form of the Quantitative Reasoning Test



Primary analysis

The approach to analysis for this trial was written in the statistical analysis plan (SAP) prior to data on post-testing being received. The SAP was approved by the EEF and the intervention team and is published on the EEF website (**https://educationendowmentfoundation.org.uk/public/files/Projects/Round_8_1st_class_at_number_SAP.pdf**).

Multilevel models were used to assess the effect of the intervention both with the primary outcome, which was the Quantitative Reasoning Test, and with the secondary outcome, which was Key Stage 1 Maths. Multilevel models were used in the analyses for both the primary and secondary outcomes in order to take into account clustering of pupils within schools, because the pupils' scores are likely to be correlated within schools. The models took into account the school membership, the pre-test scores in the Quantitative Reasoning Test, and the membership in the trial group (i.e. intervention or control). The multilevel models were carried out using R – EEF analytics for models with school membership and pre-test effects; models with more factors were implemented with MLwiN.

All primary analyses were based on an intention to treat design, which includes all the data from both the intervention and the control group available to the team. Although three control schools did not agree to the post-test, consent had been obtained for collecting KS1 Maths data from these schools and was not withdrawn, so it was possible to include the pupils from these schools in the analysis of the secondary outcome measure.

The analyses took into account the pre-test scores (fixed effect) and group allocation (fixed effect) as well as nesting in schools (random effect). In the equation, the variables are identified by names to facilitate its reading. The same equation was used for both outcome measures, Quantitative Reasoning Test, and KS1 Maths, so the equation uses the general term 'outcome measure'.

$$\textit{Outcome measure}_{ij} = \beta_0 + \beta_1 \text{Pre-test}_{ij} + u_0 \text{School}_{ij} + \beta_2 \text{Group allocation}_{ij} + e_{ij}$$

In the equation:

β_0 is the intercept

β_1 slope for pre-test effect (as a fixed effect, this coefficient is not assumed to vary between schools)

u_{0j} slope for school effect (random effect; also called level 2 residual by some)

β_2 slope of group allocation (Control vs Intervention at school level, fixed effect)

e_{ij} = level 1 residual

Effect sizes were calculated using ANCOVA controlling for pre-test scores, to increase precision and power. Hedge's g was used to calculate the effect size and employed the means corrected for pre-test differences and the total standard deviation for the whole sample (that is, including both cells in the design); the confidence interval used the traditional 95% interval. The intra-cluster correlation is reported for pre- and post-test.

Multilevel analysis is used in this project, but we have some concern about its use. The number of pupils in each cluster is small: there are only four pupils per school. This restriction was necessary because the intervention designers considered that the intervention would not be delivered as intended, and thus it would lack fidelity, if more pupils were included in the small groups for implementation. Alternatives were discussed at the stage of design of the study (such as requesting schools to run the intervention with a minimum of two groups in order to have eight pupils per school) but were discarded in view of the pressures that running two intervention groups in each school could place on schools and the possible variation in delivery to two different groups (for example, due to running the intervention a

second time). Mass and Hox (2005) suggest that the number of clusters is more important for multilevel models in providing reliable data than the number of individuals per cluster, but they do not consider any sample size smaller than five at the individual level. With the attrition of pupils in different schools, the number of pupils per school varied in the final analysis: two schools had two pupils, 25 schools had three pupils, and the remaining 103 schools had four pupils, as expected in the trial design; the total number of schools in the analyses was 130, a number that meets the recruitment target set out in the power calculation at the design stage of the trial. According to Rutterford et al. (2015), when one knows the number of pupils per cluster, and this differs, it is possible to use the mean number of pupils per cluster to calculate the minimum detectable effect size. The mean number of pupils per school was 3.78, which is rounded to 4 when entered in a power calculation; this means that the attrition cannot be said to affect the power and design of this trial. So it was decided that the analyses would be implemented in agreement with the statistical plan, as the loss of participants did not change dramatically the number of observations per cluster. As a measure of caution, though, a single level ANCOVA was also implemented in order to verify whether the same results were observed.

Non-compliance with intervention

It is important to obtain information on non-compliance in order to avoid Type III errors (rejecting a hypothesis for the wrong reason; Basch et al., 1985). There were no cases of failure to take up the intervention. Compliance with the intervention in this study involved the quality of the training offered to TAs, the participation of TAs in the training, children's attendance in the teaching sessions and the implementation of the intervention under the expected conditions. Factors relevant to the identification of non-compliance with the intervention were assessed through the instruments used in implementation and process evaluation.

The intervention team listed under the criteria for fidelity of implementation that all the lessons should be delivered to the children. The inclusion of children who participated in only a limited number of sessions could have affected the results of the impact analysis. Therefore it was relevant for understanding the impact of the intervention to implement an analysis of non-compliance with participation in all the intervention sessions.

TAs had kept records of the pupils' attendance to sessions and these records were provided to the evaluation team. For the purposes of this analysis, compliance with treatment was described as met if the children participated in 30 sessions or more (three children were offered 1, 2, or 3 extra sessions); if they participated in fewer sessions, they were excluded from the analysis.

As it may be difficult for all pupils to attend all 30 sessions, because attendance could vary due to illness or other factors unrelated to the intervention, the distribution of the number of sessions attended was investigated to see what percentage of children attended 30 sessions. The statistical analysis plan anticipated the use of quartiles for the analysis of the effect of dosage on pupils' outcomes, depending on the distribution of the numbers of missed sessions. The statistical analysis plan stated that further cut-off points for attendance would be investigated in order to extract the best information from the data. Of the 211 children (78.7% of all children in the intervention group) for whom we have information, the mean number of sessions attended was 27, which is quite close to all sessions; the mode and median were both equal to 30. Thus the use of quartiles of attendance was unlikely to reveal much about the significance of the dosage for the impact of the intervention.

No child attended less than a quarter of the sessions; three attended less than half of the sessions; eight attended between half and two thirds of the sessions; 14 attended more than half but less than 3/4 of the sessions. Thus setting a criterion of attendance at 2/3 of the sessions excluded fewer children from the analysis than setting it at 3/4, which was desirable because of the loss of power with loss of participants for the analysis. It was therefore decided to use up to three levels of dosage for the analysis of the effect of dosage on impact: (1) attendance to all 30 sessions; (2) if attendance to all sessions proved to be a mediator of implementation efficacy, attendance to at least 2/3 of the sessions (20 or

more sessions) would also be investigated; and (3) if attendance to 2/3 proved to be a mediator of efficacy, attendance to at least half (15 or more) of the sessions would also be investigated. In all these analyses, pupils who had a lower dosage than the one set by the criterion were excluded, and only those who met the criterion were compared to the control group.

Subgroup analyses

Two subgroup analyses were planned in the protocol: one in which the subgroups were defined in terms of eligibility for free school meals ((β_3 Eligible for free school meals (EVER6)$_{ij}$) [5]and the other in which the subgroups were defined by gender. The focus on eligibility for free school meals is in line with the aims of the EEF. The analyses related to gender address a question frequently asked by teachers regarding effectiveness of maths interventions for girls; failure to address this question could result in less interest among schools in the implementation of the intervention in the future. In both cases, a main effect of sub-group membership and an interaction term was added to the equations presented earlier on.

A multilevel model was used, which takes into account the intra-cluster correlation. The formula is presented here with variable names for intelligibility.

For the subgroups defined by FSM status:

$$\textit{Outcome measure}_{ij} = \beta_0 + \beta_1 \text{Pre-test}_{ij} + u_{0j}\text{School}_{ij} + \beta_2 \text{Group allocation}_{ij} + \beta_3 \text{FSM}_{ij} + \beta_4 \text{Group allocation} * \text{FSM}_{ij} + e_{ij}$$

For the subgroups defined by gender:

$$\textit{Outcome measure}_{ij} = \beta_{0j} + \beta_1 \text{Pre-test}_{i} + u_0 \text{School}_{j} + \beta_2 \text{Group allocation}_{ij} + \beta_3 \text{Gender}_{ij} + \beta_4 \text{Group allocation} * \text{Gender}_{ij} + e_{ij}$$

β_0 intercept

β_1 slope for pre-test effect (as a fixed effect, this coefficient is not assumed to vary between schools)

u_{0j} slope for school effect (random effect) (also called level 2 residual by some)

β_2 slope of group allocation (Control vs Intervention at school level, fixed effect)

β_3 slope of free school meals in the top equation and of gender in the bottom equation

β_4 slope of interaction between group allocation and the previous term (free school meal or gender)

e_{ij} = level 1 residual

Sub-group analyses were carried out with MLwiN and SPSS.

Considering that the number of pupils per cluster in these subgroup analyses was reduced because the data were split into two files, for both subgroup analyses single level ANCOVAs were also used to verify whether the same results were observed.

[5] *EVER6* is the name of the variable in the National Pupil Database that indicates whether the pupil was eligible for FSM in the last six years.

Missing data

As per the Statistical Analysis Plan (SAP) missing data were scrutinised at the time of post-test in order to verify the proportion of missing data and the relation between missingness and pre-test results (see below).

Besides analysing the proportion of missing data, we considered its nature (missing completely at random, missing at random, or missing not at random; Graham, 2009) by investigating the reasons for missing data during the process evaluation. We searched for patterns to missing data by analysing the connection between missingness with pre-test measures, with eligibility for free school meals (defined as school descriptors) and with intervention variables (trainer responsible for the trainer). As no single analysis for missing data is definitive, every effort was made during the trial to avoid attrition. Given the small number of pupils per school, every effort was made to return to schools to collect data if pupils were absent on the testing day. In this trial, testers returned to the schools to collect data on pupils absent at the time of the first visit until the only pupils who were not tested were those who had relocated or had been withdrawn from the trial; as mentioned previously, two schools in the control group did not agree to post-testing.

The statistical analysis plan indicated that, if the level of missing data were higher than 5%, an analysis of the reasons for missing data and a search for patterns in the missing data would be implemented, in order to assess whether the data were missing completely at random. If this were not the case, the conclusions from the trial might be compromised. The rate of missing data was 7.7%, thus further analyses to investigate the reasons for missing data and whether there was a pattern to missingness were carried out. The reasons for missing data were investigated by asking schools participating in the post-test to indicate why it was not possible to test some children. The profile of the control schools that did not agree to post-testing was analysed. Regression analyses were used to predict membership in the drop-out cell using different predictor variables (pre-test results; school descriptors such as above the median percentage of pupils eligible for school meals; intervention descriptors - trainer responsible for the training) in order to see whether a pattern to the missing data could be detected.

Additional analyses

One additional analysis was planned in the statistical plan: because the schools were blocked for the randomisation by geographical region and by training session, and it is possible that scores in same block are correlated, an additional model was used to test whether there was an effect of the blocks used for randomisation. Kahan and Morris (2012) suggest that failure to take into account the correlation between groups within blocks produces standard errors for the treatment effect that are biased upwards, resulting in wider 95% confidence intervals, and thus a failure to detect effects that are genuine. This analysis was carried out by using 11 dummy variables for the identification of the randomisation blocks as an additional analysis (there were 12 blocks and one is used as the reference).

Craig et al. (2008) noted that a single measure of primary outcome may not make the best use of the data in a randomised controlled trial. Although it is best to anticipate and plan for all analyses *a priori*, in this trial it was decided to implement two groups of further additional, exploratory analyses *post-hoc* in order to provide a better understanding of the outcomes of the intervention. This is one reason why the evaluation team chose to conduct additional post-hoc analyses not specified in the SAP.

The first set of analyses considered that the parallel forms of the Quantitative Reasoning Test used in the trial for the pre- and the post-test contain a total of 20 items, 12 of which are the same. The full scale score for the post-test, with 20 items, allows for measuring the impact of the intervention, but because the items differ, it does not allow for measuring growth. If the pupils' raw scores are higher in the post-test than in the pre-test (or lower), this could be due to the items being easier (or more difficult). If the pupils' percentile rank scores are higher in the post-test than in the pre-test, this could be due to the phenomenon known as regression to the mean, as the participants in this study were selected because

they were struggling with the mathematics curriculum. Thus the raw scores in a brief scale, which includes only the 12 items that are the same in the pre- and post-test, were also used to explore the impact of the intervention. An increase in the raw score from pre- to post-test measures growth and it is possible to assess whether this growth was greater in the intervention than in the control group. The same multi-level models used with the full scale were therefore used with the brief scale.

The second set of additional, *post-hoc* analysis was a sub-group analysis used to investigate whether pupils who performed below or above the median at pre-test benefitted equally from the intervention. Although teachers were asked to nominate pupils who were struggling with the mathematics curriculum, there was large variation between nominated pupils at pre-test. Therefore, in order to understand better the impact of the intervention, it was considered worthwhile to investigate the possibility of differential impact as a function of the pupils' attainment at the start of the intervention. Pupils were divided into two groups, those who performed up to the median at pre-test and those who performed above the median; the median was chosen as the reference because it provides the best approach to dividing a group into sub-groups of approximately the same number of individuals. The same multi-level models used with the full scale where therefore used with the brief scale. The group of pupils who performed up to the median included 255 children (123 in the control group and 132 in the intervention group) in 100 schools (53 in the control group and 57 in the intervention group). The group of pupils who performed above the median included 236 children (115 in the control group and 121 in the intervention group) in 104 schools (48 in the control group and 56 in the intervention group).

Because splitting the data into two files results in a smaller number of children per school, single level ANCOVAs were used to verify whether the same results would be found with a multilevel and a single level analysis. As indicated earlier on, the use of multilevel models with too few individuals per cluster is questionable and a single level analysis might be more appropriate.

Implementation and process evaluation

The aim of the process evaluation is to understand what activities and factors moderate a successful implementation of the programme and lead to best outcomes for pupils' mathematical skills. The process evaluation aimed to answer two key questions:

The main research question to be addressed by the process evaluation is:

- Does fidelity to treatment moderate the effectiveness of the 1stClass@Number intervention?

A secondary research question to be addressed is:

- Does the use of alternative treatments in the control schools involve the same contents and the same amount of resources as in the intervention schools?

If control schools invested similar resources to support the nominated pupils, and their efforts were as successful as the 1stClass@Number intervention, one could falsely conclude that the 1stClass@Number intervention is not successful. Thus control schools that invested efforts in supporting the nominated pupils were compared with those that did not do so in order to assess the effectiveness of the alternative interventions used in control schools.

The tools and measures used to address these two questions are described in this section as well as the analysies that were implemented. These can be briefly summarised as follows:

Primary question: does fidelity to treatment moderate the effectiveness of the intervention?

In order to support the evaluation team in the design of instruments for analysis of non-compliance and the process evaluation, at the time of the briefing meeting during recruitment and before the start of the intervention, the intervention team presented the evaluation team with their logic model, which contained a list of factors that enable successful implementation of their programme and moderate

outcomes. These factors are grouped here under different headings using the exact description provided by the intervention team in the bullet points in italics.

Training fidelity

- *The teaching assistant and link teacher attend all required training*

Although not mentioned by the intervention team, the evaluation team considered that the delivery of the training sessions could be a factor in the fidelity. If TAs and LTs attended the sessions, but the training session did not achieve its aim, their attendance would not contribute greatly to the success of the intervention. It was usual practice by the trainers to ask the TAs whether they felt prepared to deliver the next lessons before ending the session and they were offered the opportunity to ask further questions, if they did not feel prepared. The evaluation team also asked the TAs to answer a questionnaire that contained items about their views of whether the training they had received has prepared them well for implementing the programme. These questionnaires were answered by TAs during the period when the intervention was being implemented, not at the end of the training sessions.

Implementation fidelity

- Material conditions for the implementation:
  - *A suitable teaching area is provided where children can work undisturbed and without disturbing others;*
  - *TAs' planning and delivery time is protected; and*
  - *The teaching assistant meets at least fortnightly with the class or link teacher*

- Fidelity in the delivery
  - *The setting the scene (which is a series of activities that help the TA identify the starting point for the pupils in the group) is delivered at the beginning of each topic to children in pairs;*
  - *All the lessons are delivered to the children;*
  - *Lessons are delivered at least three times a week; and*
  - *The teaching assistant follows the programme as it is written, making adjustments as necessary using the guidance in the lesson plans*

Although not explicitly mentioned by the intervention team under fidelity in the delivery of the intervention, the evaluation team also considered the TAs' personal views of whether (a) the materials they were given were clear, (b) the lessons fitted in the time slot anticipated, (c) they understood the connections between the concepts and the activities to be implemented, and (d) they felt confident in delivering the programme. These are factors that can be expected to relate to the TAs' ability to follow the programme as it is written and to make adjustments when necessary, when they note individual differences in the children's response to the activities.

Note that a fidelity factor mentioned by the intervention team is that all lessons should be delivered to the children. This aspect of fidelity was measured at the child level, because a TA might deliver all the lessons but some children might not be present at all the lessons.

Tools and measures used in the implementation and process evaluation

The evaluation team analysed the factors described in the preceding section and planned a variety of means to collect data in order to evaluate implementation. Registers of attendance collected by the trainers provided information on whether the TAs and the LTs had attended all the training sessions. Registers of pupil attendance provided information on whether all the 30 lessons were delivered to the pupils.

Observations and questionnaires were used to collect data on fidelity of delivery of the training sessions and of implementation of the intervention by the TAs. Observation of training sessions and of lessons delivered by TAs to pupils aimed to provide a qualitative description of the implementation in order to

assess whether the trial provides a fair assessment of the intervention. However, these were unavoidably restricted to a sample and thus could not be analysed quantitatively as possible mediators of the intervention outcomes. These means of data collection were complemented by a questionnaire, answered by TAs or LTs, depending on who the appropriate respondent was, in order to reach a high number of responses, which would allow for an analysis of whether any particular aspect of fidelity mediated the intervention outcomes. For example, it was possible to observe whether the space used by a TA was in a suitable teaching area when the evaluation team went to a school to observe a lesson, but the number of observations was unavoidably limited. However, it was possible to ask all TAs whether the space where they delivered the lessons was free of disturbances, and then use this variable as a predictor to test whether TAs who used a space free of disturbance attained better outcomes for their pupils.

TAs and LTs questionnaires were available online and as a paper and pencil measure. TAs and LTs were made aware that their answers would be confidential and that only the evaluation team would have access to their answers.

LTs interviews and questionnaires were used to collect data on what business as usual was in both control and intervention schools. Interviews were inescapably confined to a sample of LTs and allowed for a more detailed description of business as usual for children struggling with maths. Questionnaires were used to obtain information from a larger number of schools which could then be used to assess whether the implementation of alternative interventions in control schools might be a factor that compromised the outcomes of this trial.

Analyses of fidelity were carried out within intervention schools. These analyses investigated whether the fidelity factors predicted pupils' outcomes.

Analyses of business as usual considered what happened in control schools; schools that offered any intervention to nominated pupils were compared to schools that had not provided additional interventions to the nominated pupils. If the offer of interventions in control schools had an impact on pupils' outcomes, this would compromise the outcomes of this trial, due to the additional efforts made by control schools during the time of this project.

A description of business as usual in intervention schools was used to verify that the intervention was an addition to the support offered to nominated pupils and did not replace the maths teaching offered to other to pupils.

An overview of these tools and measures is presented in Table 3. The tools used in data collection for implementation and process evaluation are presented in Appendix 3.

Table 3 : Overview of the tools and measures used to describe and measure the fidelity factors identified by the intervention and evaluation teams

Fidelity factor	How the information was collected	What was included as part of implementation and process evaluation
Training delivery	Observation checklist at training sessions	Observers ticked whether the presentation and timing conformed to guidance and whether the REDS model was implemented
Training – TAs perceived value of training	Questions in a TA questionnaire that created a scale	Extent to which TAs felt training had prepared them for delivery
Training – TAs' and LTs' attendance to training	Training registers provided by delivery team	Attendance at training
Implementation	Questions in a TA questionnaire that created a scale	Factors associated with delivery of programme: material conditions for the delivery, TAs' views of how easy it was to implement the programme, and TAs' confidence and experience of the programme
Implementation	Two open questions in the TA questionnaire (what was the best thing and the most challenging thing about the programme)	TAs' experience of the intervention
Implementation	Observation rating scales	Observations of delivery of sessions, including quality, material conditions and adherence to guidelines
Pupil compliance (used in compliance analysis)	Session registers provided by TAs	Pupil attendance
Business as usual	Interviews with teachers/TAs in control schools	To describe what "business-as-usual" meant for control schools
Business as usual	Survey with teachers/TAs in control schools, producing 6 variables	To describe what "business as usual" means in this trial
Business as usual	Interviews with teachers/TAs in intervention schools	Extent to which additional support was being offered (beyond the 1stClass@Number intervention) in intervention schools

Training fidelity

Observations of training delivery

Trainers are given detailed guidance on the delivery of training to TAs. A PowerPoint file is provided to be used during training and the time to be used for each of the sections during the training day is indicated in the notes for trainers. This enables a high level of consistency across trainers in training sessions. The sessions typically involve the model termed REDS (review, evaluate, discuss, and share). TAs have the opportunity to display their pupils' work, comment on implementation, and ask questions. The activities to be enacted during the training day, which involve TAs role-playing their own and the children's roles, are indicated in the notes for trainers for each training session.

The evaluation team observed five training sessions in order to assess fidelity. Sampling of sessions to be observed was intentional: each of the five training sessions that are implemented before the end of the intervention was observed once. In order to be able to cover the different sessions, four of the five trainers were observed once and one was observed twice; one trainer was not observed because priority was given to covering all themes over observing all trainers.

Before the evaluation team became familiar with the nature of the training days, it was planned to discuss with the intervention team their assumptions about the training outcomes in order to make implicit assumptions explicit. However, after the detailed script for each of the training sessions was provided to the evaluation team, it became clear that this was unnecessary. An observation sheet was designed on the basis of the detailed script for the training sessions, which allowed the observers to follow the script as the session progressed. Notes were taken as the session was delivered with respect to the use of the PowerPoint provided for the sessions, the opportunities created for interaction between the TAs and for the TAs to present their pupils' work, the TAs' participation (asking questions, carrying out the activities they were requested to do) and whether the session covered all the points planned for the training. Two observers were present at one session in order to discuss the implementation of the observation scheme and gauge inter-rater agreement. After the observation had taken place, the observers discussed whether there were discrepancies between the script and the delivery of the training session. No further inter-rater checks were carried out because the observation schedule simply required the observer to indicate whether the activities had taken place (for example that the REDS model was implemented) and the time used for the activity matched that indicated in the PowerPoint, with an allowance for variation due to questions being raised by TAs. The training observations took place between September 2016 and February 2017 (see Table 4).

Table 4: Schedule of observations of training days

Training Day	Attendees	Training content
Day 1 (30/09/2016)	TA and LT	▪ Introduction to 1stClass@Number ▪ Training for Topic 1
Day 2 (18/10/2016)	TA	▪ Review, Evaluate, Discuss and Share Topic 1 ▪ Training for topic 2
Day 3 (7/11/2016)	TA	▪ Review, Evaluate, Discuss and Share Topic 2 ▪ Training for Topic 3 and 4
Day 4 (6/12/2016)	TA	▪ Review, Evaluate, Discuss and Share Topic 3 and 4 ▪ Training for Topic 5
Day 5 (24/01/2017)	TA and LT	▪ Review, Evaluate, Discuss and Share Topic 5 ▪ Staff meeting to be delivered by TA

TA questionnaire: questions about perceived value of training

In order to assess the TAs' views of the training as expressed after they had the opportunity to deliver most of the sessions, TAs were asked to answer a questionnaire that included five questions about the training. They were asked whether: (1) they thought the training prepared them to deliver the lessons; (2) they had received clear information about children's mathematical difficulties; (3) they were provided

with clear guidance on how to help children overcome their difficulties; (4) they received clear guidance on how to adapt the lessons for children with different levels of difficulty; and (5) they found the REDS (review, evaluate, discuss, share) model employed in the training sessions useful. Each question was phrased in a positive manner (for example the training I received prepared me to deliver the sessions) and they were asked to indicate, using a five-point scale, whether they disagreed or agreed with the statement (from completely disagree to completely agree, with a middle point that indicated neither agreement nor disagreement). It was expected that these five questions would form a single scale that could be used to assess the training from the TAs' perspective.

TAs' and LTs' attendance

Data about TAs' and LTs' attendance at training sessions were provided by trainers, who scanned the registers of each session and emailed these to the evaluation team. There were no missing data. Two variables were created, one for the attendance of TAs and the other for the attendance of LTs. Training was delivered over six sessions; however, the sixth session took place after the intervention has been completed; its aim is to review the TAs' questions and it was deemed not to be relevant to the quality of delivery. TAs' attendance could therefore vary from 0 to 5. LTs are expected to attend two sessions; this variable could thus vary between 0 and 2. A third variable, binary variable, was created to distinguish those schools that had sent their TAs and LTs to all the training sessions from those that had not.

Implementation fidelity

Implementation fidelity was assessed in two ways: by means of observations of a sample of lessons and by a questionnaire that was answered by the TAs. The questionnaires and the observation schedules were presented to the intervention team before implementation and changes were made to the instruments in order to capture the intervention team's views of how to assess fidelity and non-compliance. This contribution by the intervention team to the instruments used to assess implementation and fidelity provides face validity to the measures used in the process evaluation.

Observation of a sample of lessons

An aspect of fidelity mentioned by the intervention team was delivery of the sessions according to the written guidelines. Fidelity in the delivery of the sessions was assessed through observations of a sample of lessons. TAs receive materials that clearly describe each of the activities that they will be implementing during each lesson, but there is room for considering children's answers and how to approach errors. There is explicit guidance on how to assess children at the start of a new topic (called "setting the scene") and how to adjust the activities to the pupils' performance during and after this initial assessment.

The plan included in the protocol was to collect ten observations, sampling randomly two TAs from each of the five geographical regions identified after recruitment. As one of the regions was subdivided in two training groups, the evaluation team endeavoured to carry out 12 observations, but due to scheduling difficulties it was not possible to observe a TA in one region; 11 lessons, at least one in each geographical region, were observed. The TAs were randomly sampled within the region and the lessons were chosen purposefully. It was decided not to carry out observations of lessons in the first two topics, as these would be at the start of the implementation, and TAs might feel less confident and find the presence of an observer more intrusive. Topic 3 starts in the fifth week of the intervention and after two training days. It was expected that by then TAs would have developed a good understanding of the intervention and acquired some practice in the implementation.

In order to take notes during the lessons, the evaluation team studied the training materials to be used by the TAs and the handbook provided by the intervention team to the TAs. The lessons are highly scripted so these materials were considered a key form of support for the TAs as well as the observers,

who could note whether the lesson had been implemented as written in the handbook and what sorts of adjustments had been made during delivery. Table 5 presents the list of observation dates and places, topics, and contents of the lessons. Appendix 3 presents an example of how notes were taken during the session.

