



onebillion

Evaluation Report

July 2019

Terezinha Nunes, Lars-Erik Malmberg, Deborah Evans,
David Sanders-Ellis, Susan Baker, Rossana Barros, Peter
Bryant, Maria Evangelou





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 (formerly Impetus Trust) 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:

 Jonathan Kay
Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

 0207 802 1653

 jonathan.kay@eefoundation.org.uk

 www.educationendowmentfoundation.org.uk



Contents

| | |
|---|-----|
| Contents | 3 |
| About the evaluator | 4 |
| Executive summary | 7 |
| Introduction | 9 |
| Methods..... | 19 |
| Impact evaluation..... | 33 |
| Implementation and process evaluation | 44 |
| Conclusion | 61 |
| References..... | 63 |
| Appendix A: EEF cost rating scale | 65 |
| Appendix B: Security classification of trial findings | 66 |
| Appendix 1: Protocol | 67 |
| Appendix 2: Pre-test and Post- test descriptions..... | 106 |
| Appendix 3: Details of the models and syntax for the analyses | 112 |
| Appendix 4: Instruments used for implementation and process evaluation..... | 61 |
| Appendix 5: Schools' location and dates of observations..... | 114 |
| Appendix 6: Multiple regression analyses | 116 |

About the evaluator

The project was independently evaluated by a team from the Department of Education, University of Oxford: Terezinha Nunes, Lars-Erik Malmberg, Deborah Evans, David Sanders-Ellis, Susan Baker, Rossana Barros, Peter Bryant, and Maria Evangelou.

The lead evaluator was Terezinha Nunes.

Contact details

Name: Terezinha Nunes

Address: Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 8AD

Tel: 01865 284893

Email: terezinha.nunes@education.ox.ac.uk

List of figures

Figure 1: Logic model produced by the intervention team

Figure 2: Participant flow diagram showing numbers from recruitment to post-test

Figure 3: Distribution of scores at post-test in PTM 6 (N=1089)

Figure 4: Means at post-test for the intervention and control groups by FSM status

Figure 5: School means at post-test by school means for the pre-test; the regression line is plotted for all schools independently of group

Figure 6: Frequency of pupils who attended different numbers of sessions

Figure 7: School means at post-test by school means at pre-test with schools classified by their compliance level; the regression line is based on the full sample

Figure 8: School means at post-test by school means at pre-test with differentiation of schools by the TAs' perceived role in the intervention delivery

List of tables

Table 1: Summary of impact on primary outcome

Table 2: Overview of the design

Table 3: Minimum detectable effect size at different stages

Table 4: Overview of the instruments used to measure the implementation of the enabling factors

Table 5: Measures of implementation fidelity

Table 6: Instruments used to collect information about contextual factors that could contribute to the effect of the intervention

Table 7: Instruments used to collect information about what business as usual meant in this trial

Table 8: Timeline for the implementation of the project

Table 9: Pupil and school characteristics

Table 10a: Raw means, confidence intervals (CI) and effect size for the outcome measure at post-test

Table 10b: Effect size estimation

Table 11: Post-test outcome in intervention and control groups by FSM eligibility

Table 12: 'One-off' and 'ongoing' costs (£) to implement with two groups each year over a three-year period

Table 13: Staff time resources required to implement intervention with two groups each year

Table 14: Cumulative cost per pupil and average cost per pupil, per year

Table 15: TAs' evaluation of the training they received for implementation of the intervention: mean agreement scores

Table 16: Cross tabulation of how TAs viewed their role (from TA questionnaire) and how they delivered the intervention (from the observation)

Table 17: Number of instances that a particular TA action or a particular pupil action was observed during the first 15 minutes of observation

Table 18: Number of instances that a particular TA action was observed during the second 15 minutes of observation

Table 19: Frequency of responses by TAs in the TAs' questionnaire indicating how often they thought the pupils needed support when using each of the apps

Table 20: Frequency of responses by TAs indicating how often they thought the pupils enjoyed the apps

List of boxes

Box 1: Illustrative answers by TAs who were classified as educators or observers by the evaluation team

Box 2: A sample of TAs' comments in response to open questions included in the TA questionnaire

Box 3: TAs' pedagogical and technological tips

List of appendices

Appendix A: EEF cost rating scale

Appendix B: EEF Security rating

Appendix 1: Protocol

Appendix 2: Pre-test and post-test descriptions

Appendix 3: Details of the models and syntax for the analyses

Appendix 4: Instruments used for implementation and process evaluation

Appendix 5: Schools' locations and dates of observations

Appendix 6: Multiple regression analyses

Executive summary

The project

The *onebillion* programme consists of two tablet apps, Maths 3–5 and Maths 4–6, that are designed to reinforce basic mathematical skills learned in the classroom. The apps are aimed at pupils aged 3–5 and 4–6 respectively and consist of mathematical activities organised around different topics such as counting, shape, and measures. Each topic is followed by an end-of-topic quiz which pupils are expected to pass before they move onto the next topic. The activities are aligned with the aims of the Early Years Foundation Stage and the National Curriculum in England. The apps were developed by *onebillion*, a not-for-profit organisation, and in this trial the delivery of the programme was led by a team from the University of Nottingham.

This project tested the impact of *onebillion* when it is used by schools as a targeted intervention with small groups of pupils. The programme was targeted at Year 1 pupils (aged 5–6) who had been identified by their teachers as being in the lower half of the class in mathematics at the start of the school year. Pupils worked through the apps at their own pace and sessions were usually supervised by a teaching assistant (TA) or, in some cases, a teacher. The TAs' main tasks were to ensure that all the pupils had access to the *onebillion* apps throughout the session and to solve any technical problems. TAs were only occasionally expected to provide pedagogical support, for example if the pupil was struggling to progress past an end-of-topic quiz. The programme was designed to last for 12 weeks with four 30-minute sessions per week. The delivery team recommended to schools that the intervention is delivered outside of the time allocated to normal maths lessons. The Nottingham University team provided one half-day of face-to-face training, an Implementation Manual and instructional videos to support schools to use the apps.

onebillion was evaluated using a two-arm randomised controlled trial (RCT): 113 schools were randomised to either receive the intervention or continue with business as usual teaching. The primary outcome was performance on a maths test (Progress Test in Maths [PTM], GL Assessment, 2015). A process evaluation involved interviews with TAs, observations of intervention sessions, and questionnaires given to TAs and other school staff. Recruitment for the trial started in September 2017 and the post-test took place in July 2018.

Key conclusions

1. Pupils who received *onebillion* made an additional three months' progress in maths compared to the control group. This result has very high security.
2. Pupils eligible for free school meals made 2 fewer months' progress in maths if they received *onebillion* compared to those in the control group. However, this analysis involves a smaller number of pupils so we are unable to confidently claim that this negative impact is likely to occur for FSM-eligible pupils outside of this research project.
3. The process evaluation suggested that the impact of the programme might be influenced by the amount of the pedagogical support given to the pupils during the intervention sessions. Exploratory analysis suggested that pupils tended to do better when supervised by TAs who thought that their role was to teach concepts when the pupils had difficulty.
4. In this project, teachers started with Maths 3–5 and then moved to the Maths 4–6 app. TAs reported that pupils enjoyed Maths 3–5 more and required less pedagogical support to use it.
5. Further research is needed on the nature of the pedagogical support that works best in *onebillion* sessions and the effects of the programme on the mathematics attainment of pupils entitled to FSM.

EEF security rating

These findings have a very high security rating. This was an efficacy trial which aimed to understand the impact of the programme under ideal, developer-led conditions. The trial was a well-designed and well-powered RCT. Relatively few pupils (3%) who started the trial were not included in the final analysis. The pupils in the *onebillion* schools were similar to those in the comparison schools in terms of prior attainment.

Additional findings

Pupils eligible for FSM made two fewer months' progress in maths if they received *onebillion* compared to those in the control group. These results have lower security than the overall findings because of the smaller number of pupils involved. There was no indication from the process evaluation of why there might have been a negative impact on pupils eligible for FSM.

The process evaluation suggested that TAs had different ideas about their role in intervention sessions. Some TAs believed that they should take a direct pedagogical role, while other TAs thought their role was to supervise the pupils and ensure that they stayed on task. There was some evidence of a connection between the role that TAs thought that they should adopt in intervention sessions and the pupils' scores in the post-test. Exploratory analysis suggested that pupils supervised by TAs who thought that their role was to teach the pupils how to solve problems tended to do better in the post-test.


TAs reported that pupils enjoyed participation in the intervention. TAs were also positive about their participation in the programme. Staff from 16 of the schools were initially unable to attend the face-to-face training due to travel disruption so they were trained through an online iTunesU course with seven demonstration videos and a phone call from the project team. Both forms of training received high approval ratings from participating TAs.

Cost

The average cost of *onebillion* for one school was around £3850, or £64 per pupil per year when averaged over three years. Delivery of *onebillion* required a total of 59 hours of TA time. This included time to attend training, prepare for the intervention sessions and supervise the intervention sessions.

Impact

Table 1: Summary of impact on primary outcome

| Outcome/ Group | Effect size (95% confidence interval) | Estimated months' progress | EEF security rating | No. of pupils | p value | EEF cost rating |
|-----------------------|--|----------------------------------|---|---------------|------------|-----------------|
| Maths | 0.24 (0.12, 0.36) | 3 |  | 1089 | p=0.000078 | £ £ £ £ £ |
| Maths (FSM pupils) | -0.10 (-0.33, 0.14) | -2 | N/A | 271 | p=0.43 | £ £ £ £ £ |

Introduction

Background evidence

This evaluation will assess the impact of the *onebillion maths apps* (henceforth referred to simply as ‘the intervention’ or ‘the apps’) on pupils’ mathematical outcomes in England. The apps were developed by *onebillion*, a not-for-profit organisation, and are commercially available. They were used with some indication of success in previous research in Malawi (Pitchford, 2015) and in England (Outhwaite *et al.*, 2017; Outhwaite *et al.*, 2018).

The content of the intervention is curriculum based, rather than motivated by research on learning trajectories in mathematics. It includes two levels, one labelled as ‘Maths 3–5 app’ and the second labelled as ‘Maths 4–6 app’. The Maths 3–5 app contains 10 topics and the Maths 4–6 app contains 18 topics; these are aligned with the aims of the Early Years Foundation Stage and the National Curriculum in England (Outhwaite *et al.*, 2018). Each topic has several activities (the total number of activities is 178). Some examples of topics are counting (with activities organised in different levels, taking counting up to 100), classification by different criteria (shape, colour), shape (geometrical shape vocabulary, symmetry), lines and patterns (straight or curved; repetitions of figures in a pattern), position (spatial relations vocabulary), measures (length, time, mass, capacity), addition and subtraction (arithmetic with pictures, number bonds, number line work), sharing and fractions (half, quarter). There is no overlap between the activities in the Maths 3–5 app and those in the Maths 4–6 app, and so the two apps can be viewed as one progressive sequence. For example, counting and learning numerical symbols in the Maths 3–5 app reaches 10 while the Maths 4–6 app starts with counting to 20 and continues to 100. In this trial, all pupils started with the Maths 3–5 app and moved on to the Maths 4–6 app when they had completed the Maths 3–5 app. The intervention team made this choice because, by starting with easier items, all pupils would have positive experiences with the apps initially and those with little previous experience with iPads would have the opportunity to learn more about how to use an iPad. The intervention was individually paced to reflect the way that the apps would be used when schools adopt them outside a research project.

The displays are designed to be attractive. Teaching how to execute the activity is part of the displays, which include a voice (the teacher in the app) that provides oral explanations and visual demonstration about what the pupil is expected to do. The instructions can be repeated if the pupil presses the appropriate button on the screen, a feature that allows pupils to control the number of times they want to hear an explanation about what to do in the activity. Feedback is given after the pupil’s answer by a sound if there is a mistake, or by a tick and a sound if the answer is correct.

Pupils are encouraged to work through the activities in each topic in order to master them. When a pupil completes an activity, the activity to be attempted next is indicated by it flashing on the screen. Within each topic, there are up to seven activities and an end-of-topic quiz. The Implementation Manual distributed to schools in the trial indicates that the pupils need to complete all of the activities before they can access the quiz. If the pupils answer all ten questions in the quiz correctly, they are rewarded with a certificate. If a pupil does not pass the quiz, the Implementation Manual encourages the teaching assistant (TA) to try repeating the quiz with the pupil by saying the questions, rather than just letting the pupil listen to the questions from the app. If the pupils are still challenged by some quiz questions, they may need further practice on the topic by working through the activities again. This can be done independently or with support from the adult running the intervention. However, the apps do not restrict what the pupils can access when an activity has been completed and they can access a different activity from the one intended as the next activity in the app. Once pupils have passed the quiz, they are expected to move on to the next topic. For this reason, the intervention is described by Outhwaite *et al.* (2017) as individually paced.

With respect to the content of the apps, it is noteworthy that some activities are similar to those used in research in developmental psychology (for example, the give N task; Wynn, 1990) or used in tests of cognitive skills (for example, placing pictures of events in logical order is a task used in the Wechsler Intelligence Scale for Children, WISC; Wechsler, 1992). Some of the activities are ordered according to results in developmental psychology (for example, classification by a single criterion before classification by two criteria; addition and subtraction with objects before addition and subtraction with symbols). However, developmental psychology research cannot be said to provide the basis for the choice of activities since the activities do not typically focus on promoting the learning of concepts considered central to mathematical development. (Not included as aims in the activities, for example, are learning about the counting principles, understanding the inverse relation between operations, and solving different types of addition and subtraction word problems.) Most of the activities are aligned with the National Curriculum in England and, to our knowledge, there

is no research to indicate whether there is a particular order of acquisition for these activities (for example, learning about odd and even numbers before learning how to count to 50).

The theoretical background for the design of the activities comes from studies in educational psychology. According to Outhwaite *et al.* (2018), the design of the activities incorporates the principles of active learning (as the pupil is continuously answering questions) and of direct instruction (as the teacher in the app instructs the pupil and models the responses). Feedback is immediate and contingent upon the pupil's responses; repetition is built into the apps in order to provide several opportunities for practice without the costs associated with teacher supervision. These features suggest that the apps fit the category of computer assisted instruction (CAI).

In brief, the design of the activities is based on a pedagogical theory (a theory of 'how to teach') but the content of the activities is curriculum based rather than based on a theory about pupils' trajectories in mathematics learning (a theory of 'what to teach and in which order').

In this trial the pupils worked individually, as expected, but they were organised in small groups in the same room at the same time, supervised by a nominated member of staff, who was a teacher or a TA; for brevity, the nominated member of staff is referred to as TA throughout this report. The number of pupils in the group was decided by the school; ten schools worked with groups of five pupils while the remaining schools worked with groups of nine or ten pupils.

The role of the TA in this trial (described in the implementation manual provided by the intervention team to the TAs delivering the intervention) was to prepare the apps for the session, to 'give technical support to pupils, which might mean adjusting the volume or navigating the app, and to give learning support, by making sure children are attending to the learning and working through the activities at their own pace' (Implementation Manual, page 16). In order to support the data collection, TAs were asked to track pupil attendance and progress during the intervention by filling in a specifically designed register.

Significance

The first contribution of this study is to provide robust evidence on the impact of the apps. Evidence regarding the efficacy specifically of the *onebillion* maths apps has been accumulating since 2015. Pitchford (2015) reported the first evaluation of the intervention in Malawi that used a randomised controlled trial (RCT) design. The design included the intervention group, a non-maths tablet based intervention, and a business as usual group. The randomisation procedures are described as within-class randomisation; however, the number of participants in the three groups was quite uneven (113 in Standards 1–3 [which is roughly equivalent to grade level in the local school context] in the intervention group, 112 also in Standards 1–3 in the business as usual group, and 85 in Standards 2–3 and in the non-maths tablet based intervention) and one of the groups (the non-maths tablet intervention) was not represented in Standard 1. The intervention was delivered outside the classroom and monitored by teachers over eight weeks for about ten hours. All measures were designed by the intervention team; two of them were tablet based. This is a weakness of the study, recognised by the authors. The interaction between time of testing (pre- and post-test) and group (intervention, business as usual, and non-maths tablet), which would be expected to be significant if the intervention had an effect, was not found for Standard 1, but it was found for Standard 2 on one of the measures administered on the tablet and for Standard 3 on the other tablet-based measure. Although the results are inconsistent, there is some evidence of impact of the intervention, but the design of the study has limitations.

Subsequently Outhwaite *et al.* (2018) reported a study in Nottingham carried out in 12 schools with approximately 400 participants. Within each school, the children were randomly allocated either (1) to a control group, or (2) to an intervention group that used the apps while the control group was receiving instruction in small groups from the teacher (thus this group did not receive additional time on maths instruction), or (3) to an intervention group that used the apps as additional instruction in maths. Children in the intervention groups used the apps for 12 weeks in daily sessions of 30 minutes each in a quiet area of the classroom, supervised by a member of the teaching staff. The outcome measure was a test designed independently of the intervention group, the Progress Test in Maths 5 (PTM 5). Similarly to the intervention content, the test was designed to match the aims of the English Early Years Foundation Stage and the start of Key Stage 1 (KS1). PTM 5 was used by Outhwaite *et al.* (2018) as the pre- and post-test; in the present trial, it was used only as a pre-test. The intervention group who had additional practice in maths performed significantly better at post-test than the control group (Cohen's *d* effect size=0.31; 95% CI 0.06-0.55) but did not differ from the group who used the apps without additional time on maths; this latter intervention group did not differ significantly from the control

group by a two-tailed traditional confidence level (effect size 0.21, 95% CI -0.03-0.46). Thus this trial shows a significant effect of the use of the apps when it provides additional practice with maths. However, the trial has a limitation, which is a relatively high rate of attrition (15.62%) that is not entirely random, as a whole class was unfortunately not available for post-testing.

In summary, the two studies of the efficacy of the apps provide some evidence that they have the potential to offer additional experiences with curriculum materials to a large number of children without the normal staff costs, as one teacher can supervise small groups of children working in the same room. This initial evidence is positive but the studies have limitations and so it is important to evaluate this intervention systematically using an RCT design. In this trial, the apps will be used as additional instruction in maths, as there is no control group that is offered a different intervention for equivalent amounts of time. This design allows for a demonstration of efficacy under the best circumstances, but does not rule out the possibility that any observed impact could be attained if the same amount of extra instruction were to be offered to the control children.

The second contribution is to provide evidence on how pupils react to working on activities on the apps about which they have not yet been taught in the classroom. The intervention aims to complement current teaching practice by offering pupils individually paced additional opportunities to rehearse materials that are part of the Early Years Foundation Stage and of the English National Curriculum for numeracy in Year 1. Outhwaite *et al.* (2018) suggest that it is useful to think of mathematical knowledge as involving four aspects: factual knowledge (for example, number bond combinations and properties of shapes and patterns); procedural knowledge (identifying and applying mathematical procedures)¹; mathematical reasoning (making inferences and deductions from mathematical information); and problem solving (combining and applying different areas of mathematics to solve problems in specific contexts). They further argue that the first two aspects, factual and procedural knowledge, are key to later mathematical learning and ideally should be mastered at an early age. The aim of the *onebillion* apps is to develop these two aspects of mathematical knowledge. However, there is no indication either in the Implementation Manual produced for this project or from existing research on whether the best use of the intervention is as a preparation for learning in the classroom or as a reinforcement of what has been already learned in the classroom. As it is an individually paced programme, it is likely that the synchronisation of classroom instruction and the practice with the app will vary across pupils. Some pupils might use the activities in the apps before the relevant instruction and others afterwards; this trial is not designed to test whether this difference affects the impact of the intervention.