Table 5: Details of the observation dates and schools visited

School cluster, block and date of observation	Topic and lesson	Content of lesson
Doncaster, Block1 23/11/2016 PM	Topic 3, setting the scene	Addition and subtraction, talking about addition and subtraction, understanding symbols (-, +, =).
Huddersfield, Block 3 17/01/2017	Topic 4, lesson 2	Counting in 10's then 1's using coins, addition and subtraction game, writing number sentences.
Huddersfield, Block 4 28/11/2016	Topic 3, lesson 4	Finding missing numbers on a number track, number sentences, addition and subtraction game.
Leeds A, Block 5 24/11/2016	Topic 3, setting the scene	Addition and subtraction, talking about addition and subtraction, understanding symbols (-, +, =).
Leeds A, Block 6 20/01/2017	Topic 5, lesson 3	Counting on and back in 2's, number patterns, exploring halving and remainders.
Leeds B, Block 7 18/01/2017	Topic 3, lesson 4	Finding missing numbers on a number track, number sentences, addition and subtraction game.
Leeds B, Block 8 24/11/2016	Topic 3, lesson 2	Counting in multiples of 10 using 10p coins, writing addition facts and number sentences.
Sheffield, Block 9 23/01/2017 AM	Topic 4, lesson 4	Counting on and back in 2's, number pairs to 20, subtraction from 20
Sheffield, Block 10 29/11/2016	Topic 3, lesson 5	Addition and subtraction facts for 10, counting on from and back from
Wakefield, Block 11 23/11/2016 AM	Topic 4, lesson 1	Counting in 10's, addition facts within 20
Wakefield, Block 12 23/01/2017 PM	Topic 4, lesson 1	Counting in 10's, addition facts within 20

Three rating scales were devised to summarise the overall adherence to the script. It was necessary to use summary ratings, rather than direct scoring of each factor in the session, because each topic in the intervention is delivered over six lessons and some aspects are not present in all the lessons. This variation would lead to missing data if a rigid observation rating by item had been used. Observers also asked whether the space used on the specific occasion was always used for the lessons or whether it varied, in order to complete the rating.

Two researchers discussed the ratings as they considered the notes from the observations. Each of these ratings gave origin to one variable.

- Quality of delivery, which considered whether the TA had the required materials and knew the lessons script, whether the children were engaged in the tasks as directed by the TAs, whether the mathematical language and the materials recommended in the lesson plan were used, whether the TA answered pupils' questions, provided feedback and support, when required, and followed up on children's answers, and whether there were adaptations when children seemed either to require further support or seemed to be working at a higher level than expected in the lesson. This scale was not based exclusively on the intervention team's indicators of fidelity. Although pupil engagement or TAs' consistently answering the children's questions were not explicitly mentioned by the intervention team, if a pupil did not carry out the activities planned for the lesson or the TA failed to address a pupil's question, one could say that the pupil was physically present but not attending and that the TA had followed the script but the pupils' questions were treated as not part of the lesson. The other aspects (having the materials ready, knowing the lesson script, using the mathematical language, and adapting the level of the activities) are part of the intervention team's indication that the TA should follow the written lesson plan. Ratings in this scale varied from 0 to 5.

- The material conditions of delivery during the lesson, which considered whether the TA had delivered the activities and the whole lesson according to the expected time, whether all the resources required for the lessons had been prepared and used, whether the space was adequate and whether there was space to display the children's work, which is considered a motivating factor for pupils. Ratings in this rating varied between 0 and 5.

- Adherence to the written guidelines, which considered whether all activities were covered in the lesson, whether aspects of the lesson that are expected to be delivered in pairs or with the whole group were delivered accordingly, and whether there had been interruptions that disturbed the delivery of the lesson. These were factors of fidelity in the implementation in the intervention team's logic model. Ratings in this scale varied between 0 and 10 because there was much room for noting that the TA either followed or departed from the lesson as written.

In all three scales, the higher the number, the closer the TA's delivery was to the plan. The aim of these observations was to verify whether it is reasonable to assume that the training had its expected impact on TAs and that the intervention was delivered as intended. They cannot be used to investigate whether adherence to the script mediates the impact of the intervention because only a sample of schools was involved in the observations. Although the number of observations is small, it can still be used for a description of these three different aspects of delivery.

TA questionnaire

The TA questionnaire contained questions about their views regarding training, mentioned earlier on, and questions related to the fidelity factors listed by the intervention team in their logic model. These can be grouped in three categories: the material conditions of delivery, the TAs' own perception of how easily they could implement the written guidelines, and their confidence in their understanding of the intervention programme. The tools used for assessing each one of these aspects in described in turn.

Material conditions:

Six questions about the material conditions were included in the questionnaire (four referred to the space available and two to time for preparation and meeting with the LT). It was not expected that these questionnaire items would form a scale as they measure different aspects of the material conditions; however, it was still possible that they would be correlated, as a school that protected the TA's time to

prepare and to meet with the LT might have been more likely to identify a suitable space for the delivery of the lessons.

Fidelity factors reported by TAs regarding delivery of the intervention:

These were assessed by means of the questionnaire that was answered by TAs. The questions focused on their use of the materials required during the lessons, fitting all parts of the lessons into the time slots available, number of topics covered by the end of the intervention, how setting the scene was delivered, and how easy they found it to follow the lesson plans. The answer format for these items varied: for example, when answering how the setting the scene was delivered, the TAs could tick that they either did it individually, or in pairs, or to the whole group; when answering about the topics delivered, they ticked those topics that they expect to cover by the end of the intervention; when answering how easy it was to follow the lesson plan while delivering it and how easy it was to fit all the activities in the lesson, they could tick whether it was 'very difficult', 'very easy', or 'neither difficult nor easy'. For two items, about setting the scene and about number of topics delivered, a dummy variable was created, in which 1 indicated fidelity (that is, setting the scene in pairs and delivery of all topics) and 0 indicated non-compliance. For the other three items, a three-point scale was used, in which 1 indicated low fidelity (didn't always use all the materials, it was difficult to fit the activities in the lesson and to follow the written plan while delivering the lesson) and 3 indicated high fidelity, whereas 2 was the midpoint.

Each of these aspects was coded as a variable, forming five variables, so that it was possible to investigate whether these aspects of fidelity mediate the impact of the intervention. The items that assess the TAs' perception of fidelity were not expected to form a scale: for example, a TA who covered all the topics might not have been able to deliver the lesson in the time frame expected or might have delivered the setting the scene with the whole group rather than in pairs, as recommended in the guidelines.

TAs' confidence with and experience of the programme:

Although TAs' confidence in their own ability to deliver the programme was not mentioned by the intervention team as a fidelity factor, the evaluation team hypothesised that TAs who did not feel confident in their understanding of the programme or confident to deliver it might not be able to implement it well. TAs were asked to assess their understanding of the programme (for example, understanding of the handbook and of the connection between the focus of the lessons and the activities) and their own confidence in the delivery of the intervention (confidence to adapt the lesson for different levels of pupil ability, to use this knowledge when teaching in the classroom) and their willingness to deliver it again; six variables were formed with answers to these items. In each of the items, the TAs were presented with a positive statement and they could either disagree or agree with it ('completely disagree' to 'completely agree'), forming a 5-point scale. Because these items measure subjective factors regarding the TAs' assessment of their own understanding and confidence in delivering the intervention, it was hypothesised that these items would form a scale, which could then be used to test whether the TAs' subjective experience of the intervention was a mediator of the intervention's impact.

After answering these multiple choice questions, TAs were invited to offer comments on what they thought was the best thing and the most challenging thing about the intervention. A theme analysis was carried out to identify those which were cited most often. The analysis was descriptive in order to provide an overview of the comments as offered by the TAs.

Internal consistency analyses (Cronbach's alpha) were applied to the different groups of items used in the TAs questionnaires in order to see whether they formed a single scale. When a consistent scale was identified, the score in the scale was used to investigate whether these aspects of implementation predicted pupil outcomes. When the items did not form a consistent scale, the items were analysed separately to investigate whether they related to pupil outcomes.

Pupil compliance

TAs scanned the registers for the lessons and provided to the evaluation team information on attendance, which defines the dosage of the intervention received by pupils. The effect of number of lessons attended on pupil outcomes is reported in the section about impact in the presence of non-compliance.

Secondary question: what resources and programmes were available to pupils in control schools?

Interviews and questionnaires were used to describe what "business as usual" meant in the control schools. LTs and TAs in intervention schools answered a parallel questionnaire as we wanted to check whether the pupils participating in the 1stClass@Number intervention had missed out on other opportunities offered to pupils struggling with maths because teachers thought that the intervention was replacing all other support that they required.

Business as usual

Interviews with LTs

In order to describe what "business-as-usual" meant for control schools, a sample of the LTs nominated for this role before randomisation participated in a phone interview. The aim of the interview was to describe the percentage of the nominated pupils in the control schools who had received extra support during Year 2, what type of support, if any, and what mathematical topics had been covered. LTs in intervention schools were also interviewed in order to ascertain whether the nominated pupils received the regular maths lessons and the additional support available to other pupils who had not been nominated for the trial – that is, whether the intervention was truly in addition to regular maths teaching.

Teachers in both intervention and control schools (n=10 in each group) were sampled randomly to participate in the interviews. The analysis of the interviews can only be qualitative, as not all schools in the trial participated in interviews. Due to scheduling difficulties, the LT in one control school was not interviewed.

TA and LT questionnaire for control schools

TAs and LTs in control schools were asked to complete a questionnaire to help describe what "business as usual" meant in this trial. The answers produced dichotomous variables (was any additional support offered to the nominated pupils?), nominal variables (which materials were used with the pupils, who delivered the intervention, was the intervention new to the school) and continuous variables (number of topics that are part of the 1stClass@Number intervention that were taught to the nominated children in these interventions; estimated number of sessions). Six variables were created on the basis of the answers to this questionnaire.

Our analyses aimed to evaluate whether outcomes in the control schools were associated with these variables. We tested for a difference between pupils in control schools who did vs did not receive an alternative intervention and whether pupils who were not taught about the topics included in 1stClass@Number differed from those pupils who received interventions that covered the topics included in the 1stClass@Number intervention. After identifying which alternative interventions were most often used, we assessed whether controls schools that did vs those that did not use this intervention differed from each other. The statistical analysis plan anticipated that, if the alternative intervention had an effect on pupils' outcomes, pupils in the group that received this intervention would be compared with the 1stClass@Number intervention.

Costs

Cost information was obtained directly from the intervention team and through a questionnaire answered by the LT. The intervention team provided information regarding the cost per school if the school were to sign up for training outside the trial. The LT provided information on actual costs to the school in order to implement the intervention and on the time spent for training and for delivering of the intervention.

Timeline

There were no changes to the timeline agreed at the protocol stage, and summarised in Table 6.

Table 6: Timeline

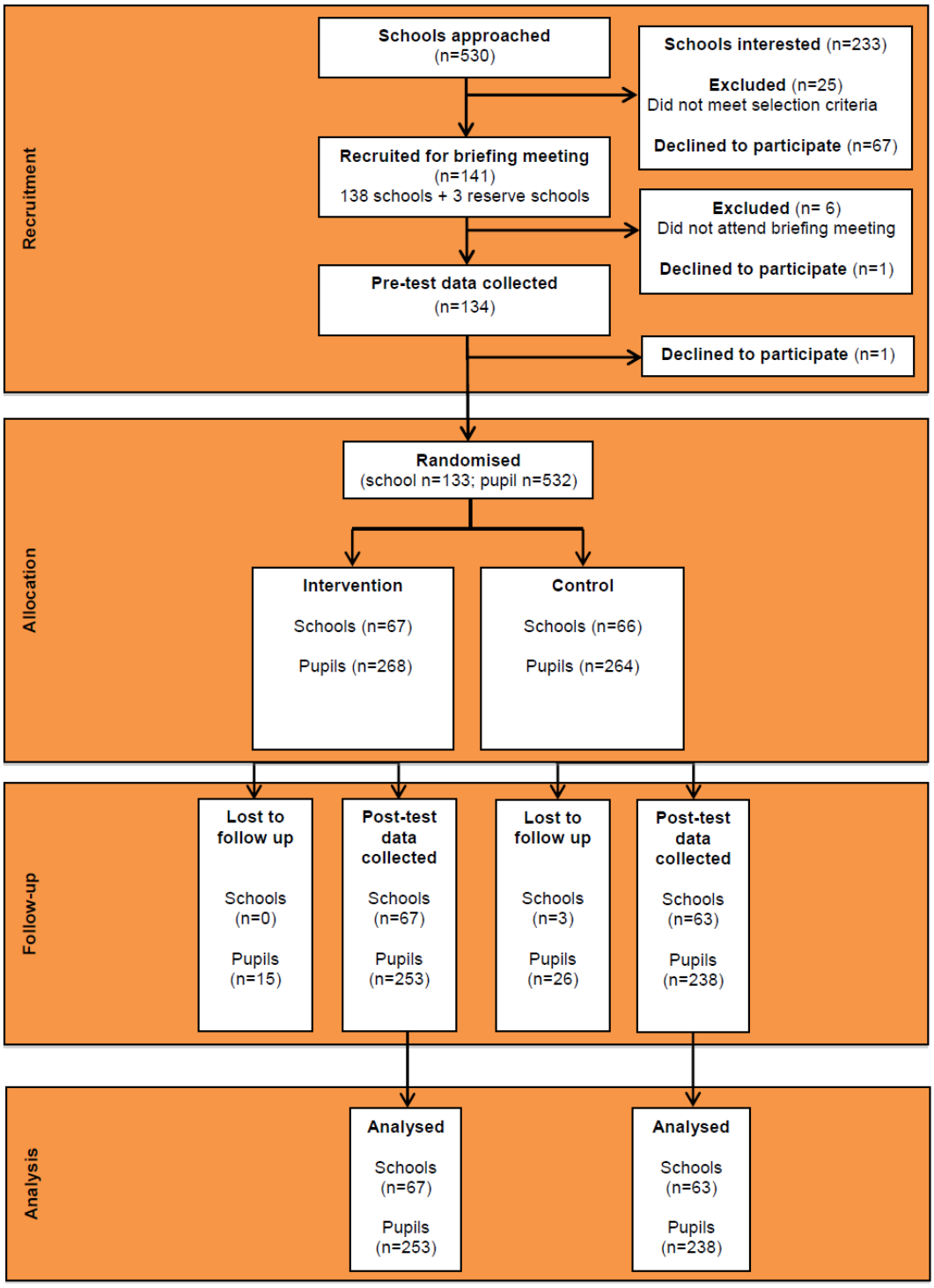
Date	Activity
01/09/2015-31/05/2016	Recruitment and briefing meetings
01/06/2016-31/08/2016	Pre-tests administered by class teachers and marked by evaluation team
01/09/2016-30/09/2016	Communication of results to schools, nomination of pupils, TAs and LTs, randomisation, communication of group membership to schools
01/10/2016-15/07/2017	Training of TAs and implementation of intervention, process evaluation, training of testers, scheduling and application of post-tests
16/07/2017-30/09/2017	Marking of post-tests, application to NPD for KS and FSM data
01/10/2017-30/04/2918	Data analysis and preparation of report

Impact evaluation

Participants

Figure 3 provides an overview of the participant flow through the different stages of the project. The intervention team approached 530 schools; 232 expressed an interest; 25 schools were told they were not eligible for the trial due to the criteria explained earlier on; 32 declined to join the trial and 15 did not meet the inclusion criteria; 138 schools were invited to a briefing meeting; 4 that were not represented at the meeting were excluded; 1 school withdrew previous to randomisation and after post-test; 133 schools were randomised (4 pupils per school: N=532 pupils); 1 school that had been assigned to the control group withdrew after randomisation due the a long-term illness of the TA who would have been implementing the alternative intervention with a different year group. After randomisation, if a child moved to another school and a new child was included in the small group to maintain its expected size of 4 children, the new child was not included in the study. The power calculation done *a priori* estimated that 120 schools would be sufficient to detect a larger effect size than that expected on the basis of past information about the intervention. The evaluation team suggested recruiting 130 in order to prevent a loss of power due to attrition. The final number of schools in the analysis was 130, which is sufficient to detect the estimated effect size of 0.22.

Figure :. Participant flow diagram



No intervention school dropped out but one intervention school was not able to complete the implementation of the intervention due to a long-term illness of the nominated TA. This was communicated to the intervention and evaluation teams after the first topic had been taught and subsequent lessons had been suspended. As this is an intention to treat design, the post-test results were collected and analysed. Three control schools (a total of 12 pupils) did not agree to participate in the post-test.

Of the 520 pupils who remained in the sample after the withdrawal of these control schools, one pupil was not post-tested because the parent withdrew the child from the trial and one pupil was not post-tested because the teacher withdrew the child from the trial due to a diagnosis of autism after pre-test; 27 pupils were not post-tested due to relocation; 491 (92.3%) pupils were post-tested and included in the analyses. The total loss of participants on the primary measure either due to school withdrawal or to loss of pupils at post-test is 7.7% (41 pupils). As this is higher than 5%, the analysis of missingness searched for patterns in the missing data, as anticipated in the statistical analysis plan. Table 7 provides a description of the minimum detectable effect size at the time of the design of the protocol, after recruitment and randomisation were completed, and at the time the analyses were implemented.

Table 7: Minimum detectable effect size at different stages

Stage	N [schools/pupils] (n=intervention; n=control)	Correlation between pre-test & post-test	ICC	Blocking/ stratification or pair matching	Power	Alpha	Minimum detectable effect size (MDES)
Protocol	120 schools/480 children (60 intervention; 60 control)	0.70	0.15	School blocking, 12 blocks	80%	0.05	0.21
Randomisation	133 schools/532 children nominated (66/264 control; 67/268 intervention)	0.70	0.15	School blocking, 12 blocks	80%	0.05	0.21
Analysis (i.e. available pre- and post-test)	130 schools/491 children (63/238 control; 67/253 intervention)	0.29	0.22	School blocking, 12 blocks	80%	0.05	0.27

As indicated earlier, the correlation between the pre- and the post-test for this sample was much lower than that observed for the whole sample tested in South and West Yorkshire (which was 0.63), as a consequence of the limited range of scores due to the sample selection. The ICC used for the protocol was 0.15, which is in line with the ICC observed for KS Maths outcomes in Yorkshire (see: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf). The ICC was also different from that expected on the basis of previous studies with mathematics attainment as an outcome measure; for this sample, it was 0.22 for the pre-test and for the post-test with the Quantitative Reasoning Test.

The correlation between the pre-test and KS1 Maths for this sample was 0.26, much lower than the correlation for the whole sample tested, which was 0.63 (identical to the correlation between the pre- and post-test in Quantitative Reasoning). The lower correlation is due to the limited range of scores in this sample because the sample was selected to fit the criterion of “struggling with maths” (that is, showing lower attainment and not covering the whole range of attainment). The ICC for KS1 Maths outcome was 0.15.

Pupil characteristics

Table 8 provides a comparison between intervention and control schools with respect to type of school (academy or Local Authority), Ofsted rating, school location (urban vs rural), percentage of pupils eligible for Pupil Premium, and the pre-test results.

Table 8: School and pupil characteristics by trial group

Variable	Intervention group (N=67)		Control group (N=66)	
School-level (categorical)	**n/N (missing)**	**Percentage**	**n/N (missing)**	**Percentage**
School Type[1]				
Academy	20/67 (0)	30%	10/63 (0)	16%
LA school	47/67 (0)	70%	53/63 (0)	84%
Ofsted rating[2]				
Outstanding	11/67 (0)	16.4%	13/63 (0)	20.6%
Good	41/67 (0)	61.2%	42/63 (0)	66.7%
Requires Improvement	7/67 (0)	10.4%	6/63 (0)	9.5%
Inadequate	4/67 (0)	6%	1/63 (0)	1.6%
New (not yet rated)	4/67 (0)	6%	1/63 (0)	1.6%
Location[3]				
Urban - (city, town and conurbation)	59/67 (0)	88%	57/63 (0)	90.5%
Rural - (hamlet, village, town and fringe)	8/67 (0)	12%	6/63 (0)	9.5%
School-level (continuous)	**n (missing)**	**Mean**	**n (missing)**	**Mean**
No. pupils per school[1]	67 (0)	267	63 (0)	280
No. eligible for pupil premium[1] **N(%)**	67 (0)	67 (25.5%)	63 (0)	89 (30.5%)
Pupil premium allocation 2016/17[1] **(£)**	67 (0)	89,809	63 (0)	117,009
Pupil-level (categorical)	**n/N (missing)**	**Percentage**	**n/N (missing)**	**Percentage**
Eligible for FSM	67 (14)	26.4	100 (12)	39.7[4]
Pupil-level (continuous)	**n (missing)**	**Mean 95% CI)**	**n (missing)**	**Mean (95% CI) [Effect Size]**
Pre-test score[6]	268 (0)	4.92 (4.62 – 5.22)	264 (0)	4.80 (4.49 – 5.11) [Cohen's d=0.05]

[6] At the time that the SAP was written (August 2016), the machine reading of the marking sheets had not been double checked. There was some error in this reading, which was identified and corrected before the information about the children's scores was provided to the teachers in September. All the analyses are based on the corrected data set.

Age in months at post-test	259 (9)[7]	87.19 (86.77 – 87.63)	244 (20)	86.67 (86.23 – 87.10)

[1]**Source:** https://www.gov.uk/government/publications/pupil-premium-conditions-of-grant-2016-to-2017, accessed 16/11/2017

[2]**Source:** https://reports.ofsted.gov.uk/**,** accessed 21/11/2017

[3]**Source:** https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2017**,** accessed 11/12/2017

[4] Percentage of valid cases; 26 parents did not agree to the release of the UPN

When the intervention and control schools were classified by type of school, it became clear that there was a larger proportion of academies in the intervention group than in the control group. Because type of school is a categorical variable, which cannot be treated as a number, the association between trial group and type of school was investigated by means of a Chi-square, which is the appropriate statistical test for assessing whether two categorical variables are associated. If the Chi-square is significant, this means that there is an association between the two variables; in this case, it would mean that there is a greater likelihood of an academy being in the intervention than in the control group. The association between the two variables in this case did not reach significance statistically (Chi-square=3.57; df=1; *p*=0.06); thus it can be concluded that it was not more likely for academies to be in the intervention than in the control group.

In the categorisation by Ofsted rating, there was variation with respect to the year when the inspections had been carried out. The most recent inspections were conducted in 2017 and the oldest was conducted in 2006. This interval of about ten years could be related to changes in the criteria for Ofsted ratings because these ratings are related to the quality of implementation of the National Curriculum and there have been changes in the mathematics curriculum in this period. When academy converters had not yet received an Ofsted inspection since converting, their previous establishment rating was used. The evaluation team considered that the variation in time of inspection and the need to use inspection ratings before a school converted to academy made the ordering of the Ofsted ratings less reliable than desired in a measure. Thus we decided that the ratings would not be treated as a number in this specific sample of schools and opted for treating them as categories. Five schools were new and had no rating yet, so they were not included in the analysis of the association between Ofsted rating and trial group. A Chi-square test showed that there was no association between Ofsted rating and trial group (Chi-square=2.05; df=3; *p*=0.53), which indicates that it was not more likely that schools in the intervention group had one type of rating than any other.

The school location, urban vs rural, is a categorical variable and it was not associated with trial group either (Chi-square= 0.20; df=1; *p*=0.66).

The national percentage of pupils in primary schools (mainstream and mainstream academy) eligible for the Pupil Premium in 2016/17 was 23.2%. The national average Pupil Premium allocated to each primary school was £78,654. In the schools in this trial, the percentage of pupil eligibility for Pupil Premium was 27.9% and the average allocation to schools was £102,991. Both control and intervention schools were above average in percentage of pupils eligible for Pupil Premium and the corresponding funding. In order to analyse whether there was an association between trial group and the percentage of pupils in the schools that were eligible for Pupil Premium (which is similar to everFSM in the NPD),

[7] Date of birth was collected from schools, not from the NDP, as date of birth is used to help match the pupils. We systematically asked the schools to complete this information when it was missing but sometime were not able to obtain it from the schools. However, we have more data on the pupils' date of birth than we could obtain through the NPD.

the number of pupils eligible for Pupil Premium in each school was estimated and added across schools in each group; the frequencies of pupils eligible and not eligible for Pupil Premium by trial group were calculated. There was no association between these two categories, trial group and eligibility for Pupil Premium (Chi-square=0.08; df=1; *p*=0.78).

Pre-test scores did not differ significantly between the two groups; the difference in pre-test scores in favour of the intervention group was equal to Cohen's d effect size of 0.05 and was not statistically significant. The schedule of post-tests purposefully mixed control and intervention schools during the testing period, but the pupils in the intervention schools turned out to be slightly older than those in the control schools at post-test; however, the difference was not statistically significant, (t=1.89; df=488; *p*=0.06). In order to assess whether this posed a threat to the validity of the trial, we calculated the correlation between age at post-test and post-test scores; this correlation was negligible and not significant (r=0.04; *p*=0.334).

In summary, even though all the comparisons between schools and between pupils at baseline indicate that the differences between the groups were not statistically significant, two of the comparisons were at the borderline for significance: the number of academies in the intervention group was higher and the number of pupils in the intervention group eligible for FSM was lower. The small difference between the intervention and the control group on pre-test scores can be attributed to chance and is too small to be a cause for concern with respect to the validity of the trial. Although the differences at school or at pupil level do not seem to pose a threat to the validity of the trial, it is advisable to investigate whether the effect of the intervention is affected by the introduction of an extra covariate in the analysis at pupil level: the eligibility for FSM.

Outcomes and analysis

Impact

The primary analysis of the impact of the intervention was implemented as proposed in the statistical analysis plan (https://educationendowmentfoundation.org.uk/public/files/Projects/Round_8_-_1st_class_at_number_SAP.pdf). The primary impact analysis is a multilevel model of covariance used to predict the scores in the Quantitative Reasoning Test at post-test, which takes into account the fact that pupils are nested within schools, and includes as a covariate the pupils' performance in the Quantitative Reasoning Test before the intervention. An analysis using the same model was also implemented with the secondary outcome measure, KS1 Maths (variable from the NPD used: KS1_MATH_OUTCOME). Effect sizes were calculated using the standard deviation for the whole sample. Table 9 presents the raw means by group in the primary and secondary measures. Raw scores were used in all the analyses.

Table 9: Primary analyses of the outcome measures

	Raw means					Effect size	
	Intervention group		Control group				
Outcome	N (missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning	253 (15)	9.75 (9.18 – 10.32)	238 (26)	9.04 (8.48 – 9.60)	491	0.18 (-0.08 to 0.43)	.09
KS1 Maths	251 (17)	3.33 (3.26 – 3.42)	252 (12)	3.30 (3.22 – 3.38)	503	0.04 (-0.18 to 0.28)	.50

All analyses were carried out using the eefAnalytics programme in R and checked by running the analyses in MLwiN. The results were always consistent; the details of the analyses with the eefAnalytics are presented in Appendix 4. The results are summarised in this section.

Results of the primary analyses

In the multilevel model with the **Quantitative Reasoning Test** as the outcome, the intra-cluster correlation (ICC) was equal to 0.22 at pre- and post-test. There was a significant contribution of the pre-test to the prediction of the outcome scores. The effect size for the impact of the intervention was equal to two months progress (0.18), but it was not significant statistically meaning this could have occurred by chance (see Table 9).[8]

In the analysis of the impact of the intervention on the secondary outcome**, KS1 Maths**, the number of participants included is not the same as the number for the primary outcome because 26 parents did not agree to releasing their children's UPNs (unique pupil identifier), so it was not possible to obtain their KS1 Maths results. However, KS1 Maths scores were available for pupils from the three control schools that withdrew from the trial because permission to obtain their UPNs and KS1 results had been obtained and was not withdrawn; scores were also available for pupils who had not been post-tested; there were missing data for three children from different schools in the NPD file; this makes the number of participants in these analyses equal to 503.[9]

A multilevel model, which took into account the fact that pupils were nested in schools, was used to investigate the impact of 1stClass@Number on KS1 Maths results. The pupils' scores in the pre-test were entered in the analysis as a covariate. The ICC was equal 0.15; the pre-test scores contributed significantly to the prediction of KS1 Maths attainment. The effect size for the impact of the intervention was very small (0.04 SD) and was not statistically significant.