The third contribution of this project is a detailed implementation and process evaluation. In a systematic review of computer assisted learning, Cheung and Slavin (2013) concluded that supplemental computer assisted interventions (CAI) had a positive, though modest, effect (about +0.18) on pupils' educational outcomes. Also, in a systematic review, Haßler *et al.* (2016) examined more specifically the use of tablets in CAI: of the 23 studies that met their quality criteria, 16 reported positive effects (that is, about two-thirds of the studies), 5 reported no effect, and 2 reported negative effects. The authors were unable to identify any contextual factors (for example, institutional characteristics, delivery conditions, sample characteristics) that could explain the difference between observing a positive effect or not, and call for research that seeks to analyse the implementation and process in greater detail. This trial aimed to assess contextual factors that might explain such different results. The apps were designed to be used by pupils with some technical support (see Figure 1, the intervention team's logic model) but relatively little learning support, which is not emphasised in the intervention team's logic model. The apps are essentially treated as an addition to the classroom, which in some instances replaces the teacher: when the topic in the apps has not been taught in the classroom, the teacher in the apps provides the first introduction of the topic to the pupils. The teacher in the apps explains the task and models the response, which the pupils are expected to imitate. Immediate feedback contingent on the pupil's response is viewed as an essential component of this teaching. However, Outhwaite *et al.* (2018) stated that 'it is important to recognize that technology alone will not lead to success; but is dependent on how the technology is integrated into the school environment' (page 11). In this project, a detailed analysis of the implementation considered whether both the material conditions of delivery (space, classroom organisation) and the input from TAs affected the impact of the intervention.

Finally, this trial examined the impact of the apps for the subgroup of pupils entitled to free school meals (FSM). Outhwaite *et al.* (2017) suggested that the intervention is particularly beneficial for pupils with lower levels of

¹Outhwaite *et al.* (2018) use the expression 'conceptual knowledge' here but their definition departs from the original source cited. Holmes and Dowker (2013) define conceptual knowledge as 'the understanding of arithmetic operations and principles, that allows one to make inferences or to relate the different aspects of arithmetic knowledge' (page 252). Thus we refer to the aspects of knowledge targeted by the apps as knowledge of facts and procedural knowledge.

performance at the start of the year. Although there was no control group in the study by Outhwaite *et al.* (2017), the hypothesis that the intervention might have an impact for pupils with lower initial levels of performance was considered worth testing in the present study. In this trial, teachers were asked to nominate for participation in the project pupils whom they considered to be in the bottom half of the class for maths in the first term of Year 1. It is often the case that pupils eligible for FSM are over-represented in the group of lower attaining pupils in maths (Nunes *et al.*, 2017); thus, if the apps are indeed particularly beneficial for pupils with lower levels of performance, the pupils eligible for FSM should show significant progress after using the apps and there should be a significant and positive interaction between membership in the subgroup of pupils eligible for FSM and membership in the intervention group.

In summary, considering the mixed findings of previous studies of CAI and the limitations of those studies related to this specific intervention, as well as the close alignment of the intervention with the aims of the Early Years Foundation Stage and the start of primary school, it seemed timely to carry out a systematic evaluation of the intervention using an RCT design.

Intervention

In this trial, the intervention was delivered over 12 weeks, four times a week (i.e. two hours a week and a total of 24 hours) to Year 1 pupils organised in small groups supervised by a TA. The pupils were selected by their teachers as those in the lower performing half of the class. In the initial sessions the pupils worked with the Maths 3–5 app and in later sessions they progressed to the Maths 4–6 app.

Previous research used a 6- and 13-week intervention (Outhwaite *et al.*, 2017), an 8-week intervention (Pitchford, 2015) and a 12-week intervention (Outhwaite *et al.*, 2018) in three separate studies, and the longer intervention tended to produce a stronger impact. In this trial, the intervention was implemented over a 12-week period (as in the study by Outhwaite *et al.*, 2018, in which the 12-week intervention was reported as effective), from the second half of the Spring term to the beginning of the second half of the Summer term in 2018. This allowed for recruitment to take place during the Autumn Term in 2017 and the pre-test, the implementation of the intervention, and post-test to take place in the Spring and Summer terms of 2018, so that the study could be carried out within a single school year.

Outhwaite *et al.* (2017), on the basis of an intervention study without an appropriate control group, suggested that the intervention has greater impact on lower attaining pupils; they also suggested that it works best as an early intervention. For this reason, the participants in this trial were Year 1 pupils identified by their teachers as performing in maths in the lower half of the class. Prior to the start of recruitment, the intervention and evaluation teams agreed that the participating schools would have a minimum number of 15 pupils in Year 1 in order to participate in the trial. If the classes were any smaller, nomination of the pupils in the lower half would reduce the number of participants in the school and affect the power of the trial. It was also agreed that all pupils would start with the Maths 3–5 app, which was likely to provide them with positive experiences and the opportunity to adapt to the use of iPads if they did not have much experience with them prior to the intervention. As the pupils progress at their own pace, pupils for whom the activities in the Maths 3–5 app were easy would quickly progress to the Maths 4–6 app.

Because the intervention is delivered through iPads, a major question for recruitment was whether each school would have ten iPads in order to implement the intervention, as it was originally envisaged that all TAs would supervise groups of ten pupils. As it became clear during recruitment that this could be an obstacle, the intervention and evaluation teams agreed that schools could sign up for the project without having the required equipment; it would be inappropriate to require schools to buy the iPads before randomisation as they might be assigned to the control group. The intervention team agreed to supplement the number of iPads for schools assigned to the intervention group by lending iPads to schools that did not have sufficient equipment to run the trial.

Prior to randomisation, the intervention team presented their logic model to the evaluation team (see Figure 1). The training of TAs (to be implemented as a one-day face-to-face training, through a support video and an Implementation Manual) is listed in the logic model as a necessary input before the intervention. The crucial elements for the success of the intervention listed in the model are the time on the app and the coverage of the topics as well as the technical support offered by the TAs during the sessions. The outcomes listed in the logic model are improved maths attainment, which is viewed as the primary outcome, as well as improved attention and independence in tackling maths tasks and confidence, which were viewed as secondary outcomes. After discussion, it was decided for pragmatic reasons that no measures of secondary outcomes would be included in the project.

Evaluation objectives

Impact analysis

This project aimed to answer the questions previously specified in the protocol (Appendix 1). The questions regarding the impact of the intervention were:

Primary research question:

- Do the pupils identified by their teachers as struggling with mathematics at the start of Year 1 who use the *onebillion* apps show better performance in Progress Test in Maths 6 (PTM 6) than pupils also identified by their teachers as struggling with mathematics at the start of Year 1 who do not use the apps?

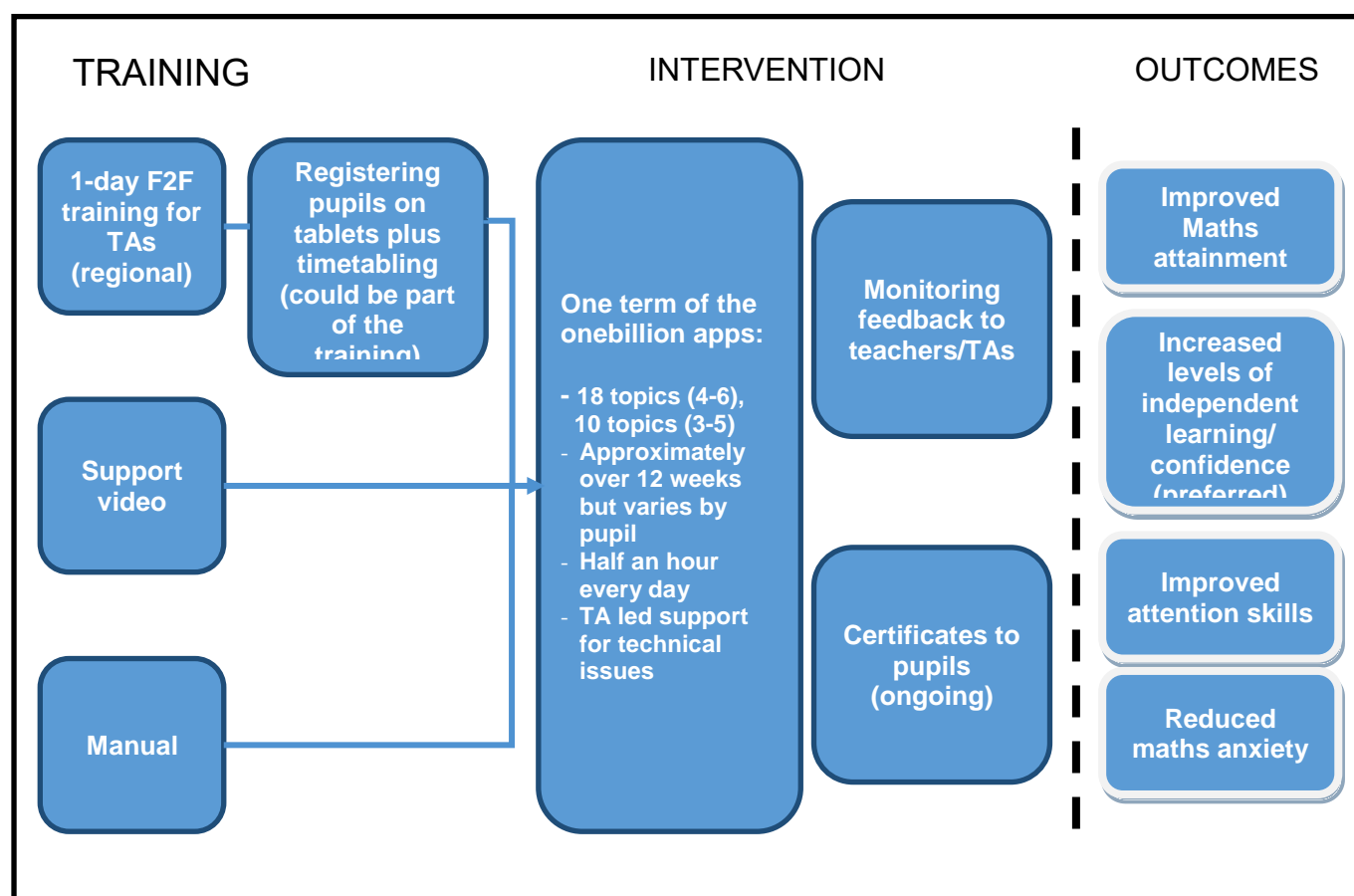
Secondary research question:

- Do pupils entitled to FSM benefit to the same extent as other pupils from using the *onebillion* apps as assessed by PTM 6?

Implementation and process evaluation

The focus of the process evaluation was to assess the fidelity of the implementation *vis-à-vis* the logic model provided by the intervention team as well as to understand the contextual factors that moderate the efficacy of the intervention. The logic model provided by the intervention team is presented in Figure 1.