KS1 Maths outcomes uses five ordered categories, and these were treated as a scale in the previous analyses, which gives more power to the analysis, uses all the data and allows for the investigation of

[8] Considering that the number of pupils per school was smaller than five, a single level ANCOVA was carried out, controlling for pre-test scores. This analysis produced the same results as the multilevel model; the 95% confidence interval for the difference between the means was from -0.10 to 1.44; p=0.09. Considering the imbalance in the percentage of pupils eligible for FSM between the groups, a second single level ANCOVA was carried out, controlling for eligibility for FSM. This analysis produced a lower effect size (Hedge's g = 0.12; CI: -0.07 to 0.30).

[9] A single level ANCOVA was also carried out with KS1 Maths outcomes, which replicated the very small and non-significant impact of the intervention; the effect size was 0.01.

interaction effects in the subgroup analyses. However, some of the assumptions for the use of parametric statistics are not met by the KS1 Maths levels and there is a risk of obtaining unreliable results with linear regression analyses. In order to test whether these results were replicated when the assumptions required for linear regression are not made, a multinomial analysis for ordered categories was implemented. Due to the low frequency of children performing at the highest level (8/503 or 1.6%), the data in this category were excluded from the analyses. The two lowest categories also had very low frequencies and were merged into a single category in order for a reasonable number to be reached for inclusion in the analysis (27/503 or 5.4%). The details of the analysis are presented in Appendix 4. The results of this new analysis converged with those obtained through a linear regression and showed a non-significant impact of the intervention (z=1.472; *p*=0.14).

Additional analyses

One further multilevel model with the Quantitative Reasoning Test as the outcome measure was carried out, as specified in the statistical analysis plan, which took into account the effect of the block used for randomisation. Kahan and Morris (2012) suggest that failure to take into account the correlation between groups within blocks could result in failure to detect effects that are genuine. In this analysis, the correlation between schools in the same blocks was not significant and thus the inclusion of the randomisation block in the analysis did not lead to any significant change in the results (see details of this analysis in Appendix 5). After controlling for randomisation block, the impact of the intervention continued to be small and was not statistically significant.

A similar analysis, which took into account the randomisation block, was carried out with KS1 Maths outcomes. After controlling for randomisation block, this multilevel model did not show a significant impact of the intervention on KS1 Maths (see details of these analyses in Appendix 5).

Thus there is no evidence of an impact of the 1stClass@Number intervention either on the Quantitative Reasoning Test or on KS1 Maths attainment when the randomisation block is taken into account.

Subgroup analyses: socio-economic background and gender

Two subgroup analyses were proposed in the statistical analysis: the first in which the subgroups were defined by the eligibility for free-school meals in the previous six years (the variable from the NPD was EverFSM) and the second in which the subgroups are defined by gender. Because 26 parents did not agree to the disclosure of their children's UPNs, which provided access to the information regarding eligibility for FSM, the number of participants in the analyses that refer to analyses with FSM eligibility is 506. Appendix 6 presents the details of the analyses with subgroups defined by eligibility for FSM and Table 10 presents the means for the subgroups and the effect sizes. Appendix 7 presents the details of the analyses with subgroups defined by gender and Table 11 presents the means and effect sizes for the subgroups defined by gender.

Comparison between pupils eligible and not eligible for FSM

In order to contextualise the differences between pupils eligible vs not eligible for FSM, the first comparison carried out was between these two subgroups at pre-test and included all pupils in the participating schools (N=5041; n= 3823 for pupils not eligible for FSM and n= 1219 for those eligible for FSM).

Pupils eligible for FSM tend to show lower attainment in school; thus it was expected that they would perform significantly lower in the pre-test than those not eligible for FSM and thus it was more likely that they would be nominated for participation in the trial than those not eligible for FSM. This was indeed the case: the mean at pre-test for the pupils not eligible for FSM was 9.89 and for those eligible for FSM was 8.19; this difference corresponds to an effect size (Cohen's d) equal to 0.37 and was statistically significant (t=11.20; df=5040; $p<0.001$).

This difference in performance led to difference in nomination rates: 13.8% of the pupils eligible for FSM were nominated for participation whereas 8.9% of those not eligible for FSM were nominated. A Chi-square showed that the association between subgroup membership and nomination was significant (Chi-square = 24.515; df=1; $p<0.001$), but this does not indicate a bias in nomination, because the pupils eligible for FSM had lower scores in the pre-test than those who were not eligible for FSM.

In the sample of pupils nominated for participation in the trial, the pre-test mean for those eligible for FSM was 4.72 and for those not eligible for FSM the mean was 4.89. This difference, which corresponds to an effect size equal to 0.07, was not statistically significant (t=0.7; df=504; $p=0.49$) and could have happened by chance. This suggests that there was no bias in the nomination of pupils as far as their eligibility for FSM is concerned.

Because in the sample of pupils nominated for the trial there was no difference between those eligible and those not eligible for FSM, one could expect that there would be no difference between these two groups in the outcome measures given at the end of the year; pupils eligible for FSM in the control group should make as much progress as those not eligible for FSM and the same should be the case for those in the intervention group. A difference would only be expected if, either in the control group or in the intervention group, the pupils eligible for FSM made less (or more) progress than those not eligible for FSM.

Table 10: Subgroup analyses by FSM status[10]

FSM status: Not eligible							
	Raw means					Effect size	
	Intervention group		Control group				
Outcome	n/n schools (missing)	Unadjusted Mean	n/n schools (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning	179/63 (8)	10.11	138/56 (14)	9.41	317 (22)	0.11 (-0.06 to 0.38)	.19
KS1 Maths	186/63 (1)	3.41	152/59(0)	3.32	338 (1)	0.14 (-0.08 to 0.35)	.20
FSM status: Eligible							
	Raw means					Effect size	
	Intervention group		Control group				
Outcome	n/n schools (missing)	Unadjusted Mean	n/n schools (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning	60/37 (7)	8.62	89/48 (11)	8.44	149 (18)	-0.03 (-0.30 to 0.36)	.92
KS1 Maths	65/39 (2)	3.14	100/52 (0)	3.27	165 (2)	-0.25 (-0.57 to 0.06)	.12

The means displayed in Table 10 show a difference in the means at post-test in favour of the pupils not eligible for FSM both in the control and in the intervention group. This is the case with respect to the primary outcome, Quantitative Reasoning Test, and the secondary outcome, KS1 Maths. In the intervention group, the difference between the pupils eligible and not eligible for FSM increased from pre- to post-test in the Quantitative Reasoning Test (SD=4.53): The Cohen's d effect size was 0.33 in the intervention group and in the control group the effect size was 0.21 at the end of the project. These results suggest that, irrespective of group membership, pupils eligible for FSM made less progress during the year than those not eligible for FSM. The effect sizes related to the impact of the intervention are positive for the pupils not eligible for FSM and negative for those eligible for FSM. However, the differences between the intervention and the control group could have happened by chance, because the differences were not statistically significant.

[10] Means in the table are raw means; the means used for calculation of the effect size are adjusted means. This explains that the raw difference between the intervention and control groups is in favour of the intervention group but the effect size is negative.

In the multilevel model that included FSM status, and an interaction term between FSM status and trial group (intervention vs control), neither the main term nor the interaction effect was significant when the analysis was run with the Quantitative Reasoning Test as the outcome. This means that the difference between the pupils eligible and not eligible for FSM remained non-significant statistically at post-test. In other words, neither the pupils eligible for FSM in the control group nor those in the intervention group made less or more progress than those not eligible for FSM from pre- to post-test.

In contrast, when the outcome measure was KS1 Maths, the main term FSM was not significant, but there was a significant and negative interaction between FSM eligibility and trial group, which indicates that the pupils eligible for FSM who received the intervention benefitted less than those not eligible for FSM from the intervention (see Appendix 6). This result indicates that the intervention did not work towards closing the gap between pupils eligible for FSM and those not eligible but rather widened it.

A sub-group analysis was carried out with each of these groups treated separately. Table 10 shows the number of pupils and number of schools in these analyses. The number of schools with at least one child eligible for FSM is 85, which is still appropriate for multilevel analyses, according to Mass and Hox (2005), but the number of pupils per school was even smaller than in the previous analyses (on average, less than two observations per unit); the number of schools with pupils not eligible for FSM was119 and the number of pupils was approximately 2.67, which is small as the number of observations per unit. Thus the results of these multilevel analyses by sub-groups must be interpreted with caution.

However, the analyses are still worth implementing because, given the relatively lower progress of the pupils eligible for FSM in comparison to those not eligible, the inclusion of pupils eligible for FSM in the overall analysis could have diluted the impact of the intervention. A sub-group analysis using a multilevel model similar to the previous ones, but including only the pupils eligible for FSM, did not show a statistically significant negative effect of the intervention; the effect size was negative but small (see Table 10). This result suggests that pupils eligible for FSM in the intervention group did not have lower KS1 Maths outcomes than those in the control group who were also eligible for FSM. This result must be considered with caution as the number of participants in the analysis is smaller than the number included in the model that investigated the interaction effect. It must also be noted that the result does not contradict the finding described in the previous paragraph: the comparison in this latter analyses is only between pupils eligible for FSM in the intervention and control group whereas in the previous paragraph the comparison was between pupils eligible for FSM and those not eligible in the intervention group.[11]

In summary, separating out the pupils by their eligibility for FSM did not alter the previous picture of the impact of the intervention: it was small for both groups and could have happened by chance. Although the negative interaction between intervention and FSM eligibility when the outcome measure is KS1 Maths gives cause for concern regarding the use of this intervention with pupils eligible for FSM, it is noted that the interaction term between FSM and intervention group was not significant when the Quantitative Reasoning Test was the outcome measure.

[11] Two single level ANCOVAs were used to test whether the results would be replicated when the nesting of children in schools was not considered as there were so few observations per school. Both analyses produced the same results as the multilevel model: the effect of trial group was not significant either for the Quantitative Reasoning or for the KS1 Maths outcomes; the interaction between eligibility for FSM and trial group was not significant when the outcome measure was the Quantitative Reasoning Test, but a significant negative interaction between trial group and eligibility for FSM was observed when the outcome measure was KS1 Maths.

Subgroup analysis by gender

In order to contextualise the analysis by gender, a comparison between boys and girls in the whole sample at pre-test was first carried out in order to see whether there is a gender effect on mathematical skills as measured by the Quantitative Reasoning Test. We had pre-test scores for 5,351 pupils (49.2% girls) at the end of Year 1. The means at pre-test were very similar for the two groups (9.59 for girls and 9.20 for boys; Cohen's d=0.08); this difference was statistically significant because the sample size is so large (t=2.98; *p*=0.003) but it is a very small difference to be educationally meaningful. The likelihood of boys and girls to be nominated for the trial differed by a small percentage: 10.1% of the boys and 9.1% of the girls were nominated for the trial. These figures suggest that the difference between the scores of boys and girls in the whole sample was not meaningful and that there was no bias in nominating pupils of either gender to participate in the trial.

Table 11. Subgroup analyses by gender

Raw means for girls							
	Intervention group		Control group				
Outcome	N (missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning	103 (6)	9.46	119 (10)	9.07	222	0.08 (-0.18 to 0.35)	0.86
KS1 Maths	100 (9)	3.40	126 (3)	3.27	226	0.13 (-0.07 to 0.46)	0.26
Raw means for Boys							
	Intervention group		Control group				
Outcome	N (missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning	150 (9)	9.95	119 (16)	9.01	269	0.22 (-0.03 to 0.46)	0.12
KS1 Maths	151 (8)	3.30	126 (9)	3.33	277	- 0.03 (-0.36 to 0.39)	0.74

The sub-group analysis by gender was carried out both with the Quantitative Reasoning Test and the KS1 Maths outcomes. The details of this analysis are presented in Appendix 7. The means displayed in Table 11 show that the differences between boys and girls are small, not significant statistically, and inconsistent in direction. In the intervention group, boys performed better than girls in the Quantitative Reasoning Test whereas girls performed better than boys in KS1 Maths; in the control group, the reverse was observed. These fluctuations indicate that there is no evidence for differences between boys and girls in the measures used in this trial.

A multilevel model was first implemented with gender as a main factor and an interaction factor between gender and trial group, intervention vs control. Neither the main effect nor the interaction factor produced a statistically significant result. The effect sizes in the comparison between the intervention

and the control group were small both for girls and for boys and could have happened by chance. Thus taking gender into account did not change the previous results in any way.

Missing data analyses

Although the percentage of missing data can be considered small (7.7%) and the causes for most of the unavailability of individual children for post-test fit with the definition by Dziura et al. (2013) of completely at random missing data (the most common cause was relocation), patterns of missingness were investigated by employing different analyses. As indicated earlier on, three control schools (4.5%) dropped out (leading to missing data for 12 pupils) as compared to no intervention schools. The evaluation team analysed whether one could find similarities between the control schools that withdrew after randomisation, as the existence of a pattern could pose threats to the validity of the trial.

Two schools, A and B, had received the Ofsted evaluation "good" in their last inspections (2015 and 2017, respectively). They had quite different percentages of pupils eligible for FSM: School A had 3.3% and 2.9% of pupils eligible for FSM in 2015/16 and 2016/17 (these are the years during which the trial took place), respectively, whereas School B had 24.8% and 29.2% in these school years. School A had not experienced a change in total number of pupils whereas School B had opened in 2015/16 and had an increase of more than 60% in pupil numbers in 2016/17. The similarity between the schools that led to their withdrawal is that, in School A, the TA who had been nominated to implement the intervention went on long term absence due to illness and, in School B, the Link Teacher who had been nominated went on long term absence due to illness. The head teacher in School B had not been directly involved with the project at the time of the briefing meeting; the TA did not attend the sessions that were offered to TAs in control schools in order to implement the alternative intervention. Thus in both schools there were staff difficulties that prevented the schools from continuing in the project.

School C received an Ofsted rating of "Inadequate" in October 2016, when the TAs were expected to start their training for the implementation of the alternative intervention. It is a much larger school than A and B (almost two times the number of pupils in School A and more than two times the number of pupils in School B) and is similar to School B in percentage of pupils eligible for FSM (28.4% in 2015/16 and 28.7% in 2016/17). Although the evaluation team did not collect information on staff change because the new head teacher was not willing to be interviewed, it is likely that the new head teacher implemented changes in staff as a consequence of the Ofsted rating received by the school. The evaluation and intervention teams agreed that, under these circumstances, it would be inappropriate to pursue further the attempts to schedule the post-test, as the head teacher indicated that the school did not have a commitment to the project after the change in headship.

In order to investigate whether attrition due to relocation was more likely in intervention or in control schools, we calculated the percentage of attrition at the individual level for intervention and control schools that remained in the trial. This seemed relevant as some schools were in geographical areas of high disadvantage and this could be associated with higher mobility if the family was reassigned to different social housing areas. After excluding the three control schools that did not remain in the trial, it was found that 4.3% of pupils in control schools and 5.6% of pupils in intervention schools were not post-tested due to relocation or withdrawal from the trial. The very similar rate of relocation between the two groups suggests that there is no pattern in the missing data at the pupil level related to trial group.

The second analysis comparing pupils who were available for post-test with those who were not was a simple comparison between the pupils' pre-test scores. The mean at pre-test for the pupils who were not available for post-test (n= 41) was 4.76 and the mean for those available for post-test (n=491) was 4.87; Cohen's d effect size was 0.04 (calculated using SD for the whole group=2.52); this difference was not significant statistically (t=0.9; p=0.72), which suggests that there is no reason to suspect a bias in withdrawal from the trial due to pre-test results.

The third analysis investigated whether eligibility for FSM (variable everFSM in the NPD file) was related to the likelihood of the pupil not being available for post-test. A Chi-square test, with eligibility for FSM and drop out/not drop out, entered as categorical variables, indicated that the eligibility for FSM was not associated with the pupil availability at post-test (Chi-square=2.83; df=1; *p*=0.09).

Finally, a logistic regression analysis was carried out to investigate whether the training group attended by the TAs predicted whether the pupils would be lost to post-test. Trainers were entered as categorical variables in this analysis. The association between training group and pupil unavailability for post-test was also not significant (all *p*s higher than 0.12).

If any of these associations had turned out to be significant, it would be possible to search for more complex patterns in the missing data by entering more than one variable in the logistic regression. As there was no association between any of the factors that might have explained the loss of participants and missing data, it seemed appropriate to conclude that the data were missing completely at random and that this low level of missing data should not cause concern with respect to the validity of the trial. Because attrition at school level was only observed in control schools, this loss of schools cannot be attributed to difficulties in implementing the intervention.

Treatment effects in the presence of non-compliers

The intervention team listed under the criteria for fidelity of implementation that all the lessons should be delivered to the children. The inclusion of children who participated in only a limited number of sessions could have affected the results of the impact analysis. Therefore it was relevant for understanding the impact of the intervention to implement an analysis of non-compliance with participation in all the intervention sessions.

TAs had kept record of the pupils' attendance to sessions and these records were provided to the evaluation team; the information on the number of sessions attended by the children was missing for 46 children (17.2%). For the purposes of this analysis, compliance with treatment was described as met if the children participated in 30 sessions or more (three children were offered 1, 2 or 3 extra sessions); if they participated in fewer sessions, they were excluded from the analysis. Of the 268 children in the intervention group, 125 (46.5% of all children in the intervention group) for whom we have information and who participated in all the session are included in the analyses.

The statistical analysis plan stated that further cut-off points for attendance would be investigated in order to extract the best information from the data. No children attended less than a quarter of the sessions; three attended less than half of the sessions; eight attended between half and two thirds of the sessions; 14 attended more than half but less than 3/4 of the sessions. Thus setting a criterion of attendance at 2/3 of the sessions excluded fewer children from the analysis than setting it at 3/4, which was desirable because of the loss of power with loss of participants for the analysis. Of the 268 children in the intervention group, 211 (78.7%) for whom we have information and who participated in at least 2/3 of the sessions are included in the second analysis that investigates the impact of the intervention excluding non-compliers.

Before implementing the multilevel models for the analyses of the impact of the intervention with these different fidelity criteria, that is, the first one in which pupils were selected if they attended all sessions required and the second one if they attended 2/3 of the sessions - the pre-test scores for the children in the intervention group who remained in the analyses and for the children excluded from the analysis were compared in order to investigate whether there was a bias in participation due to initial differences between the groups. This comparison was necessary because it was possible that children performing at a higher level (or lower level) at pre-test enjoyed the intervention more and thus attended more sessions.

The analyses with children who attended all the sessions required includes 363 children (control= 238; intervention = 125) in 111 schools (control = 63; intervention = 48); thus children from 20 intervention schools were excluded from the analysis either because of missing data or because the children did not complete all the sessions. The pre-test means were 4.71 for the children in the intervention group who completed the required number of sessions and 4.98 for those who did not. This small difference in favour of the children who did not complete all the sessions was not statistically significant (t=0.87; df=229; *p*=0.39).

The analyses with children who attended at least 2/3 of the sessions required includes 458 children (control= 238; intervention = 220) in 121 schools (control = 63; intervention = 58); thus children from ten intervention schools were excluded from the analysis either because of missing data or because they did not complete all the sessions. A similar comparison was carried out when the criterion of attending 2/3 of the sessions was implemented. The pre-test mean for the children in the intervention group who attended at least 2/3 of the sessions was 4.70 and for the children who did not attend 2/3 of the sessions was 5.50. This difference in favour of the non-compliers was not statistically significant (t=1.33; df=229; *p*=0.18) and therefore could have happened by chance. The pre-test comparisons suggest that there was no association between completing all the sessions (or 2/3 of the sessions) and the children's previous mathematical attainment as measured by the pre-test. Thus they increase the confidence that one can have in the outcomes of the analyses of impact when non-compliance is taken into account.

Table 12 presents the means for the intervention group when the children who did not meet the criteria of attendance were excluded from the analysis. The impact of the intervention when non-compliers were excluded was investigated using multilevel models based on the same equations used previously, which took into account the nesting of children in schools and used the pre-test as a covariate. The details of these analyses are presented in Appendix 8.

The analyses with the primary outcome, scores in the Quantitative Reasoning Test, showed a larger effect size of the intervention, 0.24, for pupils who attended all the sessions than the previously observed effect size, when the intention to treat design was used. This effect is at the borderline of the level of significance established in the statistical analysis plan. In contrast, when the selection criterion was attendance to 2/3 of the intervention, the effect size was similar to that observed in the previous analyses (0.17) not statistically significant.

In the analyses with the secondary outcome, KS1 Maths, the effect size remained the same as that observed with the intention to treat design (0.04 SD) and was not statistically significant (see details in Appendix 8).

Table 12: Means by trial group and effect sizes when only pupils in the intervention group who attended all the sessions or 2/3 of the sessions are included in the analyses

Attended all the sessions	Raw means					Effect size	
	Intervention group		Control group				
Outcome	N (excluded/ missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning	125 (95/48)	10.08 (9.27 – 10.93)	238 (26)	9.04 (8.48 – 9.60)	363	0.24 (0.01 to 0.45)	0.051
KS1 Maths	121 (82/48)	3.31 (3.22 – 3.38)	252 (12)	3.30 (3.19 – 3.43)	373	0.01 (-0.20 to 0.23)	0.89
Attended 2/3 of the sessions	**Raw means**					**Effect size**	
	Intervention group		Control group				
Outcome	N (excluded/ missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning	202 (18/48)	9.79 (9.60 – 10.37)	238 (26)	9.04 (8.48 – 9.13)	447	0.17 (-0.02 to 0.35)	0.08
KS1 Maths	195 (8/48)	3.32	252 (12)	3.30	447	0.03 (-0.16 to 0.22)	0.89

These analyses suggest that pupil attendance at all the sessions is associated with better outcomes. The percentage of children who did not meet the criterion of full attendance (43.2% of the children for whom we had information) suggests that intervention demands in terms of attendance to all sessions might be too high. Pupils might not attend all sessions either due to TAs' unavailability (for example, sick leave) or due to the fact that children may miss sessions because of illness or because of other factors unrelated to the intervention. However, it is possible that, if the intervention were being implemented outside the trial, TAs could have caught up with the required number of sessions for all pupils.

Additional exploratory analyses

In order to understand better the impact of this intervention, two sets of *post-hoc* additional and exploratory analyses were carried out, as outlined in the methods section. The first analysis aimed to investigate whether there was growth in the intervention pupils' scores when the brief scale of the Quantitative Reasoning Test was used, which includes only the items common to pre- and post-test. This brief scale was used as the measure at pre-test and as the outcome measure in order to detect growth in each group and to see whether this growth differentiated pupils in the intervention from those in the control group. The failure to find a statistically significant impact of the intervention could be due

to the fact that the children did not improve. Alternatively, failure to find significant results could be due to the fact that both the intervention and the control groups improved, and did so to the same extent. The same multilevel model employed in the primary analyses was therefore used to investigate whether there was detectable growth in the pupils' mathematical skills, but using as pre-test and post-test scores the raw scores obtained by adding only the items common to both testing occasions. Table 13 presents the outcomes of this analysis.

Table 13. Raw means at post-test for the intervention and control group in the brief scale of the Quantitative Reasoning Test

	Raw means					Effect size	
	Intervention group		Control group				
Outcome	N (missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	p-value
Quantitative Reasoning Brief Scale	253 (15)	6.04 (5.69 – 6.39)	238 (26)	5.52 (5.17 – 5.88)	491	0.20 (-0.05 to 0.45)	0.10

The analysis with the brief scale of the Quantitative Reasoning Test showed that both groups improved from pre- to post-test. The growth in the intervention group from pre- to post-test was equivalent to 1.15 SD and the growth in the control group was equivalent to 1.01 SD. Although these results show a solid improvement from pre- to post-test, the difference in growth between the groups is rather small. The multilevel model, which took into account the nesting of children in schools and the pre-test performance, did not show a significant effect of the intervention on this measure, even though it is quite clear that the measure is sensitive to growth in this group of participants. Appendix 9 contains the details of the analysis.

The second, additional, *post-hoc* analysis was a sub-group analysis in which the participants were split at the median at pre-test. The teachers were asked to nominate pupils who were struggling with the mathematics curriculum, but there was large variation between the pupils nominated for participation in the project at pre-test. Therefore, in order to understand better the impact of the intervention, it was considered worthwhile to investigate the possibility of differential impact of the intervention as a function of the pupils' attainment at pre-test. For these analyses, the pupils were divided into two groups, those with scores up to the median at pre-test (up to 4) and those with scores above the median at pre-test (5 or above). Separate analyses were then carried out for these two sub-groups, using the models outlined in the statistical analysis plan. The primary outcome measure, the full scale raw scores in the Quantitative Reasoning Test, and the secondary outcome measure, KS1 Maths results, were used in the analyses.

The number of participants in these subgroup analyses is smaller than the number of participants when the whole group is included, and consequently the power to detect a significant effect decreases. However, if the intervention has a differential impact on lower or higher attaining pupils within this group of pupils struggling with maths, different effect sizes would be observed. In these sub-group analyses, there were 255 children in 108 schools in the group that performed up to the median in the pre-test and 236 children in 104 schools in the group that performed above the median. The means, effect sizes and significance levels for these analyses are presented in Table 14.

The analysis for pupils who performed up to the median at pre-test showed a large effect size which was statistically significant and could not be attributed to chance. This large effect size is equivalent to

four months of extra progress in comparison to the control group; In contrast, the impact of the intervention was smaller and not significant statistically when the secondary outcome measure, KS1 Maths, was used in the analysis. [12]

The intervention did not show a statistically significant impact on either the primary or on the secondary outcome measure for pupils who performed above the median at pre-test. Appendix 10 contains the details of these analyses.

Table 14: Raw means at post-test for the intervention and control groups in the Quantitative Reasoning Test with subgroups defined by a split at the median in the pre-test

Pupils who performed up to the median at pre-test							
	Raw means					Effect size	
	Intervention group		Control group				
Outcome	N (missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	*p*-value
Quantitative Reasoning	132 (7)	8.89 (8.18 – 9.79)	123 (15)	7.63 (6.91 – 8.35)	255	0.29 (0.04 to 0.53)	0.04
KS1 outcome	130 (9)	3.20 (3.09 – 3.33)	134 (4)	3.08 (2.98 – 3.18)	264	0.19 (-0.05 to 0.44)	0.36
Pupils who performed above the median at pre-test							
	Raw means					Effect size	
	Intervention group		Control group				
Outcome	N (missing)	Unadjusted Mean	N (missing)	Unadjusted Mean	N in model	Hedges g (95% CI)	*p*-value
Quantitative Reasoning	121 (8)	10.58 (9.80 – 11.36)	115 (11)	10.54 (9.75 – 11.33)	236	0.01 (-0.25 to 0.26)	0.83
KS1 outcome	121 (8)	3.48 (3.37 – 3.59)	118 (8)	3.55 (3.44 – 3.66)	239	-0.11 (-0.37 to 0.14)	0.45

[12] Single level ANCOVAs were used to test whether they replicated the results of the multilevel analyses, one for the group of pupils who performed up to the median in the pre-test and the other for the group who performed above the median at pre-test. Both analyses replicated the results of the multilevel models: the intervention had a moderate and statistically significant impact for the group that performed up to the median at pre-test but the impact was low and not significant statistically for the group that performed above the median at pre-test.

Summary of the results

In summary, the 1stClass@Number intervention had an impact of two months additional progress in comparison to the control group on pupils' mathematical skills, as measured by the Quantitative Reasoning Test. As the difference was not statistically significant, it could have happened by chance. There is no evidence for impact on KS1 Maths outcomes. This result must be interpreted with caution because it could be due to a lack of sensitivity of KS1 Maths to growth in skills in this subgroup of children struggling with maths.