Figure 1: Logic model produced by the intervention team



The boxes under 'Training' in the figure list the enabling factors, that is the training that TAs receive to provide support to the pupils during the project. The boxes under the heading 'Intervention' list the critical factors for the intervention to succeed: the use of the apps during approximately one term; the frequency of sessions (initially set as daily but it was later acknowledged that four times a week was the expected frequency, as schools might need more flexibility); the defined length of sessions (half an hour). It was critical for the success of the intervention that TAs should be able to

provide technical support so that the pupils could access the activities productively and achieved success in the quizzes. The intervention team would support TAs in this task. Thus, in the logic model, fidelity of implementation is defined in terms of pupils achieving an appropriate dosage and TAs' ability to provide technical support. At the protocol stage (that is, before the start of the intervention), the intervention team defined three levels of compliance for this trial (see Appendix 1): (1) low compliance, defined by participation in up to 30 sessions (62.5% of the sessions in this trial) which is equivalent to 6 full weeks of intervention delivered every day; (2) medium compliance, defined by attendance to between 31 and 40 sessions; and (3) high compliance, defined by attendance to at least 40 sessions (83.3% of sessions).

Fidelity

In order to define what fidelity to treatment meant in this trial, the evaluation team considered the elements included in the intervention team's logic model as well as the recommendations recorded from the training sessions and in the Implementation Manual. This analysis led to the specification of two types of question, one to be addressed at the pupil level and the other to be addressed at the TA level.

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

- Does fidelity to treatment affect the effectiveness of the *onebillion* intervention?

This general research question was divided into more specific questions, related to the measures of fidelity used in the project.

Fidelity in dosage: At the pupil level, the project aimed to address two questions through the implementation and process evaluation:

- Do pupils who have more time on the app show better outcomes, when controlling for pre-test scores?
- Do pupils who progress further in the topics show better outcomes, when controlling for pre-test scores?

Fidelity in the enabling factors: At the TA level, the project aimed to address two questions through the implementation and process evaluation:

- Is the TAs' attendance to training related to outcomes for the pupils, controlling for pre-test results?
- Is the TAs' observance of the implementation conditions specified in the Implementation Manual (for example, place iPads on the tables before the session starts, offer technical support by adjusting the volume on the headphones, and support the pupil navigating the app; remove the pupil's headphones and give the task instructions to the pupil when the pupil fails a quiz repeatedly; page 16) related to outcomes, controlling for pre-test results?

Contextual factors that could moderate the efficacy of the apps in this trial

Business as usual: The main secondary research question to be addressed in the process evaluation is:

- To what extent do control schools use alternative treatments that involve the same content and the same amount of resources as in the intervention schools?

Because the apps are based on the English National Curriculum, it is quite possible that control schools would be teaching the same content as in the apps, but using other resources (for example, Numicon, base ten blocks or other manipulatives) with the pupils who were considered by the teachers to be struggling with maths. However, analysing business as usual was not viewed by the evaluation team as sufficient in this trial. Because some previous studies in the literature showed that CAI using tablets had positive effects, whereas other studies had no measurable effect or even negative effects, implementation and process evaluation were considered crucial in this trial in order to offer a contribution to the understanding of aspects that might contribute to the efficacy of CAI with young children. Thus further consideration was given to the contextual factors that could contribute to the success or to a negative impact of the intervention.

Contextual factors in the implementation process

The importance of contextual factors in the effectiveness of CAI was discussed in a review by Delgado *et al.* (2015), who identified different ways of integrating technology in schools. They distinguished between the use of technology to replace teaching on the one hand, and blended learning on the other hand; in the latter, device-driven and face-to-face teaching are combined in different ways depending on the aims of the school, the capabilities of the teachers and pupils, and institutional factors, such as availability of resources. The use of the apps in this intervention is defined as 100% computer based learning in the intervention team's logic model, as there is no explicit pedagogical role for the teacher. However, the apps offer extra practice on activities that are related to the curriculum; thus teachers naturally have views on how and when the topics are taught and the pupils have learning experiences in the classroom related to the activities in the app.

In this project, we aimed to examine and to measure factors that are part of the context in which the apps were used because the topics in the apps are the same as many topics taught in the classroom. The overall research question was:

- To what extent do the contextual factors (the resources offered by the schools and the TAs' capabilities) contribute to the success of the apps in promoting better outcomes for the pupils who used the apps in comparison to the pupils who did not?

The intervention team did not specify in the logic model any material resources as necessary for success and decided to provide additional iPads for schools that were randomly assigned to the intervention group. However, as discussed by Delgado *et al.* (2015), the location where the learning takes place is a factor related to the quality of the online learning, and for this reason the material conditions of implementation were a factor taken into account in the analysis of contextual factors.

Teacher capabilities were emphasised in the review by Delgado *et al.* (2015) and also in the review of best practice in blended learning by McGee and Reis (2012). It is noted here that the use of the apps in this project would not fall under the concept of blended learning proposed by McGee and Reis, as the apps were introduced as an addition to the classroom learning, without any intent to coordinate the classroom learning with the apps. With respect to the capabilities of the TAs delivering the intervention, for all the TAs involved in this trial, supervising pupils' use of the *onebillion* maths apps was a novel experience, because only schools that had not used the apps were included in the trial (with one exception, see breaches of protocol in the section on 'Trial design'; a school that had experience with the apps was included in the trial and later randomly assigned to the control group). This means that the TAs' learning about how best to carry out their job would be based on the explicit and implicit messages received from the intervention team as well as the experience that they developed while participating in the project. Thus, in order to design the instruments for analysing the process evaluation, the evaluation team analysed the explicit and implicit messages in the training elements (face-to-face sessions, the training videos and the Implementation Manual) because these represented the formal enabling factors that would lead to success in implementation. The TAs' experience would provide them with informal learning on how to succeed in implementation (for definitions of formal and informal learning by newly qualified teachers, see Williams, 2003).

There were explicit as well as implicit messages in the training sessions and in the Implementation Manual that might have been just as important for the efficacy of the intervention as the factors included in the intervention team's logic model. Two types of explicit or implicit messages that were not included in the logic model provided by the intervention team were identified in the Implementation Manual and in the videos used for training.

- (1) **Material conditions for implementation:** Pupils should wear the headphones during the sessions and their iPads should be on the tables with the app open for the pupils to use at the start of the sessions (Implementation Manual, page 16). The explicit message in the logic model is that TAs should offer technical support; the implicit message here is that there are material conditions that are considered 'Top Tips for using the *onebillion* apps' (this is the term used in the Implementation Manual). In the training videos, the pupils were sitting at tables arranged in a horseshoe with the TA at the centre; this spatial arrangement can be considered an implicit message conveyed in the training. Thus the use of headphones by all pupils during the sessions and the seating arrangements were considered in the process evaluation as possibly having a consequence for the implementation.

Further points about the ideal conditions for implementation included both in the Implementation Manual and the videos were: (1) *onebillion* best fits in the timetable outside of the usual maths teaching time; (2) the intervention works best at the start or end of a session, either morning or afternoon; (3) the sessions should each last for 30 minutes (Implementation Manual, pages 16–17).

- (2) **The TAs' role in the intervention:** Although the intervention team assigned only a technical support role to the TAs in their logic model, the Implementation Manual conveys both explicitly and implicitly the message that the TA is not simply a bystander during the sessions. The Implementation Manual includes in the 'Top tips' this description of what the adult should do: *There are 2 things the adult needs to do: (1) give technical support to children, which might mean adjusting the volume or navigating the app; (2) give learning support, by making sure children are attending to the learning and working through the activities at their own pace* (Implementation Manual, page 16). There is no further definition of 'learning support, by making sure children are attending to the learning' in the Implementation Manual. When the intervention team was asked to clarify this further at the face-to-face sessions, TAs were encouraged to use their experience in teaching the concepts in order to support the pupils' learning. The evaluation team was not certain how the interpretation of this guidance would vary among the TAs and sought to clarify it by including questions in the questionnaires that TAs were asked to answer and by taking notes from observations and interviews with TAs during the implementation phase.

The Implementation Manual also states: *If a child does not pass the quiz, try repeating the quiz with them and repeating the questions to them yourself. If they are still challenged by some quiz questions, the child may need to practise the topic further by working through the activities again. This can be independently or with support from the adult running the intervention* (page 15). This point is reinforced later on: *If a child finds a particular activity difficult the adult can take the headphones off and listen to the instructions together. Let the child touch the screen to answer the question themselves* (page 17). It is also raised later: *If a child has attempted a quiz several times and again with your support, it is okay to move the child to the next topic. They can come back to attempt that quiz at another time* (page 20). The explicit message is that, if a pupil finds a quiz challenging, the TA should repeat the instructions and, if the difficulty persists, the pupil should be allowed to move on and to return to this quiz later. The implicit message again is that the TA is not a bystander who only offers technical support to the pupils; the TA should be in the position of recognising whether the pupils are facing difficulties and of intervening in this case. These points were also made in Training Video 2. At the training day in Nottingham, the intervention team was asked by a TA what they should do if a pupil is just getting something wrong; the evaluation team's noted response is: 'If a child is getting it wrong, explain the concept, then put the headsets back on. This is pedagogical support which needs to be logged.' At the Manchester training a similar question was asked. The response written down in the notes by the evaluation team was: 'If the learning hasn't been deep, they will not pass the quiz. Do activities again or go onto the next topic, if the child is not getting the deep learning.' Thus it was mentioned explicitly in the face-to-face training that there is a pedagogical role for the TA, but this was not clarified further than the guidelines in the Implementation Manual.

It was proposed in the protocol that the role of the material conditions of delivery and the interactions between the TA and the pupils would be investigated as part of the implementation and process evaluation (Protocol, page 10) and that it would be necessary to develop instruments to capture these variations in the context of the intervention implementation, which might moderate its efficacy. These instruments are described in detail in the 'Methods' section.

Ethics and trial registration

The trial was designed and conducted according to CONSORT standards and adhered to ethics and data protection regulations from the University of Oxford Ethics Committee and the University of Nottingham. The evaluation team obtained ethical approval for the trial from the University of Oxford Central Research Ethics Committee on 16 November 2017 (Application Approval: ED-CIA-17-014). Opt-out forms were used (see Appendix 1 for parent information and consent forms). When uploading the pupil nominations, TAs were asked to confirm that they had not uploaded information about pupils whose parents had opted out of the trial or UPNs and FSM status of children whose parents opted out of providing this information.

Schools obtained parental consent for participation in the trial; heads of schools agreed to this procedure when they signed the MoU (see Appendix 1 for Memorandum of Understanding [MoU]). If a nominated pupil were to withdraw from

the trial before randomisation, schools were allowed to replace the pupil. No replacement was allowed after randomisation. Parent consent letters and the agreement between the schools and the Nottingham intervention team (MoU) were included in an appendix in the application to the Ethics Committee.

The trial was registered at The International Standard Randomised Controlled Trial Number (ISRCTN) under the number ISRCTN10520083. It can be found in the registration site (link last checked on 5/12/2018):

<http://www.isrctn.com/ISRCTN10520083>

Data protection

The University of Oxford Ethics Committee has a data protection policy that can be found at:

http://researchdata.ox.ac.uk/files/2014/01/Policy_on_the_Management_of_Research_Data_and_Records.pdf

Schools and parents were informed of the reasons for collecting the data and who would have access to the data. After the data collection was completed, the dataset was anonymised; schools and pupils are only identified in the files by identification number; date of birth has been removed and the age information is age in months. No pupils or schools can be identified in the saved dataset, which is kept in the Department of Education server, is only accessible to researchers in the project, and is password protected. A data-sharing agreement between the Oxford and Nottingham University teams was prepared for this project and is included as an appendix to the protocol (Appendix 1). However, the intervention team does not plan to access the data file.

In accordance with the General Data Protection Regulation's (GDPR) article 6, the University's legal basis for processing personal data for this project is as a public interest task. The University's public interest tasks are its 'charitable objects', as laid down in University's Statute I: (i) 'the advancement of learning by teaching and research'; and (ii) 'its dissemination by every means'. Further information on the University's compliance with GDPR can be found at:

<https://researchsupport.admin.ox.ac.uk/gdpr>

Project team

The intervention team was composed of:

- Professor Nicola Pitchford, University of Nottingham. Principal Investigator: Overall project management and leadership
- Dr Maria Neves, University of Nottingham. Programme Manager: Day-to-day management of trial implementation, key correspondent with schools and evaluation team
- Marc Faulder, Burton Joyce Primary School. Technical Specialist: Technical support to schools throughout trial, co-designer and production of training materials and iTunes U module, recruitment through Apple Distinguished Educator Network
- Anthea Gulliford, University of Nottingham. Assistance with recruitment through educational psychology networks
- Professor Geoffrey Wake, University of Nottingham. Assistance with recruitment through regional schools and maths networks

The evaluation team was composed of:

- Professor Terezinha Nunes, Emeritus Professor, University of Oxford: Project Leader. Design of the project. Planning the evaluation. Leading the work on writing the report. Carrying out statistical analyses
- Professor Lars-Erik Malmberg, University of Oxford: Principal Investigator Statistical analyses
- Dr Maria Evangelou, University of Oxford: PI Responsible for obtaining ethical approval. Design of the project and writing
- Professor Peter Bryant, Research Fellow in the Department of Education, University of Oxford: Consultant. Writing parts of the report: leading the revision of the report

- Deborah Evans, University of Oxford: Project Manager and Researcher from 1 June 2018. Recruiting contacts with schools and GL Assessment. Testing: Questionnaires: Observation: Analysis of process variables and contributing to writing
- Rossana Barros, University of Oxford: Project Administrator and Researcher, Project Manager up to the end of May 2018. Statistical analyses
- Susan Baker, University of Oxford: Researcher. Recruiting contacts with schools and GL Assessment. Testing: Questionnaires: Observation: Analysis of process variables and contributing to writing
- David Sanders-Ellis, University of Oxford: Researcher. Recruiting contacts with schools and GL Assessment. Testing: Questionnaires: Observation: Analysis of process variables and contributing to writing
- Chun Yeung Lee, University of Oxford, Postgraduate Student. Data entry and checking

Methods

Trial design

The design is an RCT, with two trial-arms (that is, an intervention and a control group) and a pre- and post-test. The year group selected for participation was Year 1 (5–6 year olds) because the apps were designed for young children; systematic implementation in pre-school was considered but discarded as pre-school settings are less formal and lack of attendance might make implementation at the right level of dosage unreliable. The apps, which were in addition to normal classroom numeracy teaching, were used with small groups of pupils supervised by one adult.

The pre- and post-test were parallel forms of Progress Test in Maths (PTM), provided by GL Assessment. After the pre-test data collection, schools were randomly assigned either to the intervention or to the control group. In order to join the project, heads of schools signed an agreement with the Nottingham team (see MoU in Appendix 1) indicating that they would accept their random assignment. If assigned to the intervention group, they would provide TAs with the necessary conditions for implementing the intervention. The incentive for intervention schools was free access to the apps. If assigned to the control group, they would continue with their usual methods of supporting pupils struggling with maths. As an incentive, control schools were offered the possibility of accessing the apps at the end of the project and using them with the new cohort of Year 1 pupils and £1000 per school after all pupils in the school had been post-tested.

Randomisation was implemented at school level. The school was chosen as the unit of randomisation to avoid the contamination that could take place in a within-school allocation. In schools that had more than one Year 1 class, only one class was randomly chosen to participate in the intervention. However, there were a few departures from the protocol in the selection of classes for participation.

Breaches of protocol in intervention schools: In one school, the teacher of the Year 1 class randomly selected was not willing to participate in the project, so the second class in the school was included instead; in a second school a different class was assigned to participate in the project from the one that had been randomly selected; in a third school, the teacher of the class that had been randomly selected was about to go on maternity leave and so the headteacher decided to designate the other class for participation as the pupils' learning was going to be disrupted by the change in teacher; finally, in a fourth school, pupils from two classes were nominated for participation in the project, although only one class had been randomly selected for participation. All four breaches of the protocol took place before randomisation. In one intervention school, one nominated pupil did not receive the intervention as the teacher inadvertently replaced this pupil with another but the nominated pupil was post-tested rather than the substitute. This breach of the protocol resulted in one pupil in the intervention group not participating in any sessions. In one school the teacher started the intervention with a non-participating pupils and not with the nominated pupil; her oversight was noted when the evaluation team received the records of attendance at intervention sessions and the nominated pupil received the intervention subsequently. None of these breaches of the protocol were serious enough to render the trial less robust.

Breaches of protocol in control schools: One control school acknowledged when answering the middle management questionnaire that they had purchased the apps and some nominated pupils had access to it. This breach of protocol was taken into account in the additional analyses of impact. Although it had been agreed that only schools that had not used the apps previously were eligible for participation, the intervention team recruited one school that had used the apps with a different cohort. The evaluation team had no awareness of this until the phone interviews with the middle management were implemented. This breach of protocol took place at the recruitment stage, before randomisation, and the school was later randomised to the control arm. The staff interviewed reassured the evaluation team that the app had not been used with this cohort of Year 1 pupils. It is reassuring to know that the response rate to the middle management questionnaire was 100% so we can be confident that other control schools did not have access to the app.

Breaches of protocol both in intervention and control schools: When the evaluation team carried out the phone interviews, one teacher indicated that she had not followed the criteria set out by the intervention team for nominating pupils but had rather chosen the pupils 'randomly' (in her words). A check of the middle management questionnaires showed that 20 intervention schools and 17 control schools indicated that they had not excluded pupils with special educational needs or pupils who had English as an additional language when they nominated participants. These problems were noted well after randomisation because the middle management questionnaire was implemented between May and June and randomisation took place in February 2018.

Intervention

TAs in schools assigned to the intervention group were invited to participate in the training for implementation of the intervention. The training included: how to find a suitable time in the daily timetable to administer the intervention; how to prepare the tablets for use (downloading the apps, registering pupils, familiarisation with the apps and their interactive features, technical trouble shooting); advice on offering pedagogical support (limited in this trial); how to record the daily information on participation and quizzes passed, as well as the support that would be offered by the intervention team to the TAs. This information was given at the training events but was also made available in the Implementation Manual and online in a private iTunesU course that was open only to TAs delivering the intervention. The iTunesU course has seven demonstration videos; a printed copy of the Implementation Manual was also given to the TAs. TAs also had access to an online forum where they could share best practice and ask questions to other TAs and to the Nottingham intervention team. The Nottingham team communicated to the schools that they needed to attend the training session for their region. Records of attendance at face-to-face meetings were provided to the evaluation team.

If it were not possible for TAs to attend face-to-face training, they could notify the intervention team and attend a training session at another region. Those schools that found it completely impossible to attend a training session were asked to arrange a phone call and follow the online training. The evaluation team subsequently asked TAs in a questionnaire whether they had accessed the online training.

The intervention was to be implemented for half an hour, 4 days per week, for 12 weeks, in addition to regular mathematics lessons. The choice of a design in which the intervention is offered in addition to regular lessons was based on a systematic review that found a small effect size for this design (+0.18), as well as the finding that, for this intervention, only a group that used it in addition to regular maths lessons showed a statistically significant impact of the intervention. The dosage of the intervention was also determined on the basis of previous research about the impact of the apps.

All children started with the Maths 3-5 app and progressed to the Maths 4-6 app, once they had completed Maths 3-5. Table 2 presents an overview of the design.

Table 2: Overview of the design

| | | |
|-------------------------------|-----------------------------|---|
| Trial type and number of arms | | RCT with two arms, intervention and 'business as usual' control. The intervention group received as an incentive free access to the maths apps and training materials; the control group received a financial incentive of £1000 to cooperate with the tests and access to the apps at the end of the project for the subsequent cohort of Year 1 pupils. |
| Unit of randomisation | | School level randomisation; if the school had more than one Year 1 class, one class was randomly selected for participation. |
| Pre-test | Variable | Attainment in mathematics. |
| | Measure (instrument, scale) | Progress Test in Maths 5 (PTM 5: a standardised test produced by GL Assessment and the University of Nottingham, School of Education). Pre-tests took place before randomisation. |
| Primary outcome | Variable | Attainment in mathematics. |
| | Measure (instrument, scale) | Progress Test in Maths 6 (PTM 6: a standardised test produced by GL Assessment and the University of Nottingham, School of Education). |

Participant selection and recruitment

School recruitment was carried out by the Nottingham intervention team, with support from the evaluation team, across four target regions: (1) East Midlands; (2) West Midlands, (3) Greater Manchester and North West; and (4) South and West Yorkshire.