When the brief scale of the Quantitative Reasoning Test (based on the subset of the questions which are the same at pre- and post-test) was used to investigate whether the intervention group had shown growth from pre- to post-test, it was found that there was a substantial growth: the effect size was 1.15 SD. However, this single group analysis does not mean that the intervention had a significant impact. When the intervention group's growth is measured against the growth shown by the control group, which improved by 1.01 SD, it was shown that the difference in growth between the two groups was not statistically significant. Thus the lack of a significant impact of the intervention on the Quantitative Reasoning Test cannot be explained by a lack of sensitivity in the measure because the Quantitative Reasoning Test detected growth in both the intervention and the control group pupils. There just was not a difference in the two groups' progress during Year 2.

The analyses of subgroups planned before the trial started showed that there was no evidence of an impact of the intervention on pupils' performance in the primary nor in the secondary outcomes, irrespective of eligibility for FSM. The multilevel model analysis showed a statistically significant and negative interaction between the intervention and eligibility for FSM when the outcome measure was KS1 Maths, which indicates that the intervention does not work as well for pupils eligible for FSM as it does for those who are not eligible. However, when the subgroup analysis was implemented, the difference between the pupils eligible for FSM in the control and in the intervention group was not statistically significant. This latter result must be considered with caution, in view of the smaller sample size when only pupils eligible for FSM are included in the analysis.

An indication of a positive impact of the intervention was noted in one analysis which was part of the original statistical analysis plan: this positive result was observed in a comparison of pupils who had attended all the intervention sessions with those in the control group. This comparison produced a small effect size, which was at the borderline of the conventional confidence interval (p=0.051), and strictly speaking is not a statistically significant result. This analysis, which takes compliance into account, is less stringent than the intention to treat design but qualifies the results of the more rigorous evaluation.

An additional *post-hoc* analysis, which divided the participants in two sub-groups split at the median at pre-test in the Quantitative Reasoning Test, also provided some evidence of a positive impact of the intervention on the pupils' mathematical skills. This impact was restricted to the group that performed up to the median at pre-test; the effect size was larger than that observed when the pupils were not split in two subgroups, and could not be attributed to chance because it was statistically significant. It corresponds to seven extra months of progress in comparison to an equivalent subgroup in the control group. The result suggests that 1stClass@Number might be an intervention most appropriate for pupils with greater difficulty than those performing above the median in the pre-test, which includes about half of the pupils in this trial. However, the post-hoc nature of the analysis suggests this result should be interpreted with caution.

Cost

In line with the EEF guidelines, the cost of this programme is evaluated by considering a three-year period, which is the average time that teachers in England spend in a school before moving to a different school, retiring, or leaving teaching (Allen, Burgess, & Mayo, 2010). In this evaluation, the costs of the training and of the follow-up visit were covered by the EEF. Thus the cost of training is not

part of what schools had to spend to participate in the trial, but it would be part of the cost of the intervention, if the schools were to pay for it in the future. Thus the cost of training is included in the cost estimates. This is a one-off cost incurred in the first of the three years in the calculation. The cost of materials not provided by Edge Hill University and the staff time were covered by schools in the trial and are part of the cost of the intervention when it is implemented without the EEF subsidy. The elements considered in the estimation of cost include the number of pupils who could receive the intervention over the three-year period, the training and the materials required outside the training, and the cost of assessments (see Table 15).

Number of pupils

1stClass@Number is designed to be delivered to up to four children per group, three sessions per week for a total of 10 weeks. Thus there is potentially time for a trained TA to deliver the intervention to two groups of children in one school year. Of the 63 intervention schools that returned information (95% return rate) about working with a second group, 64% reported that they had done so and 35% reported that the TA had not started with a second group; in the latter group, 12 (19% overall) reported that they were likely to continue to deliver 1stClass@Number before the end of the academic year. Thus approximately 83% of schools were delivering or planned to deliver the intervention to more than one group. Of these schools, 75% reported that they were likely to continue to deliver the intervention in subsequent years and 3% reported that they were unlikely to continue; the remaining 22% were undecided.

Based on this information, it is estimated a TA would deliver the intervention to eight children per year. Over three years, it would be reasonable to expect a TA to be able to deliver the intervention to 24 children.

Materials and training

The Every Child Counts team (ECC) provides four services to schools that use 1stClass@Number. The base level service is the training, teaching materials, and printed resources for the TA and link teacher. ECC also offer three optional services: a visit to the school, continued access to the ECC website, and a Continued Professional Development (CPD) module for the TA. During the evaluation, Link Teachers reported the cost of purchasing or printing additional resources that were required to implement the intervention.

The costs of all services provided by ECC, and the additional costs are presented in Table 15. As continued access to the website and the CPD module were not included as part of the evaluation they have been grouped separately. The costs of these services have not been included in the final analysis because they are optional.

Sandwell Test

The 1stClass@Number team strongly recommends that schools use the Sandwell Test to assess pupils before the intervention, to track their progress and to report the outcomes when the intervention is completed. The cost of the Sandwell involves a one-off price for the first time (handbook, ten booklets for pupils, marking tool, and a one-year license: £415.35) and additional booklets for pupils and license renewal in each year (£95.40). These costs will be included in the cost of the intervention because they are not considered optional. In this trial, the schools were not required to purchase the Sandwell because the results of the Quantitative Reasoning Test at pre-test would be used to support the pupil nomination, but some schools purchased it because it was seen as part of the usual delivery of the intervention.

In this report, only the cost of the resources and training (£900), the visit (£250) and the money spent on additional resources and printing (£31 for the first group + £15.50 for each additional group) will be included in the analysis. The training and the school visit could be considered by schools as “one-off”

investments that benefit the school by increasing the TA's skills. As these up-front costs would not apply in subsequent years, the regular cost for subsequent years is £15.50 for each group's materials.

Training

Training for the intervention consists of one full day, followed by five half-days of training, spread out over a 10 to 12-week period. The cost of training one TA and one link teacher in 1stClass@Number is £900 per school. This figure includes the cost of a pack of teaching materials and enough printed resources for four children. Costs such as transportation to the training venue were not included as these would not be incurred by schools, unless the venue were rather distant from the school, but 1stClass@Number draws on local venues for training. Supply cover for the TA and link teacher while they were attending training are not included in the calculation, although these costs might be incurred by schools, if the school cannot re-organise its day plan to avoid requiring a supply teacher.

Additional resources

Schools are expected to provide additional materials needed for the delivery of the intervention, such as counters, tins, real coins, cubes and other materials. The evaluation team asked how much additional money was spent by the school on printing and on additional resources. Most of these resources will be readily available in school, but some items may need to be purchased. The average cost reported by the 49 intervention schools that provided information on printing and additional resources was £31. There was variation as some schools spent much more than others (range: £0-£120; median: £20; mode: £50).

Spending money on additional materials can be considered a necessary, recurring cost that will need to be spent on each group of four children. As some of the resources can be reused, we estimate that the cost per year would be £31 for the first four children and then £15.50 for every extra group. This estimate takes into account the wear and tear on reusable resources and the printing of new materials. In the first year this would cost £46.50 (£31 for the first group plus £15.50 for the second group of children) and the following years would cost £31 per year (£15.50 for each group of four children). Additional printed resources can be photocopied from the original pack (provided at the training session) or downloaded from the ECC secure website to replace any that were used.

Optional resources

Schools can track the progress of the children taking part in the intervention through the website; a small fee of approximately £10 per year is charged for access after the first year. As this is optional, it was not included in the cost analysis but it can be considered an ongoing cost from the second year should a school chose to make use of it.

ECC's support includes an optional visit to the schools for the TA's trainer to observe the TA's delivery of a 1stClass@Number session and to provide a one to one tutorial to help develop the TA's skills in the delivery of the intervention. The cost of the visit is £250; while optional it has been included in our analysis as schools seem to consider this a valuable part of the training. The costs of the training and visit are considered "one off" as they are only incurred in the first year of the intervention.

TAs that complete the 1stClass@Number training are given the opportunity to gain undergraduate university credits by completing an online distance study module at Edge Hill University. The study fees for this module are currently £260 per person. This is an optional service and a "one off" cost that has not been included in the analysis; the fees would be incurred in the second year, either by the school or TA.

Table 15: "One off" and ongoing costs (£) to deliver 1stClass@Number to two groups of pupils each year over a three-year period

1stClass Costs	Year 1	Year 2	Year 3
1. ***Training, teaching materials & printed resources (one off)***	*900*	*0*	*0*
2. ***Additional resources (ongoing) per group***	*46.5 (31+15.5)*	*31 (15.5 x 2)*	*31(15.5 x 2)*
3. ***Optional visit but recommended (one off)***	*250*	*0*	*0*
4. **Optional website access (ongoing)**	0	10	10
5. **Optional CPD module for TA (one off)**	0	260	0
6. **Sandwell Test**	*415.35*	*95.4*	*95.4*
Total (1+2+3+6)	**1612**	**126**	**126**
Cumulative total	**1612**	**1739**	**1865**

Source: Training costs were taken from information provided by ECC and evaluation team questionnaire. Italic indicates costs included in the cost analysis. Cost of Sandwell Test was taken from**:** https://www.gl-assessment.co.uk/sellingline/sent?productNodeId=2241**;** last accessed in January 2018. Figures may not sum to the total due to rounding.

Staff time

The most substantial cost to schools is staff time as the intervention is to a small group of pupils outside the classroom. This is in addition to any cost incurred by arranging cover for the link teacher and TA to attend the training sessions. The time spent by the TA attending training and administering the intervention by the TA and the link teacher time preparing for delivering the intervention in the first year is presented in Table 16.

Training time (TA and link teacher)

A TA will require one whole day and five half-days (26 hours total) to attend training and a link teacher is required to attend two half-days (8 hours total); a half day has been calculated as three hours 45 minutes and a full day as 7.5 hours. TAs should attend all sessions to get the full benefit of the training. The link teacher should only attend for half of the first session and all of the fifth session. There may be administration time unaccounted for.

Preparation time (TA)

Sixty-seven TAs reported the time spent planning and preparing to deliver 1stClass@Number, which was on average one hour each week (range: 30min – 4hrs 30min), ten hours in total.

Table 16: Average staff time required to deliver the intervention to two groups in the first year

	TA		Link Teacher	
	Days	Hours	Days	Hours
Training time	*1 day & 5 half-days*	*26*	*2 half-days*	*8*
Preparation time	*N/a*	*20 (10 per group)*	*N/a*	*N/a*
Delivery time	*N/a*	*48 (24 per group)*	*N/a*	*N/a*
Supervision time	*N/a*	*6 (3 per group)*	*N/a*	*6 (3 per group)*
Total		**100**		**14**

Source: length of training taken from ECC's training events, length of delivery taken from evaluation team questionnaires to TAs and Link Teachers. Figures may not add exactly due to rounding.

Delivery time (TA)

1stClass@Number is designed to be delivered in 30 half-hour sessions (15 hours in total) to each group of up to four children. Of the 63 intervention schools that reported time spent by the TA delivering the intervention, 22% said that their TA spent more than twice that time delivering the sessions and 16% reported no extra time being used. Link Teachers reported that TAs spent an average of nine additional hours delivering 1stClass@Number bringing the total delivery time by the TA to 24 hours.

Supervision time (TA and LT)

The reported time for meetings between the TA and Link teacher was three hours. This is assumed to be a constant for each group of children as it is likely that the TA will need a tailored approach to each group and includes the general progress reports that the TA should be making to the link teacher.

In order to assess how much time is spent per year by TAs who are trained and then deliver the intervention to two groups of pupils in the same year, Table 17 presents the cumulative staff time required to deliver the intervention to two groups.

Table 17: Staff time required to deliver intervention to two groups each year

Staff time resources (hrs)	Year 1	Year 2	Year 3
Training Time (TA)	*26*	*0*	*0*
Preparation Time (TA	*10 (per group)*	*10 (per group)*	*10 (per group)*
Delivery time (TA)	*24 (per group)*	*24 (per group)*	*24 (per group)*
Supervision time (TA)	*3 (per group)*	*3 (per group)*	*3 (per group)*
Total (TA) hours each year	**100**	**74**	**74**
Cumulative total (TA) hours	**100**	**174**	**248**
Training Time (LT)	*8*	*0*	*0*
Supervision time (LT)	*3 (per group)*	*3 (per group)*	*3 (per group)*
Total (LT) hours each year	**14**	**6**	**6**
Cumulative total (LT) hours	**14**	**20**	**26**

Source: length of training taken from ECC's training events, length of delivery taken from evaluation team questionnaires to TAs and Link Teachers. Figures may not add exactly to total due to rounding.

Cost per child

The cost per child is based on the assumption that the intervention is delivered to eight pupils (two groups) each year over a three-year period. The costs are shown cumulatively, per pupil in Table 18. These figures can be used for quick comparison purposes as different interventions will have different start-up and ongoing costs. Training costs are incurred only in the first year, which results in a considerable difference in cost per child if the school uses the intervention only in the first rather than over a three-year period. According to the EEF cost rating scale, this intervention is considered "very low" cost. The cost to schools in the trial was substantially lower as they only paid for a fraction of the training costs.

Table 18: Cumulative cost per pupil and average cost per pupil, per year

Number of years using programme	Year 1	Year 2	Year 3
Cumulative number of children worked with	*8*	*16*	*24*
Cumulative cost per school (£) for materials, training, visit, and extra resources	1,612	1,739	1,865
Average cost per pupil per year (£)	201	109	77

Source: Information provided by ECC and responses to evaluator's questionnaire. Figures may not sum to the total due to rounding.

Implementation and Process evaluation

Training fidelity

Results from the observations

Observers had an observation sheet on which they were to note departures from the training guidelines, which are detailed and include the additional comments and the time that should be spent as the PowerPoint slides are presented. All sessions were delivered as planned; trainers used the notes for the discussion of the slides and covered all the planned topics. The REDS model (review, evaluate, discuss and share) was used consistently and TAs were active participants in the discussions. At the end of the sessions, TAs were asked whether they felt prepared to deliver the subsequent lessons and the response was consistently positive.

TAs' views of how useful the training was

TAs answered a questionnaire about their views of the training which included five questions, each with a five-point scale, where the lowest value (1) corresponded to the most negative view and the highest value (5) corresponded to the most positive view. The range of answers to each of the questions was consistently above 3 or 4 to 5; all means were higher than 4. These results indicated a very positive perception of the training. The items formed an internally consistent scale (Cronbach's alpha =0.822); this allows for a single score to be used to summarise the answers to these aspects of the questionnaire. The mean score in this questionnaire was 23.25 (SD=2.04), which is high as the maximum possible score in the scale was 25.

Although there was little variability in the TAs' views of the training, because the mean was so close to the maximum possible score, we explored the possibility that the TAs' views would predict the pupils' outcomes in the Quantitative Reasoning Test or in the number of sessions that the children attended. More positive TAs might deliver better sessions, engage the children more, and thus impact their attendance and outcome in mathematics learning. We used two hierarchical regressions to analyse this possibility, in which we entered the TAs' views of the training, after controlling for pre-test performance. The correlation between the TAs' views of the training and pupils' scores in the post-test was equal to 0.15, which was not statistically significant. There was no correlation between the TAs' views of the training and the number of sessions attended by the pupils. Therefore this scale did not contribute to explaining pupils' participation in the intervention nor their post-test results.

In summary, our observations showed that the training was delivered quite close to the script in all training sessions observed and the TAs were very positive about the training. Thus from the perspective of the training delivered, the trial can be seen as a good test of the 1stClass@Number intervention.

TAs' and LTs' participation in the training

Attendance to training sessions was recorded by the trainers and provided to the evaluation team. Because the last session took place after the TAs had completed the delivery of the intervention and aimed at providing the opportunity for reflection and review of the programme, for the purposes of this analysis only attendance to the first five sessions is considered. The intervention team endeavoured to offer extra training opportunities for the TAs who had been unable to attend the session in the training group to which they had been assigned by allowing them to attend the session at an alternative date and place. TAs' attendance to sessions varied between three and five; 51 attended all the sessions, 13 missed one session, and three missed two sessions. Of the 67 LTs, 51 attended both sessions and 16 attended one session only.

TAs and LTs from 42 (out of 67) schools attended all the sessions that they were expected to attend during the implementation period. In order to test whether there was a trend for TAs and LTs from the

same school to miss sessions, we looked at the results of TAs that attended all five sessions and the LTs that attended both sessions (all that they were expected to attend). Fisher's exact test for a 2x2 table showed that there was a significant association between the TAs' and the LTs' compliance (p=0.04) and their school. This suggests that there might be a school factor related to the support for the implementation of the programme.

For the purposes of analysing whether attendance at training affected the pupils' outcomes, the 42 schools whose TAs and LTs attended all the required sessions were defined as meeting this fidelity factor; all other schools were defined as not meeting this factor. An ANCOVA was carried out with attendance at training as the independent variable, the post-test scores as the dependent variable and the pre-test scores as the covariate. This analysis showed that the pupils in schools where TAs and LTs attended all the training sessions had a higher score in the Quantitative Reasoning Test (Cohen's d effect size: 0.09 SD) than the pupils in schools where TAs and LTs did not attend all the sessions that they were expected to attend. This difference was not statistically significant, and therefore could have happened by chance. However, it must be noted that the analysis only applies to schools in the intervention group, and therefore it must be interpreted with caution because of the small number of participants in the analysis.

In summary, the analyses of the training fidelity indicated that there was high fidelity in the delivery of training, high participation in the training and high satisfaction with the training. This suggests that, as far as training is concerned, the project offers a fair assessment of the intervention. The measures of training fidelity do not moderate pupils' outcomes, but this could be due to the lack of variability in the measure, because participation and satisfaction were both high.

Implementation fidelity

TA Questionnaires

The rate of response to the TA questionnaire was 98.3%; only one TA did not answer the questionnaire. The questionnaire included three types of item related to fidelity of implementation: (1) items that investigated the material conditions of delivery; (2) items that investigated the TAs' perception of how easy it was to implement the intervention; and (3) items that investigated the TAs' confidence in the use of the materials and their experience of the programme. Each of these groups of item is examined in turn.

1. Material conditions of delivery:

The items that relate to the material conditions of delivery did not form a scale (Cronbach's alpha reliability =0.16), and thus the connection between these items and pupil outcomes has to be analysed by item. Of the six items related to the material conditions of delivery, four had been listed as critical to fidelity by the intervention team: (1) the protection of the TAs' time for preparation and delivery; (2) the availability of designated space for delivery of the intervention; (3) the space available was free of disturbance; and (4) the implementation of regular meetings between the TA and the LT. The other two items (use of displays for the children's work and prominence of these displays) had not been mentioned by the intervention team but were noted by the evaluation team as part of the environment in which the intervention was delivered when the observations of sessions were carried out.

There was little variability in the items related to availability of a designated space and its freedom from disturbance: only two TAs (3%) did not have a designated space and only 16 (24%) indicated that there were occasionally disturbances to the delivery of the sessions.

There was more variability with respect to time for preparation: 38% indicated that they did not always have sufficient time to prepare, 53% felt some of the time they did not have sufficient time to prepare, and only 11% indicated that they consistently had sufficient time to prepare for the lessons. Thus

preparation time for the TAs on their own was a difficulty in this trial. There was variation with respect to time to meet with the LT, but the answers show a moderate level of fidelity in this time factor: 56% of TAs indicated that they met regularly with the LT, 18% indicated that this was less consistent, and the remaining 26% indicated that the meetings were not regular. There was a significant association between the answer to these two questions about time (Fisher's exact test: 12.2; p=0.01). When the two time factors were taken into account together, 17% of the TAs indicated that they neither had time to prepare nor met with the LT regularly, whereas the others showed intermediate levels of satisfaction with the time they had or indicated that they had sufficient time to prepare and met with the LT regularly. Overall, it seems that the majority had some preparation time either on their own or with the LT. This suggests that the trial is not biased for lack of appropriate space or preparation time.

The association between each of the items and the children's outcome was investigated using separate hierarchical regression analyses in which each item was treated as a predictor variable, after controlling for the children's pre-test scores. The outcome variable was the pupils' scores in the Quantitative Reasoning Test. In none of the four regression analyses was the variation in material conditions a significant predictor of the pupils' post-test scores, after controlling for pre-test scores.

2. TAs' perception of fidelity in their delivery of the intervention:

Most of the TAs (83%) reported that they expected to deliver all the topics by the end of the intervention; 13% expected that they would not be able to deliver the last topic (multiplication); 4% (one TA) reported not having delivered the first topic, which means that not all the topics would be delivered. Although there is limited variation in this item, it can be used to evaluated whether the number of topics delivered can be considered a mediator of the impact of the intervention. The fidelity factor was a dummy variable with the value of 1 if all the topics were delivered and 0 if not all the topics were delivered.

With respect to using the materials in each session, 98% reported doing so consistently. Thus there is not sufficient variability to implement further analysis; there is a high level of fidelity in this aspect of implementation as reported by the TAs.

Setting the scene aims to assess the children's knowledge of the topic as is expected to be implemented with pairs of children. Over half of the TAs reported following this guideline (56%); 7% indicated that they worked with individual children and the remaining 37% indicated that they worked with the whole group of four pupils. These figures suggest a moderate level of fidelity in this aspect of the implementation. This variation allows for analysing whether fidelity in the implementation of setting the scene moderates the impact of the intervention; the fidelity indicator based on this item is a dummy variable in which setting the scene in pairs corresponds to fidelity and the other answers indicate lack of fidelity.

About half of the TAs (56%) reported that it was quite difficult to fit all the activities into the lessons; 27% thought that it was easy to do so and the remaining 15% did not find it either difficult or easy. This variation in responses suggests that some TAs may have been more able to implement the intervention, and this might be a mediator of the impact of the intervention.

With respect to how easy it was to follow the lesson plans while delivering the lessons, the perceived fidelity is lower than that suggested by the previous item: 89% of the TAs found this difficult, 3% found it easy, and the remaining 8% did not find it easy or difficult.

As anticipated, the items that related to TAs' perceptions of their own fidelity in delivering the intervention did not form a single scale (Cronbach's alpha reliability =0.38), and should be analysed separately. However, there was a significant association between the TAs' perceptions of how easy it was to fit all activities in the lesson and how easy it was to follow the written guidelines while running the intervention. Kendall's tau correlation coefficient, which is appropriate for scales with a limited number of scores which can be seen as ordered, indicated a significant relation between these two

items (Kendall's tau b=0.145; *p*=0.008; Kendall's tau varies between -1 and 1, as the Spearman and Pearson correlation coefficients). This suggests that the first two items – number of topics delivered and how setting the scene was implemented - should be analysed separately whereas the latter two (how easy it was to fit all the activities and to follow the lesson plan) can be considered together in the analyses that test whether TAs' perceptions of fidelity mediates pupils' outcomes. The answers to the latter two items were added and formed a single score in the subsequent analyses.

The statistical analysis plan indicated that regression analyses would be employed to test whether the fidelity factors measured by the TA questionnaire could be seen as mediators of the impact of the intervention. Three regression analyses were implemented, the first two with the individual items related to perceived fidelity (number of topics delivered and how setting the scene was implemented) and the last one with a variable that combined the TAs' views of how easy it was to fit the activities in the lesson and to follow the lesson plan. All three analyses were hierarchical regressions that controlled for pre-test scores, and had the pupil's raw scores in the Quantitative Reasoning Test as the dependent variable. None of these analyses showed that the TAs' perceptions of fidelity was a significant predictor of the pupils' scores in the Quantitative Reasoning Test after controlling for pre-test scores. Thus none of these fidelity factors can be considered a mediator of the impact of the intervention.

In summary, TAs reported that the material conditions for implementation, in terms of adequate space and time, were satisfactory; most had a designated space with little disturbance and the majority also reported having sufficient time for preparation, either on their own or with the LT. The majority also reported that they expected to have covered all the topics by the end of the intervention and that they consistently used the materials required in the sessions. However, time to fit all the activities in the programmed half hour and ease of using the written guidelines appeared to be a source of difficulty during implementation.

It was investigated whether the outcomes for pupils who received the intervention from TAs who reported low fidelity (by not following the written guidance on how to do the setting the scene or not delivering all the topics, due to difficulty in fitting the activities into the allocated time for the lessons and in following the written lesson plan) differed from the outcomes for pupils who received the intervention from TAs that reported higher fidelity in implementation. Regression analyses showed that TAs' reported fidelity did not contribute to the prediction of pupil outcomes after controlling for pre-test scores.

3. TAs' confidence and experience of the programme

The items that investigated the TAs' perceived understanding of the intervention and their confidence in delivering it formed a reliable scale (Cronbach's alpha =0.72). There were six items in this scale, with scores that varied from 1 to 5, allowing for a maximum positive score of 30. The range of observed scores in the total scale varied from 15 (the midpoint) to 30; the mean was 26.5 (SD=3.0), the mode was 28 and the median 27. Thus no total score in this scale fell in the negative range of values and the majority of the TAs felt that they understood the principles of the intervention and felt confident about implementing it.

The items that measured the TAs' confidence in delivering the sessions and experience with the intervention were combined into a single score, which was used in a regression analysis, in order to test whether these scores predicted the pupils' post-test scores, after controlling for the pupils' pre-test scores. The TAs' scores in this scale did not show a significant correlation with the pupils' post-test scores, and thus did not predict pupils' outcomes after controlling for pre-test scores.

In summary, the different assessments of implementation and process evaluation suggest that the intervention was implemented with a high level of fidelity. Trainers implemented the training sessions very close to the written script, the majority of TAs and LTs attended the training sessions they were required to attend, and found that the sessions prepared them well to deliver the intervention. The space and time available for the intervention were perceived as satisfactory by TAs, they had the materials

they needed for implementing it, and they felt positive about the intervention. The elements assessed in the implementation and process evaluation suggest that this implementation is a good example of how the intervention works in reality in schools. Thus this trial can be considered a fair assessment of the impact of the intervention.

Observations

The original plan was to observe ten lessons, but it was possible to schedule one more observation, so 11 lessons were observed in different randomisation blocks. The means on each of the three scales were: 3.9 for quality of delivery (highest possible score=5); 4.7 for resources (highest possible score=5); 7.4 for adherence to the written script (highest possible score=10). They were thus consistently above the mid-point for each scale.

The relatively high mean for resources confirms the findings of the questionnaires, which indicated high fidelity in the material conditions of delivery; 9 of the 11 ratings were equal to 5, the maximum score.

The mean for quality of delivery is also high and indicates that the TAs were able to maintain the children engaged in the activities and provided relevant feedback, in agreement with the guidelines. There was only one TA who received a rating below the mid-point in the scale (rating=2); although the TA had all the materials and was familiar with the script, the lesson was long and the children became less engaged with the activities as the lesson progressed; feedback was not consistent (for example, one child came up with a wrong calculation repeatedly but no feedback was given).

The mean rating for adherence to the written text for the lessons was high, but it is qualified by large variation. Eight TAs delivered the lessons entirely or almost entirely as described in the written guidelines, whereas three TAs departed considerably from the script (for example, introducing sections from other lessons, taking considerably more time in one activity, or implementing an activity planned for pairs as a whole-group activity).

Looking across the three ratings, only one TA received a rating lower than the mid-point in two scales and no TA received a rating lower than the mid-point in all three scales.

It seems fair to conclude from these observations that they broadly converge with the findings from the questionnaire: appropriate material conditions were in place, TAs had the resources at hand during the lessons, they engaged well with the children, and tended to follow the written script. The concern expressed by TAs in the questionnaire with respect to time to fit all the activities was also noted in the observations: some TAs extended the lesson time beyond the expected half hour in order to fit in the activities and some had difficulty in following the script.