Seven schools outside these regions were also allowed to join the project (three in Cumbria, three in Oxfordshire, and one in Milton Keynes); the latter four were recruited through contacts of the evaluation team. Local authorities in these regions are listed in Appendix 1. These additions were agreed in order to increase the sample size.

Schools were recruited by means of the following strategies: EEF website; EEF Twitter; University of Nottingham project website; University of Nottingham Twitter; emails to schools through Apple Distinguished Educators (ADE) network and Maths Hubs network; emails to key contacts in local authorities through educational psychology networks. The Oxford evaluation team supported the intervention team with advice regarding all aspects of recruitment, including preparation of materials for inviting schools, the process of registration and the design of the Memorandum of Understanding (MoU), and by emailing schools that had been part of previous projects implemented by the Oxford team. Schools in regions not initially included in those indicated by the Nottingham team, which were approached by either team, were also accepted into the project.

Schools were eligible to participate if they had at least 15 pupils in Year 1, had not used the apps before and had a sufficient number of iPads to implement the intervention with small groups of pupils (that is, one iPad per pupil in each intervention group); 15 was decided because that allowed the teacher to select a reasonably sized number of pupils who were in the lower half for attainment in mathematics. However, the requirement for a sufficient number of iPads was found in the middle of the recruitment process to exclude schools that were eager to participate and, in view of the difficulties the intervention team was facing with recruitment, this requirement was dropped. The intervention team agreed to provide schools with additional iPads when schools that did not have sufficient iPads were assigned to the intervention group at the time of randomisation.

The children were not randomly selected for participation from each class; rather, the teacher in the randomly selected Year 1 class nominated children for participation in the trial. The Nottingham University intervention team provided written instructions to teachers in all schools on how to nominate the pupils for the project: the pupils should be in the lower half of their class for attainment in maths, according to the teacher's assessment, should not have a statement of special educational needs, and should have no difficulty in understanding spoken English.

If a class had 19–20 pupils, 10 pupils should be nominated; if a class had 17–18 pupils, 9 pupils should be nominated; if a class had 15–16 pupils, 8 pupils should be nominated. If the school had more than 14 Year 1 pupils but these were distributed across different classes, all Year 1 pupils were treated as a single cohort and the teachers from the different classes cooperated in the nomination process. The list of nominated pupils was sent to the University of Oxford evaluation team by 18 January 2018, before pre-testing and randomisation; six schools nominated nine pupils and the remaining schools nominated ten pupils.

Data collected at nomination included the pupil's name (which was removed from the dataset and replaced with a project identifier), gender, date of birth, unique pupil identifier (UPN), school (schools' names were later replaced with a number), and eligibility for FSM. Parents could allow their children to participate in the project but withhold the information on UPN and FSM eligibility status, if they wished. After nomination, pre-test results were added to the file.

Outcome measures

The outcome measure chosen for this project was Progress Test in Maths 6 (PTM 6; GL Assessment, 2015), because it is a test of pupils' attainment in topics included in the National Curriculum. An advantage of this assessment is that there is a parallel form, PTM 5, which could be used as a pre-test. The choice of outcome measure was carefully considered, taking into account the characteristics of the intervention and of the measure. A decision to use PTM 5 as the pre-test and PTM 6 as the post-test was made in view of the previous study by Outhwaite *et al.* (2018) which had used the previous versions of these tests. Outhwaite *et al.* (2018) justified their choice at length by documenting that all ten topics in the Maths 3–5 app and three of the topics in the Maths 4–6 app are assessed in PTM 5 which was the pre-test in this project.

A comparison between the 26 items in PTM 6 used at post-test and the topics covered in the intervention showed that only five had not been covered in the apps: these were four items that refer to money and one that refers to reading a digital watch display (see Appendix 2, Table 3). A comparison between the items and the topics in the English National Curriculum showed that all the items in the test (including recognising coins of different denominations and adding their values and measuring time) are part of the statutory requirements but only analogue watches are mentioned explicitly. Thus the measure for the primary outcome was closely matched to the intervention as well as to the curriculum; therefore, the measure should be able to provide clear evidence regarding the efficacy of the intervention without a bias which could be due to the inclusion of items in the test as well as in the apps but which are not part of the statutory requirements. Unfortunately, there is not a number of common items between the two tests to allow for measuring progress, but the tests are appropriate for the investigation of differences between the groups.

The questions in PTM were developed by the Mathematics Assessment Resource Service (MARS) team at the University of Nottingham, working in collaboration with GL Assessment; the intervention team was not part of the MARS group. PTM 5 contains seven items that include different parts; the maximum possible score is 26. PTM 6 contains eight items that include different parts; the maximum possible score is 31. The items cover concepts similar to those taught in the intervention (for example, height, numbers – ordering and recognition – and simple arithmetic, comparisons between sets and objects, spatial relations). There is no time limit but it is estimated that individual administration takes approximately 20–25 minutes. According to technical information from the test provider (GL Assessment, 2015), the tests have good internal consistency (Cronbach's alpha for PTM 5 = 0.87 and for PTM 6 = 0.9).

PTM 6 was standardised on a national sample of 3335 in the UK (including 1132 pupils in Northern Ireland); participating schools were asked to test all the pupils in the classes. PTM 6 was validated by its correlation with the previous version of the test, Progress in Maths (PIM), on a sample of 350 pupils: the correlation was 0.78. PTM was standardised in 2014; no information on external validity by its correlation with key stage tests was available in the Technical Manual in 2018. Gender differences were negligible in the UK standardisation sample; girls scored on average 0.3 standard age scores lower than boys.

The technical information about PTM does not report the means or standard deviations (SD) for the standardisation sample or the correlation between PTM 5, which was used as pre-test in this trial, and PTM 6. Further information about the tests can be obtained from the GL Assessment website: https://www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/?gclid=EAlaIqobChMI3situPfy3glVU6qaCh1T5wDQEAAAYASAAEgJldfD_BwE

Although PTM is designed for administration to whole classes, in view of the pupils' age at pre-test, the evaluation and intervention teams agreed that one-to-one administration would produce more valid results with such young children; for consistency, one-to-one administration was also used at post-test. The adaptations required for this individual administration were checked and approved by the test provider, GL Assessment. The evaluation team recruited testers through a teaching supply agency. The evaluation team trained supply teachers (hereafter referred to as 'testers') to administer the test on a one-to-one basis. Quality control of the administration of the test was based on the observation of a sample of testers (50%) by the evaluation team during administration of the test, both at pre- and post-test; by the end of post-test, all testers had been observed by the evaluation team. Observers provided feedback to testers; only in one case, after one pupil had been tested, the observer needed to reiterate that it is imperative to read the instructions for each item word-for-word, as in the guidance. Testers were not allowed to carry out post-tests in school where they had been employed as supply teachers, as they could inadvertently be told about the school's allocation in the project. Testers were blinded to the schools' allocations; they were asked during the training day to avoid conversations about the test with the staff as well as the pupils. TAs in the schools were reminded by written messages at the time of the scheduling of post-tests that testers should not be made aware of the school's allocation.

Raw scores (that is, the total marks attained by each pupil on the test) were used in the analyses; no transformation of the data was carried out. However, the evaluation team decided not to use one question in the calculation of the raw scores because the answer considered correct by the devisers of the test was actually mathematically incorrect (when the pupil was asked to report the number of rectangles in a pattern, squares were wrongly excluded from the number considered correct); deleting this item made the highest possible score equal to 29 because two marks were awarded for the correct answers to both parts of the item.

Fidelity of scoring at post-test

The Oxford team marked a sample of 280 pupils' post-test booklets from 29 different schools (25% of the sample) blindly with respect to the GL Assessment marking, but implementing the same marking rules. The Oxford scoring was completed by two researchers who worked independently; when their marking differed, they sought consensus by discussing how their markings had been achieved. The consensus score was then entered into the dataset for comparison with the marking by GL Assessment. The correlation between the Oxford and the GL Assessment marking was 0.993 but the same exact score was only observed on 76.8% of the papers. The differences in scores could be attributed mostly to the need to interpret the pupils' handwriting. However, because some items' scores are based on more than one answer, it is not always possible to investigate the source of any discrepancy. Appendix 2 lists the sizes of differences in scores observed. For the purposes of this project, the scores used were those provided by GL Assessment because the evaluation team only marked 25% of the sample and the correlation between the two sets of scores was almost perfect.

Sample size

The aim at the start of the project was to have power to detect an effect size for intervention relative to control equal to 0.18 SD, which was the average effect size observed by Cheung and Slavin (2013) in a meta-analysis of CAI. This seemed reasonable also because a previous evaluation of the *onebillion* apps implemented by the intervention team using a prior version of PTM showed a Cohen's *d* effect size of 0.31 (CI=0.06–0.55) after 12 weeks of implementation of the app, when it was used in addition to normal classroom practice (Outhwaite *et al.*, 2018). Considering that the design in the current project is very similar to the design used in the prior trial and a similar test will be used, the aim of detecting an effect size which is considerably smaller than that observed in the previous study was considered a conservative estimate. The correlation between pre- and post-test used in these calculations was based on information from Outhwaite *et al.* (2018), who reported a pre- to post-test correlation in their previous study of 0.67, and from Worth *et al.* (2015), who found the correlation between PIM 6 and PIM 7 to be 0.75 (calculated for this project). Because these were previous versions of PTM tests, it was decided to calculate the power for this trial using two estimates of the pre- and post-test correlation: $r=0.5$ and $r=0.7$.

Optimal Design software was used to explore the number of schools required for the trial in two different scenarios defined by these two levels of correlation. The calculations relied on the following assumptions: (i) cluster randomised trial with person level outcomes; (ii) pupil outcomes measured at pre-test and at post-test have a correlation of $r=0.7$ at pupil level for one calculation and of $r=0.5$ for the second calculation; (iii) the same correlation for a level 2 analysis; (iv) a within-school sample of ten pupils per school; (v) an intra-class correlation coefficient of 0.15 (estimated by the DfE as the intra-class correlation in mathematics assessments²); (vi) power of 0.80, alpha of 0.05 and a two-tailed significance test. The results of these calculations indicated that 128 schools were required to detect an effect size of 0.2 if the pre- to post-test correlation were 0.5 and 104 schools were required if the pre- to post-test correlation were 0.7; in both scenarios, ten pupils per school were the estimate in the calculations. The EEF decided in agreement with the evaluation and intervention teams to set the target for recruitment to a minimum of 104 schools, allowing for the inclusion of extra schools up to 128.

The actual number recruited was 113 at randomisation; six schools nominated nine pupils and the rest nominated ten. One intervention school did not implement the intervention and did not agree to post-testing, so the number of schools in the analysis was 112. A new power calculation was implemented using PowerUp (Dong and Maynard, 2013) to calculate the minimum detectable effect size (MDES) at the point of analysis. The correlation between the pre- and the post-test was 0.55 and the observed intra-class correlation was 0.2, which is higher than the 0.15 observed for the previous version of this test (Worth *et al.*, 2015). Thus the trial at the point of analysis was powered to detect an effect size of 0.24. Table 3 describes the MDES at the various points in the trial.

²This estimate was in agreement with guidance from the EEF at the time that the protocol was designed. The reference has now been withdrawn from the EEF site.

Table 3: Minimum detectable effect size at different stages

| Stage | N [schools/pupils] (N=intervention; N=control) | Correlation between pre-test and post-test | ICC | Power | Alpha | Minimum detectable effect size (MDES) |
|--|---|---|------|-------|-------|---|
| Protocol | 113 schools/1130 pupils (57 intervention; 56 control) | 0.70 | 0.15 | 80% | 0.05 | 0.19 |
| Randomisation | 113 schools/1124 pupils nominated (57/567 control; 56/557 intervention) | 0.70 | 0.15 | 80% | 0.05 | 0.19 |
| Analysis (i.e. available pre- and post-test) | 112 schools/1089 pupils (56/543 control; 56/546 intervention) | 0.55 | 0.2 | 80% | 0.05 | 0.24 |

Information regarding eligibility for FSM was obtained from the schools, with parental permission. Schools were simply asked whether the nominated pupils were eligible for FSM; because the cohort is a Year 1 cohort, this information corresponds to the same information as for Pupil Premium.

In the sample, there are 286 pupils in 88 schools who are eligible for FSM. According to Rutterford *et al.* (2015), when the number of pupils per cluster is known, and this differs, it is possible to use the mean number of pupils per cluster (3.25 in this sample) to calculate the MDES. The power calculation for the MDES for the pupils in the sample who are eligible for FSM and who can be included in the subgroup analysis, was implemented using the same assumptions as in the main analysis regarding the pre- to post-test correlation and ICC. When the proportion of schools in this subgroup that was assigned to the control and the intervention group was calculated, this turned out to be almost identical to that in the complete sample (51%). The calculation using PowerUp estimated the MDES for the subgroup analysis including only pupils eligible for FSM as 0.29. At the point of analysis, when the observed pre- to post-test correlation and ICC were entered in the calculation, the MDES was estimated to be equal to 0.335.

Randomisation

After the pre-test of the nominated pupils was concluded, the schools were randomly assigned either to the intervention or to the control group by the evaluation team, with an equal allocation of schools to each group. The allocation was completely independent of the intervention team. No blocking procedure was implemented. Random numbers were generated for all schools using SPSS. Schools were ordered by these random numbers in ascending order. Because there was an odd number of schools, the 57 schools that had received the lowest random numbers were allocated to the intervention group and the schools with the highest numbers were allocated to the control group.

The syntax used was:

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

```
EXECUTE.
```

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


Statistical analysis

The statistical analysis was run in SPSS 25 using a mixed models analysis; this is essentially a multilevel regression that analyses whether the addition of a variable to the model explains further variance taking account of differences (variance) between schools. The main analysis was also run using the EEF analytics in R as an additional check; as there was no difference between the outcomes of the analyses, the report presents the results obtained using SPSS.

The impact analysis was carried out using intention to treat; a further comparative sensitivity analysis was run excluding results from one intervention school that did not implement the intervention and from one control school that actually made the apps available to pupils in Year 1.

The pupils were nested in schools. The effects of the intervention were tested in a series of multilevel models with $post_{ij}$ as the outcome, where i = pupils nested in j = schools as follows:

(1) Variance component model $post_{ij} = b_0 + u_{0j} + e_{0ij}$

(2) Fixed effect of pre-test score $post_{ij} = b_0 + b_1pre_{ij} + u_{0j} + e_{0ij}$

(3) Fixed and random effect of pre-test score $post_{ij} = b_0 + b_1pre_{ij} + u_{0j} + u_{1j}pre_{ij} + e_{0ij}$, unstructured level-2 matrix

$$\begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{bmatrix}$$

(4) Fixed effect of pre-test score and intervention $post_{ij} = b_0 + b_1pre_{ij} + b_2intervention_{ij} + u_{0j} + e_{0ij}$

in which $post_{ij}$ is the outcome, b_0 is the grand intercept, b_1 - b_2 are the fixed effects, u_0 - u_1 are random effects and σ^2 is the variance of the residual terms. The maximum likelihood estimator was used to facilitate model comparisons using differences in the -2 log likelihood.

Additional subgroup analyses planned in the protocol assessed the interaction between allocation (intervention vs control) and subgroups defined by eligibility for FSM. These subgroup analyses were run by adding these terms to the models, first as main terms and then as an interaction with the allocation. The details of all the models run are presented in Appendix 3.

Effect sizes were calculated according to $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$, where $s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$, the variance of d is given by $V_d = \frac{n_1+n_2}{n_1n_2} + \frac{d^2}{2(n_1+n_2)}$, and the standard error of d is $SE_d = \sqrt{V_d}$. Hedges' d is then calculated by applying the correction formula $J = 1 - \frac{3}{4df-1}$, where $df = (n_1 + n_2 - 2)$. So, Hedges' g is then $g = J \times d$. The variance of g is $V_g = J^2 \times V_d$ and $SE_g = \sqrt{V_g}$ (Borenstein *et al.*, 2009, equations 4.18 to 4.25).

Compliance with the intervention

The intervention team's logic model listed as factors defining compliance with the intervention pupils' attendance to half-hour sessions, four times a week over 12 weeks, coverage of all topics in both apps, and TA technical support to the pupils. The extent of compliance with the intervention was measured in four ways; the contribution of each of these measures to explaining variance in the outcome measure was tested in the models described in the results section.

- (1) Registers were obtained from the TAs at the end of each week. These recorded whether the pupil was present or not for each of the four sessions per week and provided as the measure the number of sessions that each pupil attended (Model 1).
- (2) TAs also recorded for each pupil the 'stopping point', that is the point the pupil had reached by their final session (Model 2).
- (3) The number of quizzes the pupil had completed by the end of the programme (Model 3).
- (4) The extent to which TAs set up a suitable learning environment and provided the technical support and learning support, as specified in the Implementation Manual's 'Top tips' (Model 4).

The first three variables measure the dosage in different ways and were indicated as critical for success in the intervention team's logic model. These can be treated as interval scales and were entered in the models in order to test whether they explained significant additional variance in the outcome measure.

The fourth variable, compliance with the expectations of support, is an ordinal variable with three levels: low compliance, medium compliance, and high compliance. The Implementation Manual set up guidelines for what the TAs should do before, during, and after each session: (1) ensure that the pupils used the headphones, which generally kept the level of noise down during the session and facilitated concentration; (2) monitor iPads to ensure that they are functioning and provide technological support (for example, sometimes the apps froze and the iPad had to be restarted); (3) monitor the session (for example, check whether the pupils were distracting other pupils and maintain on task behaviour); (4) give support as the pupils work through the activities (for example, if a pupil failed a quiz several times, the TA should remove the pupil's headphones and listen to the instructions with the pupil); (5) ensure that the pupils were attending to learning, a message that received different interpretations by different TAs (some TAs thought that this meant ensuring pupils were on-task and provided no explanations to the pupils; some TAs interpreted this to mean that they could explain the concepts to the pupils; and other TAs interpreted this as meaning that they should provide further explanations about the concepts to the pupils and take on the role of teacher, bringing extra materials to provide explanations, such as manipulatives or the whiteboard). The level of compliance with this guidance was rated as low if the TA only complied with the first three of these guidelines; it was rated as medium if the TA complied with the first four guidelines; it was rated as high if the TA complied with the first four guidelines and explained concepts when the pupil was not succeeding in the activities. This information was collected through observations of sessions (one session per TA in 30 schools), and thus represents a reduced sample.

Implementation and process evaluation

The analysis of implementation and process evaluation is organised in four sections. The first section describes how the information about the enabling factors – that is, preparation for and participation in training – was collected. The second section describes how the information about fidelity to messages explicit or implicit in the training, videos, and Implementation Manual was collected. The third section describes contextual factors and covers the material conditions of implementation, who delivered the intervention and the TAs' role during the sessions. The final section describes business as usual in the control schools and considers whether there is evidence that the effect of the intervention could have been washed out due to the use of other interventions with similar content. All the instruments used for this data collection are presented in Appendix 4.

Enabling factors

The logic model produced by the intervention team placed attendance at training and being prepared for it by bringing an iPad that could be used at the training as an enabling factor. The training of the TAs was carried out in two different ways:

- (1) **Face-to-face training:** 42 members of staff from 37 schools (67.27%) attended training sessions in Manchester, Birmingham, or Nottingham; 20 other members of staff from these 37 schools, who had not attended training, also delivered the intervention. Another face-to-face course, planned to be held in Sheffield, had to be abandoned because of extremely bad weather on the day. One school (1.82%) had staff trained individually by a member of the intervention team. All schools that were represented at the face-to-face training received the training materials.
- (2) **Online training:** Members of staff from 17 schools (30.91%) were unable to attend the face-to-face training. These schools were sent the training materials and asked to watch the online training videos.

The evaluation team measured the success and the effectiveness of the training of TAs and teachers in three ways:

- (1) attendance at the face-to-face training sessions (attendance is necessary in order to know which TAs had participated in the training and keep track of response rate);
- (2) a post face-to-face training questionnaire or a post online training questionnaire – both were answered online and both dealt with the TAs' reaction to the training and their confidence about showing the pupils how to use the *one-billion* app; and
- (3) interviews with a sample of 30 TAs (54%).