Formative findings

TAs were asked to describe what they thought were the best aspects of the intervention and the most challenging. The most frequently mentioned aspect of the intervention as “the best thing” was the resources provided: of the 64 TAs who answered this open question, 32 mentioned the resources in general, 16 mentioned the lesson plans and ten mentioned the games that are part of the programme. With respect to the outcomes that they had observed, 20 TAs referred to the pupils’ increased confidence in maths and six suggested that the pupils had improved their maths skills. The positive comments that were offered in response to this open question converge with the positive attitudes described in response to the multiple choice questions.

When asked about the most challenging aspects of the intervention, 39 indicated that time was a challenge, either in terms of fitting the lesson plans into the allocated time (mentioned most often) or finding the time for preparation. Only three TAs indicated that they used extra sessions to be able to implement the complete programme; note that this is confirmed by the fact that three pupils in different schools had received extra sessions. This response to the open question also converges with the

answers to the multiple choice questionnaire, in which lower scores appeared in the items about how easy the TAs thought it was to fit all the activities in the time allotted for the intervention implementation and how much time they had to prepare, either on their own or in meetings with the LT. Only one TA mentioned the difficulties posed by pupils missing sessions, two mentioned difficulties in keeping pupils motivated and two others mentioned pupils' disruptive behaviour during the sessions as challenges to the implementation of the intervention.

The comments provided about the intervention were almost entirely positive. One TA described it as "a different but simple way to deliver basic maths topics", a description that seemed to summarise the comments by many TAs about the materials and the lesson plans.

Business as usual: policies and interventions for pupils struggling with maths in the control and intervention groups

LTs in control (n=8) and in intervention (n=10) schools answered a phone interview that aimed to assess what happened in their schools with the nominated children. Appendix 11 presents a detailed analysis of the responses during the interviews. Additionally, TAs and LTs in all control schools were asked to answer a single questionnaire, containing six items, about the interventions offered to the nominated pupils. Six control schools (9%) did not return the questionnaire.

The findings of the interviews can be summarised here in five points.

- Both intervention and control schools had a policy for supporting pupils struggling with maths at the start of Year 2. These policies were similar and involved the use of booster sessions for pupils who had difficulty in accessing a lesson; this took the form of small group sessions offered to all pupils in the same situation. The pupils participating in the 1stClass@Number intervention had received support which is additional to what is offered regularly to pupils struggling with maths in schools.

- The most common materials used in the classroom to support mathematics learning were Numicom and Base 10 Blocks. This was true of the intervention and of the control schools, and thus the materials used in the classroom cannot explain differences (nor the lack of differences) between the groups.

- Intervention schools reported more often than control schools the use of small group sessions for all pupils, including for those participating in the 1stClass@Number intervention that aimed to improve pupils' KS1 results.

- Control schools reported more often than intervention schools that homework was used as a further means to improve pupils' performance.

- LTs in control schools expressed more often than those in intervention schools that they would like to have additional interventions to improve pupils' outcomes, but were not able to implement new interventions due to lack of funds. Because the intervention schools were receiving the 1stClass@Number intervention, it is likely that the LTs in intervention schools viewed the participation in the trial as providing additional funds for new interventions.

Three main findings emerged from the questionnaires:

- In the control schools, 48% (112) of the nominated pupils had received some form of the usual support offered to pupils struggling with maths (small group catch-up sessions; pre-teaching) and 52% (119) had not received extra support. Among the control schools that had offered support to the nominated pupils, two schools were using a new intervention and the remaining had continued to implement interventions used previously. A comparison (using single level ANCOVA, controlling for pre-test results) between the nominated pupils in the control schools

who had received interventions and those who had not received any intervention (N=112) did not show any significant differences in pupils' outcomes.

- When an intervention had been delivered to nominated pupils in control schools, it usually covered the topics that are part of the 1stClass@Number interventions: 30% of the nominated pupils received teaching on all five topics, 10% on four of the five topics (multiplication was not included) and the remaining 8% of the pupils who received extra support were taught about a selection of the topics included in 1stClass@Number, but not all the topics (for example, place value only, addition and subtraction only). Because these topics are part of the national curriculum, it is expected that interventions offered to the nominated pupils would cover the same topics. Because the pupil outcomes between control schools that offered extra support on these topics and those schools which did not offer extra support was not statistically significant, it did not seem justified to compare the 1stClass@Number intervention schools only with the control schools that had not offered extra support to the nominated pupils.

- Among the control schools that had put in place an intervention for the nominated pupils, 15 schools reported using Numicom (with 59 pupils) and 13 reported using other materials, but not Numicom (with 53 pupils). A comparison (single level ANCOVA with pre-test scores as control) between pupils' outcomes in control schools that used Numicom and those that did not use Numicom showed that the difference between the two groups was not significant.

In summary, schools in the two groups interpreted "business as usual" in a very similar manner. In intervention schools, pupils nominated for participation in the 1stClass@Number intervention continued to receive support in the same way as other pupils in the school who had not been nominated for the trial; they received the intervention in addition to this support and to normal mathematics lessons. Control schools did not seem to be using alternative interventions or materials that could explain why their pupils' progress was rather similar to that of pupils in the intervention group.

An analysis of the context in which the trial took place does not suggest that control schools had made extra efforts to improve their pupils' outcomes in mathematics, which could have compromised the result of this trial.

Conclusion

Key conclusions
6. Pupils who received 1stClass@Number made two months' additional progress in maths, on average, compared to pupils in the control group. This result has a high security rating.
7. The primary result was not statistically significant. This means that, in this trial, even if the intervention had not had an impact, the probability that just by chance we would have observed an effect size as large as the one found is greater than 5%.
8. Pupils who received 1stClass@Number did not perform better in the end-of-KS1 maths test, on average, than pupils in the control group. This could be because the headline maths measure used in the trial was more sensitive than the simple five point scale available for the end-of-KS1 maths test, or because it tests specifically those skills taught in 1stClass@Number.
9. Among pupils eligible for free schools meals, those who received the intervention did not make any additional progress in maths compared to pupils in the control group. This result has lower security than the overall result because of the smaller number of pupils.
10. The intervention was implemented as intended by the developer: most TAs and Link Teachers attended most training sessions, and most of the TAs observed during the evaluation followed the written lesson plans closely.

Interpretation

1stClass@Number is a widely used intervention and its implementation has received government support. Although it is delivered outside the classroom to a small number of pupils, it is a low cost intervention, if it is delivered over three years to two groups of pupils per year. Most TAs indicated that they would like to continue to implement the intervention with further groups of children, which would keep the cost low, as estimated in this project.

The analysis of implementation and process evaluation indicates that training and intervention were delivered as intended. Trainers were faithful to the training guidelines, TAs found the training valuable, and there was a good level of attendance to training. TAs' perceptions of the intervention were positive and the material conditions of space and time were in line with the logic model provided by the intervention team. Observations of randomly sampled TAs showed variation in the implementation, as it is to be expected, but the TAs' adherence to the written guidelines was high as assessed by a comparison between the guidelines and the delivery.

The description of business as usual in control schools showed that control and intervention schools have similar policies for pupils struggling with maths. Pupils in intervention schools participated in activities for pupils requiring extra support as well as in the 1stClass@Number intervention. Less than half of pupils in control schools received extra support. It is therefore unlikely that the effects of the intervention were diluted due to the use of alternative interventions in control schools.

Two factors seem to be associated with interesting differences in impact of the intervention in this project. The first is related to pupil compliance. Pupils attended most of the sessions; the distribution of number of sessions attended was a highly negatively skewed distribution. The mean number of sessions attended was 27 sessions out of the expected 30, and the median and the mode were equal to 30. However, only 56.8% of the pupils for whom we had the information attended all 30 sessions; 17.3% attended 26 sessions or less, and 25.9% attended 27 to 29 sessions. This probably represents what really happens in schools when an intervention is delivered three times a week over ten weeks: pupils may be absent for reasons unrelated to the intervention. Although the number of missed sessions by most pupils was low, this could be a factor reducing fidelity. When non-compliers were excluded from the analysis, a larger effect size was observed, but it did not reach statistical significance. If TAs had been explicitly instructed to make up for missed sessions by scheduling extra sessions, it is possible that a different overall result would have been obtained.

The second factor associated with higher results in this project is the level of mathematical skills amongst the pupils nominated for the trial. Under usual conditions, teachers nominate pupils after the first training session; during this initial training session, the criteria for identifying pupils for whom the intervention is appropriate are discussed. In order to increase rigour in the trial and eliminate the possibility of selection bias that might affect the trial if nomination took place after randomisation (Berger & Exner, 1999; Torgerson & Torgerson, 2008), the intervention team held a briefing meeting in order to provide schools with guidance on how to identify pupils for the intervention before randomisation. This means that nomination took place before the first training meeting and teachers may not have been as clear on how to select pupils as they would have been if the usual procedure had been followed. It is quite possible that the fact that Edge Hill University has a different intervention, Numbers Count, for pupils with more severe mathematics difficulties influenced the nomination process and that some schools did not nominate their weakest pupils on the assumption that they would not be the target group for 1stClass@Number. The analysis for the subgroup of pupils who performed up to the nominated group's median at pre-test showed a moderate effect size (0.29), which was statistically significant according to a multilevel model that considered the nesting of pupils in schools and controlled for pre-test performance. This effect size corresponds to four months of extra progress in Quantitative Reasoning in comparison with the control group. This was a *post-hoc* analysis included in the project in view of the wide variation in scores in the pre-test, but it does shed additional light on the interpretation of the findings. It is therefore possible that 1stClass@Number has an impact on pupils, but that it was not appropriate for too many of the pupils nominated for participation in this trial.

Strengths and limitations

The study has several strengths. First, the small loss of participants, which seemed to be entirely at random, as it was due to relocation of participants rather than intervention related factors. The intervention team was able to keep on board even the intervention school that could not implement the intervention to the end. Thus attrition, which is a frequent problem in RCTs, did not result in threats to the validity of this trial. The loss of three control schools (5% of the control schools), which does not appear to be related to the intervention, but is explained by changes in the schools' staff, does not constitute a threat to the validity of the trial either.

A second frequent limitation of RCTs examining complex interventions is lack of fidelity in implementation. The detailed evaluation of implementation carried out in this project suggests that the intervention was implemented with a good level of fidelity. The implementation and process evaluation measures consistently showed positive ratings. An impressive aspect of implementation was the quality of the training offered to TAs and the schools' efforts in making the materials conditions for implementation as good as possible. TAs' attitudes were positive and they also reported children's attitudes to be positive. All the factors considered add up to the conclusion that the intervention was delivered with as high a level of fidelity as can be expected in the implementation of complex interventions.

Another frequent limitation that does not apply to this study is the absence of schools in rural areas. Although these represent a small percentage of the schools in the project, they were included in the sample. This suggests that the results can be generalised to the geographical region where the study was implemented, South and West Yorkshire.

A limitation of this study relates to the long interval between the schools' expression of interest in the project and its actual beginning with the nomination of pupils for participation in the trial. Under usual implementation conditions, when a school decides to join 1stClass@Number, the pupils are nominated by their Year 2 teachers as soon as the school expresses its interest in the programme and the intervention begins within about two months. Because the intervention was delivered in the context of this RCT, the interval between the school signing up for the intervention and the nomination of pupils, which marks the start of the project, was longer. Schools started to sign up for the trial about nine

months before the beginning of the intervention, as a consequence of the need to recruit a large number of schools. A briefing meeting, during which the intervention team explained the criteria for nomination of the pupils, took place in May, but the pupil nomination took place in September. The long interval between the schools' first expression of interest and the start of its implementation through the nomination of pupils might be related to factors that reduced the schools' engagement with the project, such as change in staff who were leading the schools' involvement in the project, the identification of other interventions that the staff would like to use, or a change in the schools' circumstances (such as a negative Ofsted inspection or an increase in the number of children).

A second limitation might be the process of nomination of pupils in this trial. The criteria for selection of pupils explained by the intervention team in May were phrased positively: for example, pupils selected for this intervention typically can count forwards and backwards to ten, know how to read and write some numbers, know some number facts and so on. It is possible that, because the ECC team has a more intensive intervention for pupils with severe difficulty in mathematics, schools expected 1stClass@Number to be appropriate for pupils who do not have great difficulty with mathematics. The positively phrased criteria and the existence of another intervention recommended for pupils with greater difficulty may have resulted in the nomination of a target group that included too many pupils for whom the intervention activities were not sufficiently challenging. This possibility is supported by the post-hoc analyses that compared the impact of the intervention when it was used with pupils who were weaker or stronger at pre-test, within the group of pupils struggling with maths. The intervention had a moderate and statistically significant effect when used with the weaker pupils but no effect when used with the pupils who had performed above the median.

Finally, under usual implementation conditions, schools have some flexibility in deciding whether the intervention is at the appropriate level for the pupils and can make changes to the composition of the group of pupils receiving the intervention, if required. In the context of the trial, pupils nominated for the intervention were expected to remain in the trial until post-test. These three factors – the process of nomination, and the characteristics of the pupils, and the lack of flexibility once the pupils were nominated - could explain why the intervention's impact was restricted to pupils performing up to the median at pre-test.

Future research and publications

Future research about 1stClass@Number can be carried out using this dataset as well as through new studies. It is possible to use this dataset to investigate the impact of school characteristics on the pupils' performance in the Quantitative Reasoning Test as well as on the likelihood of a greater impact of the intervention. An analysis of between-school differences carried out with the whole sample can be used to identify schools that had better results for their pupils at the end of Year 1. Although there were very large school effects on the outcomes of all pupils in the pre-test, the differences between pupils nominated for the trial were not strongly correlated with school membership. It is thus possible to investigate whether the schools with better overall results, which nevertheless have some weak pupils who participated in the trial, were more effective in implementing 1stClass@Number.

Future research should consider asking teachers to nominate for participation in 1stClass@Number pupils who meet certain criteria, as it was done in this trial, but do not meet other criteria, which was not part of the nomination process for this trial. A more precise identification of pupils for whom the intervention is appropriate might clarify whether it is an effective intervention for pupils within a certain range of mathematical skills.

References

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, D. F. and Kolbe, L. J. (1985) 'Avoiding type III errors in health education program evaluations: a case study', *Health Education Quarterly,* 12**,** pp. 315–331.

Berger, V. W. and Exner, D. V. (1999) 'Detecting selection bias in randomized clinical trials', *Controlled Clinical Trials*, 20, pp. 319–327: https://doi.org/10.1016/S0197-2456(99)00014-8

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. and Petticrew, M. (2008) 'Developing and evaluating complex interventions: the new Medical Research Council guidance', *British Medical Journal,* 337**,** pp. 979–983.

Department for Education (2017) 'National Statistics, Early Years Foundation Stage profile results for England, 2017', SFR 60/2017: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652602/SFR60_2017_Text.pdf [accessed 21 April 2017].

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

Dziuraa, J. D., Posta, L. A., Zhao, Q., Fub, Z. and Peduzzib, P. (2013) 'Strategies for dealing with missing data in clinical trials: From design to Analysis', *Yale Journal of Biology and Medicine*, 86**,** pp. 343–358.

Edge Hill University (2016) *1stClass@Number Handbook,* Ormskirk, Lancashire: Edge Hill University.

Graham, J. W. (2009) 'Missing data analysis: Making it work in the real world', *Annual Review of Psychology,* 60**,** pp. 549–576.

Rutterford, C., Copas, A. and Eldridge, S. (2015) 'Methods for sample size determination in cluster randomized trials', *International Journal of Epidemiology***,** 44 (3), pp. 1051–1067.

Kahan, B. C. and Morris, T. P. (2012) 'Improper analysis of trials randomised using stratified blocks or minimisation', *Statistics in Medicine,* 31**,** pp. 328–340.

Maas, C. J. M. and Hox, J. J. (2005) 'Sufficient Sample Sizes for Multilevel Modeling', *Methodology,* 1**,** pp. 86–92.

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 Maths in Primary School', *British Journal of Developmental Psychology*, 25, pp. 147–166.

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., Bryant, P., Evans, D. and Barros, R. (2015) 'Assessing quantitative reasoning in young children', *Maths Teaching and Learning,* 17**,** pp. 178–196.

Torgerson, D. and Torgerson, C. (2008) *Designing and running randomised trials in health, education and the social sciences*, London: Palgrave Macmillan.

Torgerson, C. J., Wiggins, A., Torgerson, D. J., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hodgen, J., Hulme, C. and Tymms, P. (2010) 'Every Child Counts: the independent evaluation Technical report', Department for Education (DFE), Research Report DFE-RR091a.

Appendix 1. Agreement with schools and consent form

Agreement with schools

Randomised Control Trial of 1stClass@Number

Agreement

This agreement is between the School named below and Edge Hill University (EHU) about a randomised control trial of the 1stClass@Number intervention that will take place in the school year 2016-17, funded by the Education Endowment Fund and evaluated by a team from Oxford University.

Name of School	
Address Postcode	
Head Teacher	
Telephone	e-mail

Edge Hill University	Edge Hill University Every Child Counts Woodlands Centre Southport Road Chorley PR7 1QR	W: everychildcounts.edgehill.ac.uk E: ecc@edgehill.ac.uk T: 01257 517 190	*Edge Hill University runs Every Child Counts on a not-for-profit basis.*

The School

- The School understands that it will be randomly allocated in September 2016 to either the 1stClass@Number group or the control group (receiving training for a Key Stage 2 intervention); it commits to full participation in either group.
- The School's head teacher or representative will attend a briefing meeting in May 2016, accompanied by a link teacher who will support the implementation of 1stClass@Number.
- The School will seek permission from the parents of Year 1 pupils in May/June 2016 to share data about their progress with the evaluation team.
- The School will administer a mathematics test provided by the evaluation team to all Year 1 pupils in June 2016.
- In early September 2016, the School will nominate four Year 2 pupils to receive 1stClass@Number support, a teaching assistant who will attend training and deliver the intervention, and a link teacher who will attend a part of the training and support the teaching assistant, in case the School is allocated to the 1stClass@Number group.
- If the School is allocated to the 1stClass@Number group, it will support the teaching assistant and link teacher to attend all training sessions and to manage and deliver the intervention in accordance with guidance from EHU.
- The School will communicate fully and promptly with EHU and the evaluation team, share appropriate data and ensure that questionnaires and surveys are completed and returned.

- The School will facilitate visits to the school by EHU to support the implementation of its intervention and by the evaluation team to observe and interview staff during 2016-17 and to administer a post-test to all Year 2 pupils in June 2017.

- The School understands that it will receive intervention training, resources and support, which normally cost approximately £1,200, free of charge in return for complete participation in the trial as set out in this agreement. If the School receives any element of the training and does not complete the trial in full, then the School will pay an £800 training fee to EHU.

Edge Hill University

- EHU will provide a briefing meeting in May 2016 to inform all schools about the trial procedures.

- EHU will provide an intervention training programme for the School's teaching assistant and link teacher, delivered by an accredited Every Child Counts trainer:

  - if the evaluation team allocates the School to the 1st Class@Number group, the training will begin in September 2016;

  - if the evaluation team allocates the School to the control group, the training will be for either 1st Class@Writing or Success@Arithmetic and will begin in the autumn term 2016;

  - if the evaluation team allocates the School to the control group and the school has no Key Stage 2 classes, EHU will provide training for 1st Class@Number in the autumn term 2017.

- EHU will inform the School of the venue and dates of its training programme when the evaluation team has made the allocation, endeavouring to allocate the School to its first preference wherever possible.

- EHU will provide will provide written guidance and a comprehensive teaching resource pack to support the School's delivery of its intervention.

- EHU will provide on-site support for the School's delivery of its intervention through a support visit to the School by an accredited Every Child Counts trainer.

- EHU will provide e-mail and telephone helplines and website support for all trial schools.

Signed for and on behalf of:

School _____	**EDGE HILL UNIVERSITY**
Signed _____	**Signed** _____
Name *(print)*_____	**Name** **Nick Dowrick**
Position _____	**Position** **ECC Director**
Date _____	**Date** _____

Thank you for offering to join the trial.

Please return 2 signed copies of this agreement to the address above.
One copy will be signed and returned to you if your application is successful.

Appendix 2. Consent forms

Consent form

UNIVERSITY OF OXFORD

DEPARTMENT OF EDUCATION

Professor Terezinha Nunes

15 Norham Gardens

Oxford OX2 6PY

Dear Parent/Carer,

I am writing to let you know about an exciting project that aims to find out the effect of a mathematics programme, 1stClass@Number, that focuses on improving children's number skills. 1stClass@Number promotes core aspects of number in the National Curriculum (number sense, place value and calculation). The programme was developed by Edge Hill University, where other successful programmes to promote children's mathematics learning were developed. The project is funded by the Education Endowment Foundation (EEF); a team from Oxford University will be evaluating the effectiveness of the programme.

Your child's school has kindly agreed to be part of the study. In order to evaluate the programme, all children will complete a short assessment carried out by their teacher at the end of Year 1. Subsequently, some children will participate in the 1stClass@Number project. They will receive support, in small groups of four children, from a trained TA for 3 weekly half hour sessions over 10 weeks. These sessions are in addition to usual, daily classes of mathematics. At the end of the project, all children will participate again in an assessment, in order to evaluate the programme. This assessment does not influence your child's placement in school. It is necessary only for the research. All schools in the project will have the opportunity to use materials from one of Edge Hill's programmes; some schools will receive 1stClass@Number whereas other will have the opportunity to use a different mathematics or a writing programme. The assignment to each programme will be done randomly. There are no expenses to be incurred by parents from participation.

Pupil data and test responses will be collected and accessed by Oxford University. No information collected by the researchers about individual children will be made available to anyone outside the research team and your child's teacher, who will carry out the assessment. The data will be kept confidential, in accordance with the Data Protection Act. Only average results of the programme evaluation will be published. We will not use your child's name in any report arising from the research.

We are seeking your permission to use your child's data and to obtain your child's Unique Pupil Number (UPN) to complement the assessment of the 1stClass@Number programme. The UPN is part of the Department for Education records. It will allow us to link test results to the National Pupil Database and share data with the EEF, EEF's data contractor FFT Education, the Department for Education, and to the UK data archive for research purposes. Once this information is included in the data set, the data will be anonymised and no one will be able to identify individual children. If you agree for your child to be part of the research and for their UPN to be used, then you need not do anything. If you **DO NOT** wish your child to participate in the project and for the research team to have access to your child's UPN, please complete and return the attached form to your child's class teacher by the 27th May 2016.

We expect that your child will enjoy being part of the project. If you have any questions you would like to ask before replying, please do not hesitate to contact the lead of the evaluation team, Professor Terezinha Nunes (**terezinha.nunes@education.ox.ac.uk**). If you have any concerns about ethical procedures at any point during the research, please contact the Head of the Ethics Committee in the Department of Education, Dr Liam Gearon (**liam-gearon@education.ox.ac.uk**). Please keep this letter for your records.

Kind regards,

Prof Terezinha Nunes

University of Oxford

Title of Project: An evaluation of the 1st Class@Number programme

If you agree for your child to be part of the research and their UPN to be used, then you need not do anything.

If you agree for your child to participate but DO NOT agree to releasing your child's Unique Pupil Number (UPN), please tick the box below.

☐ I agree to my child's participation in the research but DO NOT consent to the Unique Pupil Number to be released to the research team.

Child's name: ……Date of birth: ……

Child's class Teacher: ……..

School:……

Parent/carer name (BLOCK CAPITALS) ……

Parent/carer signature: ……

Date ……

Appendix 3. Tools used in the data collection for implementation and process evaluation and an example of observation record

Observation form for 1stClass@Number Observation Form for Session 'Setting the scene: Topic 3'

Notes:

School		**Date/time**	
TA		**Observer**	
Cluster		**Topic and Lesson Number**	

Preparation and resources	Does the TA have a suitable area to deliver the intervention?		Can you see a 1stClass@Number display?	
	Has the TA prepared the area in advance?		Is the TA using the resources?	
	Do the children keep their books/1stClass@Number Record sheets?			
Notes on school				

Instructions: Highlight details for appropriate lesson before session starts.	
Things to watch out for	**Notes**
• How do they count the cubes? • Do they count them all in ones? • Do they count on from the largest number? • Do they touch them as they count? • Do they recognise small amounts without counting? • Do they use their fingers to help them count on? • Do they recognise small amounts without counting? • Can they explain how they counted them? • Do they touch them as they count? • Do they recognise small amounts without counting? • Do they use their fingers to help them? • Do they count back? **NB:** Some children may find this difficult, so try to have some fun even if the answers are wrong by praising their attempts. You can help them succeed by using very small numbers e.g. 3 - 1 = ?	

	What to look for	Tick if yes	Notes	Time spent
Part 1: doing (CHILD FOCUS)	Do all children Participate/Are engaged with the task?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	
	Do they Explain their answers?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	

• Make a note of the children's approaches to addition and subtraction. • Are there any common errors or misconceptions?			Part 2: talking (TA FOCUS)	Does TA reinforce mathematical vocabulary?		Note any vocabulary used:	
				Does TA give feedback?		Count: Notes on Praise: Notes on guidance:	
• Are they familiar with the symbols? • Do they show some understanding of their meaning?			Part 3: making	Do all children Participate/Are they engaged with the task?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	
Connective model		**Tick if used**		**Deviations/adaptations from lesson plan?**			

Physical Objects	e.g. counters, coins cubes			
Language	(see mathematical vocabulary above)			
Symbols	e.g. numbers and operations (1, 2, 3...+, =, - ,etc)			
Pictures	e.g. number lines, numbers squares			
Notes				

Observation form for 1st Class@Number Observation Form for Session 'Setting the scene: Topic 4'

Notes:

School		**Date/time**	
TA		**Observer**	
Cluster		**Topic and Lesson Number**	

Preparation and resources	Does the TA have a suitable area to deliver the intervention?		Can you see a 1stClass@Number display?	
	Has the TA prepared the area in advance?		Is the TA using the resources?	
	Do the children keep their books/1stClass@Number Record sheets?			
Notes on school				

Instructions: Highlight details for appropriate lesson before session starts.	
Things to watch out for	**Notes**
• How do they count the cubes? • Do they count them all in ones? • To find the total, do they count on from the largest number? • Do they touch them as they count? • Do they recognise small amounts without counting? Do they understand that each time they shake the tray the total stays the same? • If you turn the tray around do they understand that 6 + 4 is the same as  4 + 6? • Can the children use a number sentence to record their ideas?	

	What to look for	Tick if yes	Notes	Time spent
Part 1: doing (CHILD FOCUS)	Do all children Participate/Are engaged with the task?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	
	Do they Explain their answers?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	

• Make a note of the children's approaches and known facts. • Are there any common errors or misconceptions?			Part 2: talking (TA FOCUS)	Does TA reinforce mathematical vocabulary?		Note any vocabulary used:	
				Does TA give feedback?		Count: Notes on Praise: Notes on guidance:	
• How do they count their stamps? • Do they count them all in ones? • Do they easily recognise that 10 and 10 equals 20?			Part 3: making	Do all children Participate/Are they engaged with the task?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	
Connective model		**Tick if used**		**Deviations/adaptations from lesson plan?**			

Physical Objects	e.g. counters, coins cubes			
Language	(see mathematical vocabulary above)			
Symbols	e.g. numbers and operations (1, 2, 3...+, =, - ,etc)			
Pictures	e.g. number lines, numbers squares			
Notes				

Observation form for 1st Class@Number Observation Form for Session 'Setting the scene: Topic 5'

Notes:

School		**Date/time**	
TA		**Observer**	
Cluster		**Topic and Lesson Number**	

Preparation and resources	Does the TA have a suitable area to deliver the intervention?		Can you see a 1stClass@Number display?	
	Has the TA prepared the area in advance?		Is the TA using the resources?	
	Do the children keep their books/1stClass@Number Record sheets?			
Notes on school				

Instructions: Highlight details for appropriate lesson before session starts.	
Things to watch out for	**Notes**
• Are the children able to count in multiples of 5 to 20? 50? 100? • Can they count backwards in multiples of 5? • Are they beginning to predict the next multiple of five without counting in ones? • Are they able to recognise multiples of 5? • Are they able to recognise numbers which are not multiples of 5?	

	What to look for	Tick if yes	Notes	Time spent
Part 1: doing (CHILD FOCUS)	Do all children Participate/Are engaged with the task?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	
	Do they Explain their answers?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	

• Can they see the pattern that is emerging? • Can they describe it? • What mathematical vocabulary do they use to share their thinking?			**Part 2: talking (TA FOCUS)**	Does TA reinforce mathematical vocabulary?		Note any vocabulary used:	
				Does TA give feedback?		Count: Notes on Praise: Notes on guidance:	
• Are the children familiar with positioning the numbers on the number track? • How accurate are they? • Do they calculate where the numbers should be or are they "guessing?" • Are they beginning to predict the next number on the number line without counting in ones? • Can they see the pattern that is emerging? • Can they describe it? • Do they seem aware that 2 lots of five equals ten?			**Part 3: making**	Do all children Participate/Are they engaged with the task?		Child 1: Child 2: Notes: Child 3: Child 4: Notes	

• Are they able to use the multiples of 5 to help them divide by 5?							

Connective model		**Tick if used**
Physical Objects	e.g. counters, coins cubes	
Language	(see mathematical vocabulary above)	
Symbols	e.g. numbers and operations (1, 2, 3...+, =, - ,etc)	
Pictures	e.g. number lines, numbers squares	
Notes		

Deviations/adaptations from lesson plan?

Observation form for 1stClass@Number Observation Form for Sessions '1-5: Topic 3'

Notes:

School		**Date/time**	
TA		**Observer**	
Cluster		**Topic and Lesson Number**	

Preparation and resources	Does the TA have a suitable area to deliver the intervention?		Can you see a 1stClass@Number display?	
	Has the TA prepared the area in advance?		Is the TA using the resources?	
	Do the children keep their books/1stClass@Number Record sheets?			
Notes on school				

Instructions: Highlight details for appropriate lesson before session starts.		
Things to watch out for	**Mathematical focus of lesson**	**Mathematical vocabulary (Tick off vocabulary when used)**
Lesson 1,2 & 3: Check that the children are **counting on** accurately. Sometimes children include the number that they start from. So 5 + 3 should be **5…**6, 7, 8, not **5,** 6, 7. **Lesson 3**: Encouraging the children to check if they have found **all** the combinations that make 5. **Lesson 4**: Make sure that the children are making the link between the calculations and counting up or down the number track of houses **Lesson 5**: Can the children recall addition and subtraction facts for 10? Can the children understand the inverse link between subtracting and adding or do they have to subtract each time?	**Lesson 1**: Count things and name them as we count. **Lesson 2**: Count in multiples of ten. Count 10p coins, indicating the quantity in pence. **Lesson 3**: Count forwards and backwards along a Number Track. Count backwards from 10 and 20 without looking. **Lesson 4**: Count forwards and backwards to 30 using a Number Track. Count 1p coins as they are added to a tin. **Lesson 5**: Count on and back in ones on the Hundred Square Plus, firstly within 30 and then from and to different starting and finishing numbers.	**Lesson 1**: Count, add, how many, altogether, count on from, larger number. **Lesson 2**: Add up, count on from, fact, larger **Lesson 3**: Total, count on from, fact, altogether **Lesson 4**: Count back from, count on from, subtract, take away, add, number sentence **Lesson 5**: Altogether, total, pairs, add, subtract, more, take away, number sentence, forwards, backwards, count back, count on

	What to look for	Tick if yes	Notes	Time spent
Counting	Do they Start with counting?		If children are doing it wrongly, please note the feedback from TA:	
	Does TA Reinforce mathematical vocabulary?		Note vocabulary used:	
Main Learning (TA FOCUS)	Does TA give feedback?		Count: Notes on Praise: Notes on guidance:	

					Does TA reinforce mathematical vocabulary?		Note vocabulary used:	
Adapting the lesson (Fidelity)		**Tick if made easier**	**Tick if learning extended**		Do all children Participate/Are engaged with the task?		Child 1: Child 2: Child 3: Child 4: Notes:	
Lesson 1:	**Making it easier...**Keep the numbers very small and show the children several times how to count on. **Extending the learning…**Use larger numbers							
Lesson 2:	**Making it easier…**Keep the numbers small. **Extending the learning…**Count in 10ps, crossing the 100p/£1 barrier. Use larger numbers for counting on. Encourage the children to put the larger number in their head and count on.				Do they Explain their answers to each other?		Child 1: Child 2: Child 3: Child 4: Notes:	
Lesson 3:	**Making it easier…**Give the children other target totals to find, e.g. 3p or 4p. Use coins next to each stamp to support their counting. **Extending the learning…** Give the children other target totals to find, e.g. stamps with a total of 7p: 4p + 3p = 7p, 5p + 2p = 7p. **Making it easier…**Put more numbers on the houses in Counting Street. **Extending the learning…**If the children can manage numbers beyond 10 you can tape 2 Counting Streets together to make a street with 20 houses. Ask the children to record the calculations they do in Counting Street.				How much time does TA focus on each pair?		Pair 1: Pair 2: Notes:	
Lesson 4:	**Making it easier…**Keep the counting within 30. **Extending the learning…**Count in ones within or beyond 100 using the 200 Grid (Topic 1). Use two empty Parcel Vans to find number pairs to 20.							

Using What has been Learnt (CHILD FOCUS)

Lesson 5:			

Connective model		Tick if used
Physical Objects	e.g. counters, coins cubes	
Language	(see mathematical vocabulary above)	
Symbols	e.g. numbers and operations (1, 2, 3...+, =, - ,etc)	
Pictures	e.g. number lines, numbers squares	
Notes		

To Finish	Is there a special delivery to take home?		Notes:	
	Do the children record today's learning?		Notes:	
Notes				

Observation form for 1stClass@Number Observation Form for Sessions '1-5: Topic 4'

Notes:

School		**Date/time**	
TA		**Observer**	
Cluster		**Topic and Lesson Number**	

Preparation and resources	Does the TA have a suitable area to deliver the intervention?		Can you see a 1stClass@Number display?	
	Has the TA prepared the area in advance?		Is the TA using the resources?	
	Do the children keep their books/1stClass@Number Record sheets?			
Notes on school				

Instructions: Highlight details for appropriate lesson before session starts (Topic 4).		
Things to watch out for	**Mathematical focus of lesson**	**Mathematical vocabulary (Tick off vocabulary when used)**
Lesson 1: It can be difficult to change the count when counting in tens and ones. For example, 23 is 10, 20, (change the count) 21, 22, 23. Make this change very explicit.	**Lesson 1**: Explore addition facts within 20.	**Lesson 1**: More, altogether, more than, fewer than, most, fewest
Lesson 2: Changing the count from 10s to 1s when counting on in 1s over the tens boundary e.g. 10, 20, 30 - 31, 32, 33. If children are having difficulty with 10p and 1p coins, then repeat work from Exploring Place Value lessons using lolly sticks to help children to see the individual units of 10, or else support with the Empty Stamp Sheets.	**Lesson 2**: Use tens frames to find the different ways of writing number sentences for pairs of numbers to 10. – Ten Nice Things	**Lesson 2**: Teen number, ty number, count on, how much, altogether, total, number sentence, pairs, pattern
Lesson 3: Changing the count from 10s to 1s when counting on in 1s over the tens boundary e.g. 10, 20, 30 - 31, 32, 33. Ensure that children are able to say and record their number sentences correctly using the – and = symbols. Ensure that children understand the use of zero (an empty set).	**Lesson 3**: Solve problems to help us to understand subtraction as 'take away'.	**Lesson 3**: Teen number, ty number, count on, how much, altogether, total, how many, left, number sentence, subtract, subtract from, take away
Lesson 4: Do the children use number pairs that make 10 to calculate and derive pairs that make 20?	**Lesson 4**: Use number pairs that total 10 to derive pairs that total 20. Calculate change from 20p.	**Lesson 4**: Count on, total, change, take away, subtract, how many left, number pair.
Lesson 5: Many children may find 'missing number' / 'empty box' problems challenging. Explain to them that here the empty box just means we have to find the number that is missing from the number sentence. The amounts on each side of the = sign should have the same total. For example 20 = 15 + 5 Each side of the = sign has the same total of 20.	**Lesson 5**:Use knowledge of addition and subtraction facts to solve number problems with a missing number.	**Lesson 5**: How many more? Empty box, add equals, total, subtract, take away.

	What to look for	Tick if yes	Notes	Time spent
Counting	Do they Start with counting?		If children are doing it wrongly, please note the feedback from TA:	
	Does TA Reinforce mathematical vocabulary?		Note vocabulary used:	
Main Learning (TA FOCUS)	Does TA give feedback?		Count: Notes on Praise: Notes on guidance:	

						Does TA reinforce mathematical vocabulary?		Note vocabulary used:	
Adapting the lesson (Fidelity)		**Tick if made easier**	**Tick if learning extended**			Do all children Participate/ Are engaged with the task?		Child 1: Child 2: Child 3: Child 4: Notes:	
Lesson 1:	**Making it easier...**In the Main Learning, find pairs that total 10. **Extending the learning...**In Using What We Have Learnt, use two Empty Stamp Sheets and find number pairs to 20. Pair the children and use three or four Empty Stamp Sheets to explore number facts beyond 20								
Lesson 2:	**Making it easier...**Count amounts of money with just 10p coins or just 1p coins in the purses. **Extending the learning...**Introduce a different colour of counter and write number sentences to 10, e.g. 5 + 3 + 2 = 10. Use 2 Empty Stamp Sheets and write number sentences to 20.					Do they Explain their answers to each other?		Child 1: Child 2: Child 3: Child 4: Notes:	
Lesson 3:	**Making it easier...** Count amounts of money with just 10p coins or 1p coins into the tin/money box. Only use the easier problems cards. **Extending the learning...** Count other amounts of money into the tin/money box, e.g. start with a 20p or 50p coin. Use 20 letters instead of 10 letters.				**Using What has been Learnt (CHILD FOCUS)**				
Lesson 4:	**Making it easier...**Display a number track and circle the multiples of 2 to support the count. Use pairs that make 10, not 20. **Extending the learning...**If children are secure with 12 + 8 = 20, can they deduce what should be added to 22 to make 30? Encourage them to use the pairs that make 10 or 20 to derive the new facts.					How much time does TA focus on each pair?		Pair 1: Pair 2: Notes:	
Lesson 5	**Making it easier …**The children use 20 cubes to help solve the missing number problems. **Extending the learning …** Use pairs that total 30.								

Connective model		Tick if used
Physical Objects	e.g. counters, coins cubes	
Language	(see mathematical vocabulary above)	
Symbols	e.g. numbers and operations (1, 2, 3...+, =, - ,etc)	
Pictures	e.g. number lines, numbers squares	
Notes		

To Finish	Is there a special delivery to take home?		Notes:	
	Do the children record today's learning?		Notes:	
Notes				

Observation form for 1stClass@Number Observation Form for Sessions '1-5: Topic 5'

Notes:

School		**Date/time**	
TA		**Observer**	
Cluster		**Topic and Lesson Number**	

Preparation and resources	Does the TA have a suitable area to deliver the intervention?		Can you see a 1stClass@Number display?	
	Has the TA prepared the area in advance?		Is the TA using the resources?	
	Do the children keep their books/1stClass@Number Record sheets?			
Notes on school				

Instructions: Highlight details for appropriate lesson before session starts (Topic 4).		
Things to watch out for	**Mathematical focus of lesson**	**Mathematical vocabulary (Tick off vocabulary when used)**
Lesson 1: When counting, children know that 4 equals 2 and another 2. **Lesson 2:** The children may not be familiar with the terms 'row' and 'column'. Do the children understand there is a range of language that can all mean multiply? **Lesson 3**: When counting back, emphasise the teen numbers. Know that if you are finding half of a number you divide it into two equal parts (numbers). A half refers to one of the two equal parts. Halving a number is the same as dividing it by 2. **Lesson 4**: Say teen and ty numbers correctly when counting on in 5s. **Lesson 5**: Changing the count from counting in fives to counting in	**Lesson 1**: Explore repeated addition and its relationship to multiplication. Multiply by 2 by counting on in multiples of 2. **Lesson 2**: Make and read arrays. Multiply and divide by 5. **Lesson 3**: Develop an understanding of a half as equal sharing into two parts (÷ 2). Find a half of small numbers. **Lesson 4**: Explore doubling and its relationship to multiplication (x 2). Explore the inverse relationship between doubling and halving.	**Lesson 1**: 3 **groups of 2** equals 6 / 3 **lots of 2** equal 6 / 2 **x** 3 = 6/2 **multiplied by** 3 equals 6 **Repeated addition** is 2 + 2 + 2 An **array** is… **x** is the **multiplication** symbol 3 **times** 2 equals 6 4 is a **multiple of** 2 **Lesson 2**: 20 is a **multiple** of 5 / 4 **rows of** 5 equals 20 / 4 **groups of** 5 equals 20 / 5 **times** 4 equals 20 / 5 **multiplied by** 4 equals 20 /5 **x** 4 = 20 /20 **divided by** 5 is 4 and can be written as 20 ÷ 5 = 4 Multiplying is a quick way to calculate **repeated addition.** 5 + 5 + 5 + 5 is the same as 5 **times** 4 or 5 **multiplied by** 4 and can be written as 5 x 4. **Lesson 3**: **Half** of 6 is 3 6 ÷ 2 = 3 / 6 **divided/shared into** 2 groups is 3 **A half** is one of two equal parts **Lesson 4**: **Double** 5 is 10 / **Half** of 10 is 5 / 2 **groups of** 5 is 10 / 2 **lots of** 5 is 10 / **Twice** 5 is 10 / 5 + 5 = 10 / 5 x 2 = 10 / 5 **times** 2 is 10

	What to look for	Tick if yes	Notes	Time spent
Counting	Do they Start with counting?		If children are doing it wrongly, please note the feedback from TA:	
	Does TA Reinforce mathematical vocabulary?		Note vocabulary used:	
Main Learning (TA FOCUS)	Does TA give feedback?		Count: Notes on Praise: Notes on guidance:	

twos can be tricky.	**Lesson 5**: Reflect upon 1.Class@Number and evaluate progress. Revisit children's favourite game/activity **SPECIAL DELIVERY**	**Lesson 5**: Use vocabulary from the selected game			Does TA reinforce mathematical vocabulary?		Note vocabulary used:	

Adapting the lesson (Fidelity)		**Tick if made easier**	**Tick if learning extended**
Lesson 1:	**Making it easier...**Use a number line to count in 2s. **Extending the learning...** Use larger sums of money. Buy two or more stamps - find the total (in Using What We Have Learnt).		
Lesson 2:	**Making it easier...**Use towers of linking cubes to support counting in 5s. Make the rows in the arrays the same colour. **Extending the learning...**Explore arrays arranged in rows of 2 or 4.		
Lesson 3:	**Making it easier...** Use even numbers to 10. **Extending the learning...** Use larger numbers including odd numbers.		
Lesson 4:	**Making it easier...**Use made dominoes to work out doubles **Extending the learning...**In the Doubles and Halves game, start by laying out one of the rows of cards face down, with only the number showing.		
Lesson 5:	**Making it easier …**The children count 5p or 2p coins. **Extending the learning …** Count a selection of 10p, 5p and 2p coins. Extend the count beyond 100p.		

Using What has been Learnt (CHILD FOCUS)

Do all children Participate/Are engaged with the task?		Child 1: Child 2: Child 3: Child 4: Notes:	
Do they Explain their answers to each other?		Child 1: Child 2: Child 3: Child 4: Notes:	
How much time does TA focus on each pair?		Pair 1: Pair 2: Notes:	

Connective model		Tick if used
Physical Objects	e.g. counters, coins cubes	
Language	(see mathematical vocabulary above)	
Symbols	e.g. numbers and operations (1, 2, 3...+, =, - ,etc)	
Pictures	e.g. number lines, numbers squares	
Notes		

To Finish	Is there a special delivery to take home?		Notes:	
	Do the children record today's learning?		Notes:	
Notes				

Observations: Scoring guidance used for the scoring of each observation

Aspect	Sub-Aspects	Max Point(s) per sub-aspect	Description of point(s) given if
Quality	1. Following theoretical model by using: Symbols, Language, Concrete materials, Pictures	1	All of the connective model suggested in the lesson plan is delivered
	2. Children's engagement	1	All children are engaged and following instructions
	3. TA Follows up children's answers	1	TA follows up answers with praise, encouragement and further questions
	4. TA preparation	1	TA is prepared to deliver the session (has materials and consults the script without having to read every line)
	5. Time and pace	1	TA delivers the lesson in an appropriate time and pace
Resources	1. Time provided by school	3	0- TA does not have enough time to deliver the lesson 1- TA over runs time slightly 2- TA finishes before end of time 3- TA uses all of the time available
	2. Materials	1	TA uses appropriate materials for the lesson
	3. Display	1	There is a 1CN display where children can see their work displayed
Adherence to the script	1. Lesson follows plan/script	5	0- TA does not follow lesson plan at all 1- TA significantly deviates from plan unnecessarily 2- TA deviates from plan unnecessarily 3- TA adapts lesson incorrectly to children's levels 4- TA does not adapt lesson to children's level 5- TA adapts plan to children's level
	2. Quality of space to deliver 1CN	5	0- 1CN delivered in inappropriate space such as busy hallway or shared room

			1- 1CN regularly delivered in an inappropriate space 2- 1CN occasionally delivered in an inappropriate space 3- 1CN delivered in a reasonable space with expected levels of disruption 4- 1CN delivered in good space with rare levels of disruption 5- 1CN delivered constantly in an ideal space such as empty classroom or quiet intervention room

Examples of scoring of three observations

Aspect	Sub-Aspects	Description of point(s) given if	SCHOOL 1	SCHOOL 2	SCHOOL 3
Quality	6. Following theoretical model by using: Symbols, Language, Concrete materials, Pictures	All of the connective model suggested in the lesson plan is delivered	0	1	1
	7. Children's engagement	All children are engaged and following instructions	1	1	1
	8. TA Follows up children's answers	TA follows up answers with praise, encouragement and further questions	1	1	1
	9. TA preparation	TA is prepared to deliver the session, has materials and consults the script without having to read every line	1	1	1
	10. Time and pace	TA delivers the lesson in an appropriate time and pace	1	1	0
Resources	4. Time provided by school	4- TA does not have enough time to deliver the lesson 5- TA over runs time slightly 6- TA finishes before end of time 7- TA uses all of the time available	3	2	3
	5. Materials	TA uses appropriate materials for the lesson	1	1	1
	6. Display	There is a 1CN display where children can see their work displayed	1	1	1
Adherence to the script	3. Lesson follows plan/script	6- TA does not follow lesson plan at all 7- TA significantly deviates from plan unnecessarily 8- TA deviates from plan unnecessarily 9- TA adapts lesson incorrectly to children's levels 10- TA does not adapt lesson to children's level	5	5	0

		11- TA adapts plan to children's level			
	4. Quality of space to deliver 1CN	6- 1CN delivered in inappropriate space such as busy hallway or shared room 7- 1CN regularly delivered in an inappropriate space 8- 1CN occasionally delivered in an inappropriate space 9- 1CN delivered in a reasonable space with expected levels of disruption 10- 1CN delivered in good space with rare levels of disruption 11- 1CN delivered constantly in an Ideal space such as empty classroom or quiet intervention room	5	3	3

1st Class@Number Observation protocol

		Training Day Observations					
Cluster	Trainer	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Doncaster	Judith Copley	-	-	TN & RB on 7/11/2016	-	-	-
Huddersfield	Susie Nicholson	-	-	-	DSE on 6/12/2016	-	-
Leeds A	Helen Laflin	-	-	-	-	-	-
Leeds B	Helen Laflin	-	RB on 18/10/2016	-	-	-	-
Sheffield	Jamie Heathcote	DSE on 30/09/2016	-	-	-	DSE on 24/01/2017	-
Wakefield	Louise Matthews	-	-	-	-	-	-

Training Day Observations: One or more researchers from Oxford University will observe one training day in each cluster.

TA Intervention School Observations: We will undertake 12 observations of TA's teaching a 1st Class@Number lesson. The first set of observations will take place after training for topic 3 the second will take place after training for topic 5.

Schools in each cluster were divided into two blocks using the median for the percentage of primary children eligible for pupil premium from the database 2016-2017. Schools in the higher block had a higher percentage of pupils eligible for the pupil premium than the median of their cluster schools in the lower block had a lower percentage of pupils eligible for the pupil premium.

Random Selection of school observations: Selection will occur by randomly allocating numbers from 101-200 to each cluster, this will determine whether schools are selected from the higher block or lower block for the 1st and 2nd visit.

The three clusters with the highest random numbers will have a school selected from the **higher block for the 1st visit** and from the **lower block for the 2nd visit**.

The three clusters with the lowest random numbers will have a school selected from the **lower block for the 1st visit** and from the **higher block for the 2nd visit**.

Every school will then be randomly allocated a number between 101 and 200 with the highest number in each block being selected for observation.

In total, 12 observations in 12 different schools will occur. Each cluster (N=6) will have a school visited twice (once for the first observation, once for the

second).
Each block (N=12) will have a school visited once.

After randomisation the schools will be contacted to inform them that an observation will be taking place. The Evaluation team will request the dates and times over the next week/s for the planned 1stClass@number lessons.

Random numbers were simultaneously generated using the MS Excel formula =RANDBETWEEN(101,200)

The observer will contact school prior to visiting to ensure that the 1stClass@Number lesson is going ahead that day and that there will be at least 2 of the nominated children present. They will plan to arrive between 15 and 30 minutes before the lesson is due to start or during the break/lunch period before.

Upon arrival they will liaise with the link teacher/ TA to organise the logistics of the observation.

TA Observations						
Cluster	1st Observation School and block	1st Observation (November 2016)		2nd Observation School and block	2nd Observation (February 2017)	
		Date/ time	Observer		Date/ time	Observer
Doncaster	Anonymised School	23rd Nov @2pm	DSE	Anonymised School		
Huddersfield	Anonymised School	28th Nov @1pm	DSE	Anonymised School	17th Feb @2:30pm	DSE
Leeds A	Anonymised School	24th Nov @1:30 pm	RB	Anonymised School	20th Feb @1:45pm	DSE
Leeds B	Anonymised School	24th Nov @2:15pm	DSE	Anonymised School	18th Feb @1pm	DSE
Sheffield	Anonymised School	29th Nov @2.45pm	DSE	Anonymised School	23rd Feb @9am	DSE
Wakefield	Anonymised School	23rd Nov @8:30am	DSE	Anonymised School	23rd Feb @1:30pm	DSE

Random allocation

Cluster	Rand No.	1st Observation	2nd Observation
Doncaster	121	Lower (1)	Higher (2)
Huddersfield	188	Higher(4)	Lower(3)
Leeds A	137	Lower(5)	Higher(6)
Leeds B	153	Higher(8)	Lower(7)
Sheffield	173	Higher(10)	Lower(9)
Wakefield	102	Lower (11)	Higher(12)

1st **Class@Number phone interview script**

1stClass@Number Link Teacher Phone Interview

Introduction

Hello <Link teacher name>, I hope that you are having a nice day? My name is <researcher name> and I am calling on behalf of Oxford University regarding the phone interview for 1stClass@Number that was arranged via email last week.

EITHER

That is quite alright we are aware that these unexpected things can crop up. Do you have another date/time in mind when I will be able to call you back so that we can complete the phone interview?

OR
Thank you for agreeing to speak to me today, I have some brief questions that I would like to go through with you so that the evaluation team can get a better idea of how 1stClass@Number has been implemented in your school.

Preamble

Before we begin I would like to check whether you have received the information sheet and consent form regarding the phone interview that was included in our email?
YES/NO

To summarise the information sheet members of the evaluation team at Oxford are speaking to a selection of Link Teachers in both control and intervention schools. We are interested in individual experiences and thoughts about 1stClass@Number, both positive and negative. We combine all the data we collect to provide an overall picture of 1stClass@Number and its implementation. Any comments in the report are attributed very generally, for example, as "A (Year 2) teacher commented that…" and nothing that you say will be reported back to your school or to the 1stClass@Number team.

Do you have any questions about the project or how we use your comments?
YES/NO

Ethics

The interview will take about 30 minutes. You do not have to answer any questions that you are not comfortable with and you can stop at any time; no explanation needed. If any question doesn't make sense, please feel free to ask me for an explanation.

Are you able/willing to sign the consent form?
YES/NO

Explain procedure

I will begin in a moment by reading the first question. All your details will be anonymised when the data is transcribed. Do you have any questions before we start?
YES/NO

*****Start of Questions*****

NOTE: Bullet points are questions that can be used to prompt interviewees who struggle to answer the initial question or to elaborate their answer. Do not ask if they have already answered with their initial response.

Besides 1stClass@Number, what activities are in place to support Y2 children's mathematics learning?

• Maths days (outdoor maths), Maths/homework club

• How is the TA used in the Y2 classroom

Besides 1stClass@Number, do Y2 children receive one-to-one or small group supplementary maths instruction?

• Did the school change how they would work with these children from previous years?

• Did this include any specific work with the nominated children in preparation for their KS1 SATs?

If so, what programmes are used with them?

• Other interventions e.g. catch up, Mathematics Mastery, LA or school led interventions?

• Brief description; how often they take place/who takes part, who teaches the interventions, whether they received formal training in the interventions?

• Do you have a system to track children's progress as a potential effect of participating in these interventions?

• Do you have any in school mathematics evaluation procedures?

What resources are used in the classroom during maths lessons?

• Concrete materials e.g. coins, cubes, Numicon, base 10, cubes,

Picture materials e.g. Number tracks/lines, 100 squares, number cards

INTERVENTION: Have you received any feedback from the parents of the children who have taken part in 1stClass@Number?

CONTROL: Have you received any feedback from the parents of the children who were nominated (received specific programme mentioned above)?
• Positive and negative, requests from other parents to take part

INTERVENTION: Do you have mid and long terms plans for the use of 1stClass@Number in your school? What would need to be in place for these to happen? Do you think it is likely that this will go ahead?

CONTROL: Do you have mid and long terms plans for the use of Maths interventions with year 2 children (or specific programme if mentioned above) in your school? What would need to be in place for these to happen? Do you think it is likely that this will go ahead?
• Will your school continue to dedicate the time and resources needed to continue delivering 1stClass@Number/ programme/maths intervention?
• Given the potential for staff turnover would your school consider training another TA in the 1stClass@Number/programme/maths intervention?

Is there anything else that you would like to add for 1st class at number and how it has been implemented in your

Thank you very much for taking the time to speak to me, your help with our evaluation it is greatly appreciated.

Post interview pleasantries

If you have any further questions please do not hesitate to contact me.

Thank you again for your time and assistance with our evaluation.

1st Class@Number information and consent form for phone interview

Dear 1stClass@Number Evaluation Link teacher,

As part of the evaluation of 1stClass@Number it is important to get an idea of what each school classifies as business as usual. We combine all the data we collect to provide an overall picture of 1stClass@Number and its implementation. In order to gain an understanding of what this means to the schools in the evaluation a selection of schools from both the control group and intervention group were randomly chosen to conduct a structured telephone interview with the school's Link Teacher.

The phone interview should take no longer than 30 minutes and can be completed during or outside of school hours between the 3rd and the 7th of April or the 19th and the 26th April depending on when your school has Easter holidays. If you are willing to take part please email **david.sanders-ellis@education.ox.ac.uk** to confirm your participation and indicate 2 dates and times that would be most suitable for yourself.

Times between normal working hours 09:00-17:30 would be preferred however early morning and late evening can also be arranged. If you would prefer to be contacted on your mobile or home telephone rather than through the school's phone number please include this information in your email.

We are interested in individual experiences and thoughts about 1stClass@Number and how maths is taught in schools that did not receive the intervention. Any comments in the report are attributed very generally, for example, as "A (Year 2) teacher commented that…" and nothing that you say will be reported back to your school or to the 1stClass@Number training team.

To help you reflect on your schools position we have included the list of questions below.

INTERVENTION Schools:

- *Besides 1stClass@Number, what activities are in place to support Y2 children's mathematics learning?*
- *Besides 1stClass@Number, do Y2 children receive one-to-one or small group supplementary maths instruction?*
- *If so, what programmes are used with them?*
- *What resources are used in the classroom during maths lessons?*
- *Have you received any feedback from the parents of the children who have taken part in 1stClass@Number?*
- *Do you have mid and long terms plans for the use of 1stClass@Number in your school? What would need to be in place for these to happen? Do you think it is likely that this will go ahead?*

CONTROL Schools:

- *What activities are in place to support Y2 children's mathematics learning?*
- *Do Y2 children receive one-to-one or small group supplementary maths instruction?*

- *If so, what programmes are used with them?*
- *What resources are used in the classroom during maths lessons?*
- *Have you received any feedback from the parents of the children who were nominated (received specific programme mentioned above)?*
- *Do you have mid and long terms plans for the use of Maths interventions with year 2 children (or specific programme if mentioned above) in your school? What would need to be in place for these to happen? Do you think it is likely that this will go ahead?*

The consent form below can be returned digitally (scanned copy) via email, or through the post.

– –

Name of teacher ………………………………………………………………...

School …………………………………………………………………………………….. ....

☐ I confirm that I have read and understood the information sheet for the above evaluation and have had the opportunity to ask questions

☐ I give consent to take part in the telephone interview for 1stClass@Number and understand my rights to refuse to answer any of the questions and/or withdraw my consent at any point before, during or after the interview.

Signature of teacher…..……………………………………………….............
Date…..………………......................

Thank you for agreeing to take part in this research. Please retain a copy for your files and return this consent form to:

David Sanders-Ellis

University of Oxford

Department of Education
15 Norham Gardens

Oxford

OX2 6PY

OR

david.sanders-ellis@education.ox.ac.uk

1st Class@Number link teacher questionnaire (intervention schools)

Nominated link teacher:

If different please write your name: .....

School

School ID

Instructions:

We would be very grateful if you could please fill in this questionnaire by filling in the bubble with a black pen. Example:

	Question	Scale				
		Head Teacher	Deputy Head	Class Teacher	SENCO	Maths Coordinator
1	Which of these describes your role in the school (Please tick all that apply)	☐	☐	☐	☐	☐
2	Which members of the school were involved with the decision to take part in the evaluation of 1stClass @Number (Please tick all that apply including yourself)	☐	☐	☐	☐	☐
3	What members of staff have continued to be involved with 1stClass@Number? E.g. through facilitating or supporting its implementation. (please tick all that apply including yourself)	☐	☐	☐	☐	☐
4	Was the decision to take part in the evaluation related to a previous Ofsted Report?	YES ☐		NO ☐		
5	Was the decision to take part based on the Education Endowment Foundation's call for TA involvement?	YES ☐		NO ☐		
6	Does 1stClass@Number integrate with your school's existing maths policy?	YES ☐		NO ☐		
7	Has your school used 1stClass@Number with children who were not nominated for the project?	YES ☐		NO ☐		
8	How likely is your school to continue to use 1stClass@Number next year?	Not at all ☐	Unlikely ☐	Undecided ☐	Likely ☐	Very Likely ☐

9	1stClass @ Number consists of 30, approximately half-hour long lessons. Approximately how many additional hours were spent by the TA delivering 1stClass@Number	TA _____Hours
11	Approximately how many hours did you spend meeting with the TA?	TA _____Hours
13	How much does your school pay per hour to a TA who could deliver 1stClass@Number	£_____
14	What was the estimated cost of your involvement in the intervention? (Your Estimated Hourly Salary x time spent supporting the implementation of 1stClass@Number)	£_____
15	Approximately how much did it cost the school to employ supply cover for the TA and Link teacher to attend training for 1stClass@Number? (Please enter "0" if cover was not employed)	TA Cover £_____ Link Teacher Cover £ _____
16	Approximately how much was spent by the school on printing and additional resources (cubes, coins etc) that were not provided by1stClass@Number?	£_____

THANK YOU

1st Class@Number TA questionnaire (control schools)

Nominated TA: <name>

If different please write your name: .....

School

School ID

Instructions:

Last year your school nominated 4 children for the 1stClass@Number project. Your school was assigned to the control group and the children did not participate in the 1stClass@Number teaching. We are trying to find out whether they participated in alternative interventions. We would be very grateful if you could please fill in this questionnaire by filling in the bubble with a black pen. Example:

Question	List of nominated children	Yes	No
1. Have these children received any other maths intervention?	<name1>	☐	☐
	<name2>	☐	☐
	<name3>	☐	☐
	<name4>	☐	☐

		TA	Teacher	Maths specialist	Other
2. How was the intervention delivered	<name1>	☐	☐	☐	☐
	<name2>	☐	☐	☐	☐
	<name3>	☐	☐	☐	☐
	<name4>	☐	☐	☐	☐

3. Approximately, how many hours overall did each child received from the intervention		1-3 hours	4-6 hours	7-9hours	10- or more
	<name1>	☐	☐	☐	☐
	<name2>	☐	☐	☐	☐
	<name3>	☐	☐	☐	☐
	<name4>	☐	☐	☐	☐
4. Was this intervention new to the school or used in previous years?	New ☐		Used before ☐		
5. Tick the topics you have covered during your sessions.	Number systems ☐	Place value ☐	Addition and subtraction ☐	Multiplication and division ☐	
6. Tick the materials that you used in your sessions	Hundred squares plus ☐	Manipulatives (e.g. linking cubes, counters) ☐	Games ☐	Numicon ☐	

THANK YOU

1st Class@Number TA questionnaire (intervention schools)

Nominated TA:

If different please write your name: .....

School

School ID

Instructions:

We would be very grateful if you could please fill in this questionnaire by filling in the bubble with a black pen. Example:

Question					
	Number systems	Place value	Addition and subtraction	Multiplication and division	
1. Tick the topics you have covered during your sessions.	⬚	⬚	⬚	⬚	
	Never Clear	Not very clear	Neither	Somewhat clear	Always clear
2. In the Mathematical Overview of your topic books, each lesson has a mathematical focus. How clear to you was the connection between this focus and the activities you delivered?	⬚	⬚	⬚	⬚	⬚
	Hundred Square plus	Manipulatives (e.g. linking cubes, counters, straws		Games	
3. Tick the materials that you used in your sessions.	⬚	⬚		⬚	
	Very difficult	Somewh at difficult	Neither	Somewhat easy	Very easy
4. The lesson plans use concrete materials, symbols,	⬚	⬚	⬚	⬚	⬚

mathematical language and pictures. How easy was it to include these in your sessions?					
	1 to 1	In pairs	4 Children at a time		
5. How did you deliver the "Setting the scene" session(s)?	☐	☐	☐		
	Very difficult	Somewhat difficult	Neither	Somewhat easy	Very easy
6. How easy were the 1stClass@Number lessons to fit into the time you had available for each session?	☐	☐	☐	☐	☐
7. How easy were the 1stClass@Number lesson plans to follow while you were teaching the session?	☐	☐	☐	☐	☐
	Not useful at all	Not very useful	Neither	Somewhat useful	Very useful
8. How useful did you find the guidance for adapting the lessons so that they were suitable for the child with the lowest maths skills in the small group?	☐	☐	☐	☐	☐
9. How useful did you find the guidance for adapting the lessons so that you could challenge the child at the highest level?	☐	☐	☐	☐	☐
	Disagree	Somewhat disagree	Neither	Somewhat agree	Agree
10. The 1stClass@Number training prepared me to deliver the sessions.	☐	☐	☐	☐	☐
11. The training gave me clear information about children's difficulties in maths.	☐	☐	☐	☐	☐

12. The training provided clear guidance on how to help children overcome their difficulties in maths.	☐	☐	☐	☐	☐
13. I found that the REDS (Review, Evaluate, Discuss and Share) sessions during training improved my delivery of the subsequent topics.	☐	☐	☐	☐	☐
14. The feedback I received from my trainer following their visit was very useful.	☐	☐	☐	☐	☐
15. Did you have a designated area to work with the nominated children?	YES ☐		NO ☐		
	Never	Sometimes	Often	Very often	All the time
16. Were your sessions significantly disturbed by other children or school activities?	☐	☐	☐	☐	☐
17. Were you able to display the children's work in a prominent area?	YES ☐		NO ☐		
	Less than 1 hour	1 to 2 hours	2 to 3 hours	3 to 4 hours	4 hours or more
18. How much time did you use for planning and preparation time each week?	☐	☐	☐	☐	☐
	Weekly	Once every 2 weeks	Monthly	Once per term	Never
19. How often did you meet with your Link Teacher or the Class Teacher/s to discuss 1stClass@Number?	☐	☐	☐	☐	☐
	Every session	Every week	Every 2 weeks	Once per month	Never

20. How often did you or the children record their work?	☐	☐	☐	☐	☐
	Disagree	Somewhat disagree	Neither	Somewhat agree	Agree
21. After teaching 1stClass@Number, I feel more confident in teaching maths to small groups.	☐	☐	☐	☐	☐
22. After teaching 1stClass@Number I feel able to apply the skills I developed in my role in the classroom.	☐	☐	☐	☐	☐
23. Taking part in 1stClass@Number has encouraged me to look for further professional development.	☐	☐	☐	☐	☐
24. I would like to continue delivering the programme with future year 2 children.	☐	☐	☐	☐	☐
25. What was the best thing about 1stClass@Number?					
26. What was the most challenging aspect of delivering 1stClass@Number					

Appendix 4. Details of impact analyses with the primary and the secondary outcome measures

Model 1. Multilevel model used to test the impact of 1stClass@ number –

Dependent variable: *Quantitative Reasoning Test*

USING R ANALYSIS (eefAnalytics package)

Table 1 Fixed effects calculated with eefAnalytics R package

Fixed effects: Estimates and confidence intervals for predictors specified in the model.			
	Estimate	**95% LB**	**95% UB**
Intercept	6.36	5.32	7.39
Group allocation	0.68	-0.29	1.65
Pre-test	0.55	0.39	0.71

Table 2 Covariance Matrix using eefAnalytics

Vector of variance decomposition into between cluster variance (Schools) and within cluster variance (Pupils). It also contains the intra-cluster correlation (ICC).			
Schools	Pupils	Total	ICC
4.06	14.72	18.78	0.22

Table 3 Effect size calculated with eefAnalytics R package

Hedges' g effect size for the intervention, the confidence intervals are 95%			
	Estimate	95% LB	95% UB

Within	0.18	-0.08	0.43
Total	0.16	-0.07	0.38

USING MlwiN software

Figure 1 Equation representation of model 1 using MLwiN

$$post\_test_{ij} \sim N(XB, \Omega)$$
$$post\_test_{ij} = \beta_{0ij}cons + 0.551(0.080)pre\_test_{ij} + 0.678(0.492)Intervention\ group_j$$
$$\beta_{0ij} = 6.360(0.523) + u_{0j} + e_{0ij}$$
$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 3.921(1.016) \end{bmatrix}$$
$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 14.695(1.091) \end{bmatrix}$$
$$-2*loglikelihood(IGLS\ Deviance) = 2803.318(491 \text{ of } 532 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 4 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	6.36	0.52	12.16		5.34	7.38
Pre-test	0.55	0.08	6.89		0.39	0.71
Group allocation	0.68	0.49	1.38	0.17	-0.28	1.64

u **School**	3.92	1.02	3.86			
e	14.70	1.09	13.47			
Deviance	2803.32					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

USING R ANALYSIS (eefAnalytics package)

Table 5 Fixed effects calculated with eefAnalytics R package

Fixed effects: Estimates and confidence intervals for predictors specified in the model.			
	Estimate	**95% LB**	**95% UB**
Intercept	2.96	2.81	3.11
Group allocation	0.03	-0.11	0.16
Pre-test	0.07	0.05	0.10

Table 6 Covariance Matrix using eefAnalytics

Vector of variance decomposition into between cluster variance (Schools) and within cluster variance (Pupils). It also contains the intra-cluster correlation (ICC).			
Schools	Pupils	Total	ICC
0.06	0.35	0.41	0.15

Table 7 Effect size calculated with eefAnalytics R package

Hedges' g effect size for the intervention, the confidence intervals are 95%			
	Estimate	95% LB	95% UB
Within	0.05	-0.18	0.28

Total	0.04	-0.17	0.25

USING MlwiN software

Figure 2 Equation representation of model 2 using MLwiN

$$KS1\_Maths_{ij} \sim N(XB, \Omega)$$

$$KS1\_Maths_{ij} = \beta_{0ij}cons + 0.072(0.012)Pre\_test_{ij} + 0.028(0.068)TrialGroup_j$$

$$\beta_{0ij} = 2.959(0.076) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.060(0.020) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.350(0.026) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 965.679(503 \text{ of } 532 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 8 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	2.96	0.08	37.00		2.80	3.12
Pre-test	0.07	0.01	7.00		0.05	0.09
Group allocation	0.03	0.07	0.43	0.67	-0.11	0.17

u **School**	0.06	0.02	3.00			
e	0.35	0.03	13.46			
Deviance	965.68					

Model 3. Multinomial analyses of the impact of the intervention treating KS1 outcomes as ordered categories

KS1 outcomes are released by the NPD as five categories that can be ordered. We analysed the effect using a multinomial analysis with ordered categories in which we merged the two lowest categories into a single category and eliminated from the analysis the highest level, which only included 8 participants (approximately 1% of the sample). The highest category is for pupils performing above the level expected at the end of Year 2, which is unlikely to be observed in a sample of pupils selected at the start of the year because they were struggling with maths. The distribution by trial group is presented here in frequencies in the graph and in percentages as well as frequencies in the cross-tabulation.

The equation and outcomes of this analysis are presented below.

USING MlwiN software

Figure 3 Equation representation of model 3 using MLwiN

$$\text{resp}_{ijk} \sim \text{Ordered Multinomial}(\text{cons}_{jk}, \pi_{ijk})$$
$$\gamma_{2jk} = \pi_{2jk};\ \gamma_{1jk} = \pi_{2jk} + \pi_{1jk};\ \gamma_{0jk} = 1$$
$$\text{logit}(\gamma_{1jk}) = + h_{jk}$$
$$\text{logit}(\gamma_{2jk}) = + h_{jk}$$
$$h_{jk} = 0.577(0.061)\text{Pre\_test.12}_{jk} + 0.496(0.337)\text{TrialGroup.12}_{k}$$
$$\text{cov}(y_{sjk}, y_{rjk}) = \gamma_{sjk}(1 - \gamma_{rjk})/\text{cons}_{jk} \quad s <= r$$

N=495

The standard errors are displayed in parentheses after each parameter estimate.

Table 9 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
Pre-test	0.577	0.061	9.46		0.46	0.70
Group allocation	0.496	0.337	1.47	0.14	-0.16	1.16

Table 10 Pupil level at KS1 Maths (categorical)

Variable	Intervention group		Control group		Effect size		
Pupil level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage	n in model (intervention; control)	OR (95% CI)	Cohen's d approx. (95% CI)
Writing							
GDS	120/1326 (23)	9.1%	135/1300 (51)	10.4%	2626 (1326, 1300)	0.85 (0.5, 1.35)	-0.07 (-0.28, 0.13)
EXS	788/1326 (23)	59.4%	789/1300 (51)	60.7%			
WTS	351/1326 (23)	26.5%	302/1300 (51)	23.2%			
PKG	52/1326 (23)	3.9%	61/1300 (51)	4.7%			
PKE	12/1326 (23)	0.9%	4/1300 (51)	0.3%			
PKF	13/1326 (23)	0.2%	8/1300 (51)	0.6%			
BLW	0/1326 (23)	0.0%	1/1300 (51)	0.1%			

Appendix 5. Details of the impact analysis with the primary and secondary outcome measures including randomisation block

USING MlwiN software

Model 1. Multilevel model used to test the impact of 1st Class@ number on pupils classified by block used for randomisation –

Dependent variable: *Quantitative Reasoning Test*

Figure 4 Equation representation of model 1 using MLwiN

$$\text{Post\_test}_{ij} \sim N(XB,\ \Omega)$$

$$\text{Post\_test}_{ij} = \beta_{0ij}\text{cons} + 0.553(0.080)\text{Pre\_test}_{ij} + 0.984(1.051)\text{TrialGroup}_j + 0.015(0.100)\text{Block}_j + -0.046(0.140)\text{TrialGroup.Block}_j$$

$$\beta_{0ij} = 6.245(0.859) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) : \Omega_u = \begin{bmatrix} 3.920(1.016) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) : \Omega_e = \begin{bmatrix} 14.691(1.091) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 2803.194(491 of 532 cases in use)

The standard errors are displayed in parentheses after each parameter estimate.

Table 11 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	6.25	0.86				
Pre-test	0.55	0.08				
Group allocation	0.98	1.05	0.93	0.35	-1.08	3.04
Block	0.02	0.10	0.20	0.85	-0.18	0.22
Group*block	-0.05	0.14	-0.36	0.72	-0.32	0.22
u **School**	3.92	1.02				
e	14.69	1.09				
Deviance	2803.19					

Model 2. Multilevel model used to test the impact of 1st Class@ number on pupils classified by block used for randomisation –

Dependent variable: KS1 Maths

Figure 5 Equation representation of model 2 using MLwiN

$$KS1\_Maths_{ij} \sim N(XB, \Omega)$$

$$KS1\_Maths_{ij} = \beta_{0ij}cons + 0.070(0.012)Pre\_test_{ij} + -0.217(0.143)TrialGroup_j + -0.023(0.014)Block_j + 0.038(0.019)TrialGroup.Block_j$$

$$\beta_{0ij} = 3.120(0.117) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.056(0.019) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.350(0.026) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 961.679(503 \text{ of } 532 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 12 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	3.12	0.12				
Pre-test	0.07	0.01				
Group allocation	0.22	0.14	1.57	0.12	-0.05	0.49

Block	-0.02	0.01	-2.00	0.05	-0.04	0.00
Group*block	0.04	0.02	2.00	0.05	0.00	0.08
u **School**	0.06	0.02				
e	0.35	0.03				
Deviance	961.68					

Appendix 6. Details of the analyses with subgroups defined by eligibility for FSM

USING MlwiN software

Model 1. Multilevel model used to test the impact of 1st Class@ number on pupils eligible for FSM –

Dependent variable: *Quantitative Reasoning Test*

Figure 6 Equation representation of model 1 using MLwiN

$$\text{Post\_test}_{ij} \sim N(XB,\ \Omega)$$

$$\text{Post\_test}_{ij} = \beta_{0ij}\text{cons} + 0.527(0.083)\text{Pre\_test}_{ij} + 0.823(0.563)\text{TrialGroup}_{j} + \text{-0.533(0.582)EVERFSM\_6\_SPR17}_{ij} + \text{-0.696(0.866)TrialGroup.EVERFSM\_6\_SPR17}_{ij}$$

$$\beta_{0ij} = 6.676(0.593) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) : \Omega_u = \begin{bmatrix} 3.418(1.001) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) : \Omega_e = \begin{bmatrix} 15.118(1.162) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 2664.628(466 \text{ of } 532 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 13 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval

Intercept	6.68	0.59				
Pre-test	0.53	0.08	6.63		0.37	0.69
Group allocation	0.82	0.56	1.46	0.15	-0.28	1.92
FSM	-0.53	0.58	-0.91	0.36	-1.67	0.61
Group*FSM	-0.70	0.87	-0.80	0.42	-2.41	1.01
u **School**	3.42	1.00				
e	15.12	1.16				
Deviance	2664.63					

Model 2. Multilevel model used to test the impact of 1st lass@ number on pupils eligible for FSM –

Dependent variable: *KS1 Maths*

Figure 7 Equation representation of model 2 using MLwiN

$$KS1\_Maths_{ij} \sim N(XB, \Omega)$$

$$KS1\_Maths_{ij} = \beta_{0ij}cons + 0.073(0.012)Pre\_test_{ij} + 0.109(0.079)TrialGroup_j + 0.005(0.082)EVERFSM\_6\_SPR17_{ij} + -0.308(0.123)EVERFSM\_6\_SPR17.TrialGroup_{ij}$$

$$\beta_{0ij} = 2.952(0.084) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.061(0.020) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.341(0.025) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 954.783(503 of 532 cases in use)

The standard errors are displayed in parentheses after each parameter estimate.

Table 14 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	2.95	0.08	36.88		2.79	3.11
Pre-test	0.07	0.01	7.00		0.05	0.09
Group allocation	0.11	0.08	1.38	0.17	-0.05	0.27
FSM	0.01	0.08	0.13	0.90	-0.15	0.17
Group*FSM	-0.31	0.12	-2.58	0.01	-0.55	-0.07

u **School**	0.06	0.02	3.05			
e	0.34	0.03	13.64			
Deviance	954.78					

Sub-group analyses by socio-economic background defined by eligibility for FSM

USING MlwiN software

Set 1 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of children eligible for FSM

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children = 149 (control= 89; intervention = 60)

Number of schools = 85 (control = 48; intervention = 37)

Table 15 Raw means for FSM eligible

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
Quantitative Reasoning	60	8.62	89	8.44	149

Figure 8 Equation representation of model 1 using MLwiN

$$\text{Post\_test}_{ij} \sim N(XB,\ \Omega)$$

$$\text{Post\_test}_{ij} = \beta_{0ij}\text{cons} + 0.526(0.142)\text{Pre\_test}_{ij} + -0.076(0.798)\text{Intervention}_{j}$$

$$\beta_{0ij} = 6.169(0.814) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) : \Omega_u = \begin{bmatrix} 3.201(2.334) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) : \Omega_e = \begin{bmatrix} 16.211(2.679) \end{bmatrix}$$

$$-2*\text{loglikelihood(IGLS Deviance)} = 862.555(149 \text{ of } 149 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 16 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	6.17	0.81	7.58		4.57	7.76
Pre-test	0.53	0.14	3.70		0.25	0.80
Group allocation	-0.08	0.80	-0.10	0.92	-1.64	1.49
u **School**	3.20	2.33	1.37			
e	16.21	2.68	6.05			
Deviance	862.56					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children = 165 (control= 100; intervention =65)

Number of schools = 91 (control =52; intervention =39)

Table 17 Raw means for FSM eligible

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
KS1 Maths	65	3.14	100	3.27	165

Figure 9 Equation representation of model 2 using MLwiN

$$KS1\_Mathsrecoded_{ij} \sim N(XB, \Omega)$$

$$KS1\_Mathsrecoded_{ij} = \beta_{0ij}cons + 0.066(0.021)pre\_test_{ij} + -0.179(0.114)Intervention_j$$

$$\beta_{0ij} = 2.984(0.116) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.071(0.048) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.359(0.056) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 326.443(165\ of\ 165\ cases\ in\ use)$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 18 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	2.98	0.12	25.72		2.76	3.21
Pre-test	0.07	0.02	3.14		0.02	0.11
Group allocation	-0.18	0.11	-1.57	0.12	-0.40	0.04
u **School**	0.07	0.05	1.48			
e	0.36	0.06	6.41			
Deviance	326.44					

Set 2 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of children not-eligible for FSM

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children = 317 (control= 138; intervention =179)

Number of schools = 119 (control = 56; intervention =63)

Table 19 Raw means for not-eligible for FSM

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
Quantitative Reasoning	179	10.11	138	9.41	317

Figure 10 Equation representation of model 1 using MLwiN

$$post\_test_{ij} \sim N(XB, \Omega)$$

$$post\_test_{ij} = \beta_{0ij}cons + 0.521(0.100)pre\_test_{ij} + 0.768(0.584)Intervention_j$$

$$\beta_{0ij} = 6.778(0.663) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 4.626(1.365) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 13.482(1.338) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 1799.658(317\ of\ 317\ cases\ in\ use)$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 20 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	6.78	0.66	10.22		5.48	8.08
Pre-test	0.52	0.10	5.21		0.33	0.72
Group allocation	0.77	0.58	1.32	0.19	-0.38	1.91
u **School**	4.63	1.37	3.39			
e	13.48	1.34	10.08			
Deviance	1799.66					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children = 338 (control=152; intervention =186)

Number of schools = 122 (control = 59; intervention =63)

Table 21 Raw means for not-eligible for FSM

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
KS1 Maths	186	3.41	152	3.32	338

Figure 11 Equation representation of model 2 using MLwiN

$$\text{KS1\_Mathsrecoded}_{ij} \sim N(XB, \Omega)$$

$$\text{KS1\_Mathsrecoded}_{ij} = \beta_{0ij}\text{cons} + 0.076(0.015)\text{pre\_test}_{ij} + 0.103(0.080)\text{Intervention}_j$$

$$\beta_{0ij} = 2.944(0.094) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.074(0.026) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.314(0.030) \end{bmatrix}$$

$$-2*\text{loglikelihood(IGLS Deviance)} = 627.651(338 \text{ of } 338 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 22 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	2.94	0.09	31.32		2.76	3.13
Pre-test	0.08	0.02	5.07		0.05	0.11
Group allocation	0.10	0.08	1.29	0.20	-0.05	0.26
u **School**	0.07	0.03	2.85			
e	0.31	0.03	10.47			
Deviance	627.65					

Appendix 7. Details of the analyses with subgroups defined by gender

USING MlwiN software

Model 1. Multilevel model used to test the impact of 1st Class@ number on pupils classified by gender - Dependent variable: Quantitative Reasoning Test

Figure 12 Equation representation of model 1 using MLwiN

$$\text{post\_test}_{ij} \sim N(XB, \Omega)$$

$$\text{post\_test}_{ij} = \beta_{0ij}\text{cons} + 0.561(0.080)\text{pre\_test}_{ij} + 0.322(0.638)\text{Intervention group}_{j} + 0.128(0.526)\text{Boy}_{ij} + 0.580(0.746)\text{Boy.Intervention group}_{ij}$$

$$\beta_{0ij} = 6.249(0.588) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 3.831(1.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 14.684(1.091) \end{bmatrix}$$

$$-2*\text{loglikelihood(IGLS Deviance)} = 2801.488(491 \text{ of } 532 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 23 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	6.25	0.59	10.59		5.09	7.41
Pre-test	0.56	0.08	7.00		0.40	0.72
Group allocation	0.32	0.64	0.50	0.62	-0.93	1.57
Gender	0.13	0.53	0.25	0.81	-0.91	1.17

Group*gender	0.58	0.75	0.77	0.44	-0.89	2.05
u **School**	3.83	1.00	3.82			
e	14.68	1.09	13.46			
Deviance	2801.49					

Model 2. Multilevel model used to test the impact of 1st Class@ number on pupils classified by gender –

Dependent variable: *KS1 Maths*

Figure 13 Equation representation of model 2 using MLwiN

$$KS1\_Maths_{ij} \sim N(XB, \Omega)$$

$$KS1\_Maths_{ij} = \beta_{0ij}cons + 0.071(0.012)Pre\_test_{ij} + 0.110(0.092)TrialGroup_j + 0.083(0.078)Gender_{ij} + -0.149(0.112)Gender.TrialGroup_{ij}$$

$$\beta_{0ij} = 2.922(0.086) + u_{0j} + e_{0ij}$$

$$\left[u_{0j}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.062(0.020)\right]$$

$$\left[e_{0ij}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.347(0.025)\right]$$

$$-2*loglikelihood(IGLS\ Deviance) = 963.891(503 \text{ of } 532 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 24 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	2.92	0.09	32.44		2.74	3.10
Pre-test	0.07	0.01	7.00		0.05	0.09
Group allocation	0.11	0.09	1.22	0.22	-0.07	0.29

Gender	0.08	0.08	1.00	0.32	-0.08	0.24
Group*gender	-0.15	0.11	-1.36	0.17	-0.37	0.07
u **School**	0.06	0.02	3.10			
e	0.35	0.03	13.88			
Deviance	963.89					

Sub-group analyses by gender

USING MlwiN software

Set 1 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of boys

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children = 269 (control=119; intervention =150)

Number of schools = 123 (control = 60; intervention = 63)

Table 25 Boys raw means by group allocation

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
Quantitative Reasoning	150	9.95	119	9.01	269

Figure 14 Equation representation of model 1 using MLwiN

$$post\_test_{ij} \sim N(XB, \Omega)$$

$$post\_test_{ij} = \beta_{0ij}cons + 0.532(0.115)pre\_test_{ij} + 0.927(0.589)Intervention_j$$

$$\beta_{0ij} = 6.533(0.704) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 3.900(1.476) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 13.762(1.570) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 1526.609(269\ of\ 269\ cases\ in\ use)$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 26 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	6.53	0.70	9.28		5.15	7.91
Pre-test	0.53	0.12	4.63		0.31	0.76
Group allocation	0.93	0.59	1.57	0.12	-0.23	2.08
u **School**	3.90	1.48	2.64			
e	13.76	1.57	8.77			
Deviance	1526.61					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children = 277 (control=126; intervention =151)

Number of schools = 124 (control =62; intervention =62)

Table 27 Boys raw means by group allocation

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
KS1 Maths	151	3.30	126	3.33	277

Figure 15 Equation representation of model 2 using MLwiN

$$\text{KS1\_Mathsrecoded}_{ij} \sim N(XB, \Omega)$$

$$\text{KS1\_Mathsrecoded}_{ij} = \beta_{0ij}\text{cons} + 0.079(0.018)\text{pre\_test}_{ij} + -0.027(0.082)\text{Intervention}_j$$

$$\beta_{0ij} = 2.962(0.103) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.027(0.030) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.389(0.043) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 541.734(277 of 277 cases in use)

The standard errors are displayed in parentheses after each parameter estimate.

Table 28 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	2.96	0.10			2.76	3.16
Pre-test	0.08	0.02	4.39		0.04	0.11
Group allocation	-0.03	0.08	-0.33	0.74	-0.19	0.13
u **School**	0.03	0.03	0.90			
e	0.39	0.04	9.05			
Deviance	541.73					

Set 2 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of girls

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children = 222 (control=119; intervention =103)

Number of schools = 117 (control =60; intervention =57)

Table 29 Girls raw means by group allocation

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
Quantitative Reasoning	103	9.46	119	9.07	222

Figure 16 Equation representation of model 1 using MLwiN

$$post\_test_{ij} \sim N(XB, \Omega)$$

$$post\_test_{ij} = \beta_{0ij}cons + 0.563(0.108)pre\_test_{ij} + 0.122(0.667)Intervention_j$$

$$\beta_{0ij} = 6.316(0.702) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 4.216(1.907) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 15.478(2.048) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 1285.609(222 \text{ of } 222 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 30 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	6.32	0.70			4.94	7.69
Pre-test	0.56	0.11	5.21		0.35	0.77
Group allocation	0.12	0.67	0.18	0.86	-1.19	1.43
u **School**	4.22	1.91	2.21			
e	15.48	2.05	7.56			
Deviance	1285.61					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children = 226 (control=126; intervention =100)

Number of schools = 119 (control =61; intervention =58)

Table 31 Girls raw means by group allocation

	Intervention group		Control group		
Outcome	N	Unadjusted Mean	N	Unadjusted Mean	N in model
KS1 Maths	100	3.40	126	3.27	226

Figure 17 Equation representation of model 2 using MLwiN

$$KS1\_Mathsrecoded_{ij} \sim N(XB, \Omega)$$

$$KS1\_Mathsrecoded_{ij} = \beta_{0ij}cons + 0.064(0.016)pre\_test_{ij} + 0.108(0.096)Intervention_j$$

$$\beta_{0ij} = 2.950(0.101) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.104(0.039) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.297(0.039) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 425.786(226 \text{ of } 226 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 32 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	p-value	confidence interval	
					LB	UB
Intercept	2.95	0.10			2.75	3.15
Pre-test	0.06	0.02	4.00		0.03	0.10
Group allocation	0.11	0.10	1.13	0.26	-0.08	0.30
u **School**	0.10	0.04	2.67			
e	0.30	0.04	7.62			
Deviance	425.79					

Appendix 8. Details of the analyses for children who met the attendance criteria of attending all sessions or 2/3 of the sessions

USING MlwiN software

Set 1 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of children who attended 30 sessions or more

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children = 363 (control= 238; intervention = 125)

Number of schools = 111 (control = 63; intervention = 48)

Table 33 Raw Mean, Standard Deviation and number of children per group in model 1

Outcome	Group allocation	Mean	SD	N
Quantitative Reasoning	**Control**	9.04	4.41	238
	Intervention	10.08	4.75	125
	Total	9.40	4.55	363

Figure 18 Equation representation of model 1 using MLwiN

$$post\_test_{ij} \sim N(XB, \Omega)$$

$$post\_test_{ij} = \beta_{0ij}cons + 0.634(0.095)pre\_test_{ij} + 1.077(0.551)Intervention\ group_j$$

$$\beta_{0ij} = 5.963(0.567) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 3.027(1.106) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 15.198(1.342) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 2072.842(363\ of\ 363\ cases\ in\ use)$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 34 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

					confidence interval	
	coefficient	standard error	z-ratio	P-value	LB	UB
Intercept	5.96	0.57	10.52		4.85	7.07
Pre-test	0.63	0.10	6.67		0.45	0.82
Group allocation	1.08	0.55	1.96	0.05	-0.003	2.16
u **School**	3.03	1.11	2.74			
e	15.20	1.34	11.33			
Deviance	2072.84					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children = 373 (control= 252; intervention = 121)

Number of schools = 111 (control = 65; intervention = 46)

Table 35 Raw Mean, Standard Deviation and number of children per group in model 1

Outcome	Group allocation	Mean	SD	N
KS1 Maths	**Control**	3.30	0.659	252
	Intervention	3.31	0.684	121
	Total	3.31	0.666	373

Figure 19 Equation representation of model 2 using MLwiN

$$\text{KS1\_Mathsrecoded}_{ij} \sim N(XB, \Omega)$$

$$\text{KS1\_Mathsrecoded}_{ij} = \beta_{0ij}\text{cons} + 0.079(0.015)\text{pre\_test}_{ij} + 0.032(0.081)\text{Intervention group}_{j}$$

$$\beta_{0ij} = 2.921(0.084) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.059(0.023) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.345(0.030) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 710.736(373 of 373 cases in use)

The standard errors are displayed in parentheses after each parameter estimate.

Table 36 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	P-value	confidence interval	
					LB	UB
Intercept	2.92	0.08	34.77		2.76	3.09
Pre-test	0.08	0.02	5.27		0.05	0.11
Group allocation	0.03	0.08	0.40	0.69	-0.13	0.19
u **School**	0.06	0.02	2.57			
e	0.35	0.03	11.50			
Deviance	710.74					

USING MlwiN software

Set 2 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of children who attended 20 sessions or more

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children =440 (control= 238; intervention = 202)

Number of schools = 120 (control = 63; intervention = 57)

Table 37 Raw Means, Standard Deviation and number of children per group in model 1

Outcome	Group allocation	Mean	SD	N
Quantitative Reasoning	**Control**	9.04	4.41	238
	Intervention	9.79	4.62	202
	Total	9.38	4.52	440

Figure 20 Equation representation of model 1 using MLwiN

$$\text{post\_test}_{ij} \sim N(XB, \Omega)$$

$$\text{post\_test}_{ij} = \beta_{0ij}\text{cons} + 0.550(0.086)\text{pre\_test}_{ij} + 0.780(0.518)\text{Intervention group}_j$$

$$\beta_{0ij} = 6.363(0.546) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 3.954(1.075) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 14.637(1.156) \end{bmatrix}$$

$$-2*\text{loglikelihood(IGLS Deviance)} = 2511.660(440 \text{ of } 440 \text{ cases in use})$$

UNITS:

SchoolID: 120 (of 120) in use

The standard errors are displayed in parentheses after each parameter estimate.

Table 38 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	P-value	confidence interval LB	UB
Intercept	6.36	0.55	11.65		5.29	7.43
Pre-test	0.55	0.09	6.40		0.38	0.72
Group allocation	0.78	0.52	1.51	0.13	-0.24	1.80
u **School**	3.95	1.08				
e	14.64	1.16				
Deviance	2617.09					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children = 447 (control= 252; intervention = 195)

Number of schools = 122 (control = 65; intervention =57)

Table 39 Raw Means, Standard Deviation and number of children per group in model 1

Outcome	Group allocation	Mean	SD	N
KS1 Maths	**Control**	3.30	0.66	252
	Intervention	3.32	0.68	195
	Total	3.31	0.67	447

Figure 21 Equation representation of model 2 using MLwiN

$$\text{KS1\_Mathsrecoded}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{KS1\_Mathsrecoded}_{ij} = \beta_{0ij}\text{cons} + 0.076(0.013)\text{pre\_test}_{ij} + 0.025(0.073)\text{Intervention group}_{j}$$

$$\beta_{0ij} = 2.936(0.080) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.067(0.022) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.342(0.027) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 853.786(447 of 447 cases in use)

UNITS:

SchoolID: 122 (of 122) in use

The standard errors are displayed in parentheses after each parameter estimate.

Table 40 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	P-value	confidence interval	
					LB	UB
Intercept	2.93	0.08	37.62		2.78	3.09
Pre-test	0.08	0.01	5.92		0.05	0.10
Group allocation	0.03	0.07	0.40	0.69	-0.11	0.17
u **School**	0.07	0.02	3.05			
e	0.34	0.03	13.15			
Deviance	853.79					

Appendix 9. Details of the analysis with the brief scale of the Quantitative Reasoning Test

Model 1. Multilevel model used to test the impact of 1stClass@ number –

Dependent variable: *Brief scale, Quantitative Reasoning Test*

USING R ANALYSIS (eefAnalytics package)

Table 41 Fixed effects calculated with eefAnalytics R package

Fixed effects: Estimates and confidence intervals for predictors specified in the model.			
	Estimate	**95% LB**	**95% UB**
Intercept	4.17	3.60	4.74
Group allocation	0.48	-0.12	1.08
Pre-test	0.49	0.36	0.63

Table 42 Covariance Matrix using eefAnalytics

Vector of variance decomposition into between cluster variance (Schools) and within cluster variance (Pupils). It also contains the intra-cluster correlation (ICC).			
Schools	Pupils	Total	ICC
1.51	5.78	7.29	0.21

Table 43 Effect size calculated with eefAnalytics R package

Hedges' g effect size for the intervention, the confidence intervals are 95%			
	Estimate	95% LB	95% UB
Within	0.20	-0.05	0.45
Total	0.18	-0.04	0.40

USING MlwiN software

Figure 22 Equation representation of model 1 using MLwiN

$$\text{Post\_Test\_12items}_{ij} \sim N(XB, \Omega)$$

$$\text{Post\_Test\_12items}_{ij} = \beta_{0ij}\text{cons} + 0.494(0.070)\text{Pre\_Test\_12items}_{ij} + 0.484(0.304)\text{Intervention group}_{j}$$

$$\beta_{0ij} = 4.174(0.289) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 1.455(0.387) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 5.770(0.430) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 2340.671(491 of 532 cases in use)

The standard errors are displayed in parentheses after each parameter estimate.

Table 44 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	p-value	confidence interval	
Intercept	4.17	0.29	14.38		3.60	4.74
Pre-test	0.49	0.07	7.00		0.35	0.63
Group allocation	0.48	0.30	1.60	0.10	-0.11	1.07
***u* School**	1.46	0.39	3.76			
e	5.77	0.43	13.42			
Deviance	2340.67					

Appendix 10. Details of the analyses for pupils up to and above the median in the pre-test of Quantitative Reasoning

USING MlwiN software

Set 1 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of children that scored up to the median at pre-test

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children = 255 (control= 123; intervention = 132)

Number of schools = 108 (control = 51; intervention = 57)

Figure 23 Equation representation of model 1 using MLwiN

$$\text{Post\_test}_{ij} \sim N(XB,\ \Omega)$$

$$\text{Post\_test}_{ij} = \beta_{0ij}\text{cons} + 0.516(0.267)\text{Pre\_test}_{ij} + 1.252(0.616)\text{TrialGroup}_{j}$$

$$\beta_{0ij} = 6.104(0.887) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 2.781(1.466) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 16.083(1.805) \end{bmatrix}$$

$$-2*\textit{loglikelihood(IGLS Deviance)} = 1467.981(255 \text{ of } 277 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 45 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	P-value	confidence interval	
					LB	UB
Intercept	6.10	0.89	6.88		4.37	7.84
Pre-test	0.52	0.27	1.93		-0.01	1.04
Group allocation	1.25	0.62	2.03	0.04	0.05	2.46
u **School**	2.78	1.47	1.90			
e	16.08	1.81	8.91			
Deviance	1467.98					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children = 264 (control= 134; intervention = 130)

Number of schools = 110 (control = 53; intervention = 57)

Figure 24 Equation representation of model 2 using MLwiN

$$KS1\_Maths_{ij} \sim N(XB, \Omega)$$

$$KS1\_Maths_{ij} = \beta_{0ij}cons + 0.066(0.041)Pre\_test_{ij} + 0.086(0.094)TrialGroup_j$$

$$\beta_{0ij} = 2.888(0.137) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.081(0.034) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.355(0.039) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 522.310(264 \text{ of } 277 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 46 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	P-value	confidence interval	
					LB	UB
Intercept	2.89	0.14	21.08		2.62	3.16
Pre-test	0.07	0.04	1.61		-0.01	0.15
Group allocation	0.09	0.09	0.92	0.36	-0.10	0.27
u **School**	2.78	1.47	1.90			
e	16.08	1.81	8.91			
Deviance	1467.98					

USING MlwiN software

Set 2 of analyses. Multilevel models used to test the impact of 1st Class@ number with the subgroup of children that scored above the median at pre-test

Model 1. Multilevel model used to test the impact of 1st Class@ number – Dependent variable: *Quantitative Reasoning Test*

Number of children =236 (control= 115; intervention = 121)

Number of schools = 104 (control = 48; intervention = 56)

Figure 25 Equation representation of model 1 using MLwiN

$$\text{Post\_test}_{ij} \sim N(XB,\ \Omega)$$

$$\text{Post\_test}_{ij} = \beta_{0ij}\text{cons} + 0.437(0.129)\text{Pre\_test}_{ij} + 0.123(0.613)\text{TrialGroup}_{j}$$

$$\beta_{0ij} = 7.524(0.987) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 2.472(1.469) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 15.601(1.845) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 1349.319(236\ of\ 255\ cases\ in\ use)$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 47 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 1

	coefficient	standard error	z-ratio	P-value	confidence interval	
					LB	UB
Intercept	7.52	0.99	7.62		5.59	9.46
Pre-test	0.44	0.13	3.39		0.18	0.69
Group allocation	0.12	0.61	0.20	0.83	-1.08	1.32
u **School**	2.47	1.47	1.68			
e	15.60	1.85	8.46			
Deviance	1349.32					

Model 2. Multilevel model used to test the impact of 1st Class@ number –

Dependent variable: KS1 Maths

Number of children =239 (control= 118; intervention = 121)

Number of schools = 104 (control = 48; intervention = 56)

Figure 26 Equation representation of model 2 using MLwiN

$$KS1\_Maths_{ij} \sim N(XB, \Omega)$$

$$KS1\_Maths_{ij} = \beta_{0ij}cons + 0.019(0.019)Pre\_test_{ij} + -0.067(0.088)TrialGroup_j$$

$$\beta_{0ij} = 3.415(0.142) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.059(0.030) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.304(0.036) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 431.062(239 \text{ of } 255 \text{ cases in use})$$

The standard errors are displayed in parentheses after each parameter estimate.

Table 48 Coefficient, standard error, Z-ratio and confidence interval for each parameter in model 2

	coefficient	standard error	z-ratio	P-value	confidence interval	
					LB	UB
Intercept	3.42	0.14	24.05		3.14	3.69
Pre-test	0.02	0.02	1.00		-0.02	0.06
Group allocation	-0.07	0.09	-0.76	0.45	-0.24	0.11
u **School**	0.06	0.03	1.97			
e	0.30	0.04	8.44			
Deviance	431.06					

Appendix 11. Detailed analysis of the responses to the interviews with Link Teachers in the Control and Intervention schools

School Interviews

A phone interview was conducted to find out what "business as usual" meant for both the control and intervention schools. The interviews were conducted over the phone by a member of the evaluation team with the Link Teacher (an experienced teacher in the school who was nominated to liaise between the school and the evaluation team). The evaluation team wanted to find out what activities were in place to support Year 2 children's mathematics learning, what resources were used in the classroom, whether children had received one-to-one or small group supplementary instruction, and what programmes were used with them.

The phone interviews were carried out between April and July 2017. It was originally planned that interviews would be conducted before Key Stage 1 SATs, however some Link Teachers did not want the interviews to take place until after SATs had finished. Twenty schools were randomly selected and contacted to take part, however only (n=8) control interviews were able to be conducted. The remaining schools (n=2) had agreed to take part but one cancelled the interview on two occasions due to unforeseeable incidents in the school and the other was unable to be contacted despite repeated attempts from both the evaluation and intervention teams. Of the randomly selected intervention schools (n=10), all responded and took part in the phone interview.

The interviews followed a structured design and were conducted by a member of the evaluation team who made notes on the Link Teachers' answers. A report of each interview was then written by the researcher. Analysis of these reports highlighted several similarities in their approach to teaching mathematics and what constituted "business as usual" for their school.

Interview Themes

Three superordinate themes arose from the interviews with both the intervention and control schools: (1) activities, supplementary instruction, and programmes; (2) resources used in the classroom; and (3) satisfaction with their current provision. A summary of the percentage frequency of reported occurrence for each theme can be found in Table 6.1 of this Appendix.

1. Activities, Supplementary Instruction, and Programmes

Many of the teachers that were interviewed provided information about activities, supplementary instruction and the programmes that they used. For this reason they were collated to form the superordinate theme. Within the answers provided by the interviewees, several recurring themes arose relating to the activities that the school employs and considers "business as usual".

- *Level of support* –Most intervention schools (9/10) and control schools (6/8) reported that support is provided within the class or in groups outside the class; 1 of the intervention schools and 2 of the control schools reported that they provided support at a class and group level as well.
- *Who delivers the activities and support* – Intervention schools reported that their additional maths activities/instruction were generally delivered by a TA (6/10), by either a teacher or TA (3/10), and by an intervention teacher employed to carry out additional work (1/10). Control schools reported that their additional maths activities were delivered by a TA (1/8), delivered by either a teacher or TA (6/8) and by the class teacher (1/8).

- *Type of activity*
  - Booster Sessions – A booster session is delivered in small groups or one to one by either the TA or the class teacher, based on the content of the maths lesson from the previous or same day. Children are selected based on the teacher's assessment of whether each child had understood and accessed the lesson and whether they needed to address misconceptions. In this case, it is not only the lowest achieving children who receive additional instruction, but all those who were not able to access the lesson at their differentiated level. In addition to 1st Class@Number, the majority of the intervention schools (7/10) reported that they were using booster sessions as standard in their year 2 classes. All 8 of the control schools reported using booster sessions.
  - SATs Preparation – In addition to 1st Class@Number and any booster sessions, almost all intervention schools (9/10) reported that their school was also running specific sessions with their year 2 children in preparation for their KS1 SATs. Of the control schools interviewed (3/8) were delivering additional SATs specific sessions in addition to running booster sessions.
  - External intervention(s) – In addition to 1st Class@Number, 2 intervention schools reported that they had employed a class level programme in the mornings, such as Mathletics, to help their children with maths recall and 2 intervention schools were following a specific intervention with their lowest achieving pupils in addition to 1st Class@Number. Three control schools reported that they had employed a class level programme in the mornings, such as Mathletics, to help their children with maths recall.
- *Duration of activity* – Half of the intervention and of the control schools were unable to provide an estimate of the duration of their additional activities as they used whatever length of time was required for each session. Thus the analyses reported here are based on 5 intervention and 4 control schools. The mean duration reported by intervention schools for their additional sessions and activities sessions was 25 minutes. One of the intervention schools ran a further hour long improvement session once a week with the whole class so some of this time may have been directed to further maths instruction. For control schools the mean time was 35 minutes. One of the control schools ran also an hour long 'catch up' session once per week with the whole class; a second one ran an hour long session after school attendance was based on who the teacher thought needed support, and was split between maths and literacy.
- *Frequency of activity* – Many intervention schools (5/10) and control schools (5/8) were unable to provide an estimate of how often the children in their schools received or took part in the supplementary instruction, particularly in relation to booster sessions where attendance related to the children's progress in the lesson. Four intervention schools and 3 control schools reported that whilst sessions were daily or weekly, but they were unable to provide a reliable estimate of children's attendance due to the reactive nature of the intervention and the difficulty of keeping records.
- *Homework* – Half of the intervention schools (3/5) and most of the control schools (7/8) reported that they were sending out maths homework to their year 2 children on a regular basis. Of the intervention schools that sent out homework, 3 also ran a homework club where the children could receive additional one-to-one help from a TA or class teacher. No control schools reported having a homework club that the children could attend.

2. Resources used in the classroom

Part of understanding what business as usual meant included understanding how similar the resources that are used for 1stClass@Number were to those being used in maths lessons. Two sets of materials stood out as being used regularly in the schools

- *Numicon* –The majority of the intervention schools (9/10) and all 8 of the control schools reported that their children had access to Numicon in the classroom.
- *Base 10 Blocks* (Dienes blocks) –over half of the intervention schools (6/10) and the control schools (6/8) reported that their children had access to Base 10 blocks in maths lessons. Two of these intervention schools and 1 of the control schools reported either that Base 10 blocks were used extensively or that they were the primary concrete resource used to help children.

3. Satisfaction with current provision

This superordinate theme arose from the answers that the interviewees provided in relation to the questions regarding parental feedback and whether they believed that the school would continue to use its current provision, either 1stClass@Number or "business as usual".

- *Feedback from parents* – Parental feedback was reported by intervention schools (8/10) and control schools (7/8). Five of the intervention schools that reported receiving parental feedback said that it had been positive or very positive and the remaining 3 schools that had not received direct feedback knew that the children were completing the 1stClass@Number homework intervention. Feedback in the control schools was limited to parents evening; however, parents were generally happy with the progress that their children were making.
- *Lack of available funding* – Funding issues were reported by both the intervention schools (4/10) and the control schools (5/8). Two of the intervention schools reported that they would be unable to fund the training for another TA to deliver 1stClass@Number if their current TA were to leave. The 5 control schools reported that their budget for interventions is tight and they would not be able to fund an intervention.
- *Continuation of current provision* – All 10 of the intervention schools interviewed were pleased with 1stClass@Number and were very keen to continue with the intervention; almost half (4/10) reported that they had already begun working with a second group. All 8 of the control schools expected to carry on with their current provision; 4 control schools would be interested in adopting a maths intervention if it was effective and especially if it was offered free of charge as part of an evaluation.

Analysis of the Link Teacher interviews suggests that there are many similarities between the intervention schools' and the control schools' definitions of "business as usual"; both groups had activities and extra support for pupils on a regular basis.

The LT in control schools who were interviewed reported a higher rate of the use of booster sessions (8/8 compared with 7/10 in intervention schools). Intervention schools reported a much higher rate of the use of SATs preparation sessions ((9/10 compared with 3/8 in control schools). A higher proportion of control schools found finding funding for interventions hard to come by (5/8 compared to 4/10 in intervention schools), but were more likely to say that they would look at adopting another intervention suggesting that there is still a desire for KS1 maths interventions in these schools.

Intervention school interventions were more likely to be delivered by a TA than those in the control schools.

In summary, the themes reported in the interviews recurred regularly, which leads the evaluation team to be sufficiently satisfied that the interviews are representative of the intervention and the control

schools procedures for working with pupils who are struggling with maths at the start of Year 2. The analysis of the interviews suggested that intervention and control schools interpreted "business as usual" similarly.

Table: Comparison of interviewed schools that reported similar themes. Highlighted themes show a difference of >20% between the Intervention and Control schools.

Superordinate Themes, *Themes,* Sub Themes		Percent responses for each theme
Activities, Supplementary Instruction, and Programmes		
Level of support		100
Delivered by		100
Type of activity	Booster	70
	SATs Prep	90
	External	40
Duration		90
Frequency		90
Homework		50
Resources used in the classroom		
Numicon		90
Base 10 blocks		60
Satisfaction with current provision		
Feedback from parents		80
Lack of available funding		40
Continuation with current provision		100

Superordinate Themes, *Themes,* Sub Themes		Percent responses for each theme
Activities, Supplementary Instruction, and Programmes		
Level of support		100
Delivered by		100
Type of activity	Booster	100
	SATs Prep	37.5
	External	37.5
Duration		100
Frequency		100
Homework		87.5
Resources used in the classroom		
Numicon		100
Base 10 blocks		75
Satisfaction with current provision		
Feedback from parents		87.5
Lack of available funding		62.5
Continuation with current provision		100

Control schools.

NCETM, Maths hub	**800-** School has received maths mastery training from the White Rose maths hub. **3800-** School is looking at getting involved with the local maths hub. **3900-** School does dip into NCETM as an overview but combines it with the "maths- no problem" Singaporean based maths workbooks. **5800-** School has been involved in the NCETM and dips into their maths mastery model for curriculum planning. **5900-** School currently follows the NCETM for its long term planning and materials. **11600-** School does not follow the NCETM maths mastery model but uses some of the training such as bar modelling and fractions when it relates to the lesson being taught. **11900- N/A** **12000- N/A**

Intervention schools.

NCETM, Maths hub	**900-** The school has been following the NCETM guides to help with the planning of their lessons. **5200-** The whole school has been following the NCETM Maths Hub scheme **6200- N/A** **7100- N/A** **7700-** The school follows the NCETM Maths Hub Maths mastery model for teaching mathematics. **9300- N/A** **11100- N/A** **11300- N/A** **11500-** The school is following the NCETM maths hub maths mastery lessons based on the South Yorkshire.

	50000- N/A

Appendix 12: Padlock rating

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 **[n/a]**		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 **[n/a]**		
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:** lowest of the three ratings for design, power and attrition = 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 13: 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
£ £ £ £ £	*Very low:* less than £80 per pupil per year.
£ £ £ £ £	*Low:* up to about £200 per pupil per year.
£ £ £ £ £	*Moderate:* up to about £700 per pupil per year.
£ £ £ £ £	*High:* up to £1,200 per pupil per year.
£ £ £ £ £	*Very high:* over £1,200 per pupil per year.

Education
Endowment
Foundation

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