An overview of how the information was collected is presented in Table 4.

The aim of these measures was not to compare the forms of training but to find out what percentage of the TAs felt that they were appropriately prepared for implementing the intervention. The post-training questionnaire provided information on the TAs' immediate perception and the interview provided information on how they felt about the training after they had implemented the intervention for more than half the time.

The evaluation team carried out direct observations of the three training sessions, which sought to complement information about the expectations of the intervention team for the implementation of the intervention, particularly when the attendees asked clarifying questions. After the training, TAs completed an online questionnaire. They were reassured that their answers were confidential and that the questionnaires would be anonymised later. The full post-training questionnaire is presented in Appendix 4. Due to the cancellation of one face-to-face training session, not all TAs participated in this mode of training.

Table 4: Overview of the instruments used to measure the implementation of the enabling factors

| Enabling factor | How the information was collected | What was included as part of implementation and process evaluation |
|--|--|--|
| Training completion | Attendance log from the face-to-face sessions taken by the intervention team and provided to the evaluation team TA questionnaire | Percentage attending face to face training Percentage accessing online resources independently |
| Perceived value of the training – face-to-face | TA post training questionnaire; interview questions following observation | TAs evaluated the training and the extent they felt prepared for the apps |
| Perceived value of the training – online | TA post self-training questionnaire | TAs' perceived value of online resources and the Implementation Manual for training and extent they felt prepared for the apps |

Implementation fidelity factors

The methods that were used to measure the fidelity and the context of implementation are divided in this report into two groups; the first relates to fidelity of implementation while the second describes the contextual factors that could moderate the effect of the intervention. Two instruments provided information on fidelity: (1) TAs provided session logs for measuring each pupil's time and success on the apps; (2) observation of an implementation session in a sample of 30 schools. An overview of the information collected for fidelity is presented in Table 5.

Table 5: Measures of implementation fidelity

| Implementation fidelity factor | How the information was collected | What was included as part of implementation and process evaluation |
|---|-----------------------------------|--|
| Dosage measured by time on the apps | Session logs provided by TAs | Number of sessions each pupil attended |
| Dosage measured by progress on the apps | Log of last session | How far each pupil progressed on the app; how many quizzes the pupil succeeded with |
| TA implementation fidelity | Session observations | TA implementation: low, medium or high fidelity as outlined in the Implementation Manual |

TAs' session logs: The intervention team provided TAs with a format for recording each pupil's attendance at each session; the logs included the pupils' identification, the date and time of the session; information on whether each pupil had needed technical or pedagogical assistance during the session; quizzes on which the pupil succeeded. The information on attendance was condensed into a single number to represent the time on the apps. The number of quizzes passed and the stopping point (that is, the last activity concluded by the pupil) were also treated as variables that described how far the pupil had progressed on the apps.

Session observations: The protocol had originally indicated that the evaluation team would observe ten sessions in order to describe the context of implementation. Observation grids were prepared (see Appendix 4) which guided how the samples of behaviour would be collected. The observer would initially ask the TA to indicate a pupil who was quite comfortable with the apps, one who was finding the apps rather challenging, and one who was about average in how he/she was getting on. The first 15 minutes of the observation was dedicated to monitoring each of these pupils for 5 minutes each and noting on the observation grid difficulties and successes, requests for support, off-task behaviour and interaction with other pupils. The second part of the observation focused on the TA: the observer noted how the TA managed the group (for example, whether the TA was watching the pupils and identifying those who needed help; whether the guidelines in the Implementation Manual were followed when a pupil failed a quiz repeatedly by removing the earphones and explaining concepts). Observers also noted the use of headphones (or not), whether the level of noise interfered with the pupils' concentration (particularly when no headphones were used), where the session took place, and the seating arrangement (some TAs had no space to position themselves behind or next to the pupils in order to look at the iPad with the pupil, whereas others were able to see all the pupils at the same time and also their iPads; some TAs sat away from the pupils and were engaged in other tasks).

The first five observation sessions indicated such variability in implementation that the evaluation team, in agreement with the EEF, decided to increase the sample of observations. Members of the evaluation team directly observed a *onebillion* teaching session in 30 out of the 57 schools that had been allocated to the intervention group. The selection of schools for observation was intentional rather than random. The post-training questionnaires allowed the evaluation team to identify three levels of previous use of iPads with pupils in schools, combined with TA level of confidence with iPads (low, middle and high) and to identify different forms of training for using the app (face-to-face training versus online video resources only). These elements were combined into a table from which the schools were chosen. The schools were also selected to illustrate a variation in the proportion of nominated pupils eligible for FSM with different levels of previous familiarity with iPads in school. The purpose of the 30-minute observations was to record information regarding TAs' compliance with the guidance provided by the intervention team and to provide information about the context of implementation. Appendix 5 presents the schools' locations and dates of observations.

Data was analysed by comparing the TAs' behaviour during the sessions with the guidelines from the Implementation Manual (listed as 'Top Tips' for successful implementation).

Contextual factors

Information about the contextual factors that could affect the implementation was collected using three instruments: (1) week 9 online questionnaire for TAs; (2) observation (described above) and (3) interviews with TAs after the observations. Table 6 provides an overview of these instruments and of the information collected.

Table 6: Instruments used to collect information about contextual factors that could contribute to the effect of the intervention

| Contextual factor | How the information was collected | What was included as part of implementation and process evaluation |
|---|--|--|
| Material conditions | TA week 9 questionnaire; observation | Access to iPads and headphones; type of intervention space; schedule for sessions; apps and headphones functioning |
| Fit with school planning | TA week 9 questionnaire | Location; preparation time; fit with timetable |
| TAs' perceived role in the implementation | TA week 9 questionnaire and interviews | Open answers categorised by the evaluation team |
| TAs' observed style during implementation | Observation | TAs' behaviour during the session |
| Pupils' enjoyment and engagement | Observation; TA week 9 questionnaire | Pupils' behaviour: reaction to success/failure in quizzes; off-task behaviour, use of headphones |

Week 9 online questionnaire: In week 9 of the intervention, all the TAs were asked to complete an online questionnaire about the implementation of the programme in order to provide information on aspects of delivery, including: group size; ease of use of the materials and materials conditions under which the intervention was delivered; how TAs perceived their role; how often they gave pedagogical or technological support and what this support entailed. As noted earlier, all TAs were implementing this intervention for the first time; the questionnaire was proposed to them in week 9 to maximise the chances of TAs having developed some expertise and insight into the use of the apps. The questionnaire contained both closed and open questions (their perception of their role in the intervention, what they did in the sessions, what they found most challenging, and what they found the best thing about the programme; their 'Top tips' for another TA who wished to use the apps in the future). Data from the open-ended questions was analysed qualitatively into categories to provide a description of how the TAs perceived the intervention. The categories formed to describe the TAs' perceptions of their role in the intervention were validated by examining the correlation between these categories with the observation data of the TAs' compliance with the guidelines. The variable based on the TAs' perception of their role was used in a regression analysis that investigated whether the TAs' perceived role was related to the pupils' outcomes.

Post-observation interviews with TAs: The member of the evaluation team who carried out the observation in a particular school also interviewed the TA. The interview took place after the observation for the most part, but sometimes before the observation if the TA's timetable made that necessary. The interview included questions about the effectiveness of the training the TA received, any issues arising with running the intervention, the pupils' perceived enjoyment of the app, technical support commonly given, pedagogical support given, confidence of the TA with daily and weekly tasks, and any further comments that TAs wanted to make about the intervention.

Business as usual

Intervention effectiveness is measured against progress in control schools where ‘business as usual’ is expected to take place. This means that effective changes in practice in control schools could make the intervention look less effective than it really is. One possible such change is contamination, that is the use of aspects of or of the whole of the intervention in the control schools. In this trial, a second possibility is that, even though the apps were to be used in addition to normal mathematics lessons, the intervention schools could have used them instead of mathematics lessons that other pupils were receiving (for example, pre-teaching of topics; post-teaching after a pupil found a concept difficult during the lesson). In order to assess the likelihood that these events could have distorted the results of the trial, a middle management questionnaire as well as phone interviews with teachers and middle management staff were used to collect relevant information in both control and intervention schools. An overview of the items in these instruments is presented in Table 7.

Table 7: Instruments used to collect information about what business as usual meant in this trial

| Business as usual | How the information was collected | What was included as part of implementation and process evaluation |
|--|--|---|
| Business as usual for this trial in control and intervention schools | Phone interviews with teachers in control schools and intervention schools Online questionnaire for teachers in control schools | Year 1 use of iPads School Improvement Plan: IT and maths priorities Maths interventions apart from <i>onebillion</i> used with Year 1 pupils requiring support in maths (intervention schools) Maths interventions used with Year 1 pupils requiring support in maths (control schools); confirmation of non-exposure to <i>onebillion</i> apps |
| Implementation of selection criteria in intervention and control schools | Online questionnaire for teachers and TAs in intervention and control schools | Selection criteria of pupils – adherence to guidelines Resources – iPads, headphones, and expenditure on IT hardware and software |

Middle management telephone interviews – intervention and control groups: The evaluation team conducted 20 telephone interviews with the teachers (ten from control schools and ten from intervention schools) who had been nominated at the recruitment stage as the link teacher for the project. The ten control schools were chosen randomly; the ten intervention schools were chosen randomly from the group of schools that had not been visited for observations. The reason for this strategy was to maximise the number of schools that the evaluation team contacted during the project. Questions for middle management staff from the intervention schools included clarifications regarding: (1) whether the pupils were using the apps instead of normal maths lessons; (2) whether the nominated pupils were receiving maths interventions offered to other Year 1 pupils who needed additional support; and (3) whether the school was intending to use the apps in the next academic year. Questions for middle management staff from control schools included clarification about: (1) the use of iPads in Year 1; (2) maths interventions implemented for Year 1 pupils needing support in maths; and (3) whether the nominated pupils had had access to the apps. Both groups were asked to explain how they had selected the pupils for the project at the nomination phase and were asked about their priorities for maths and IT and their School Improvement Plan. The response rate was 100%.

Middle management/link teacher questionnaires – intervention and control groups: An online questionnaire was completed by a middle management member of staff in the intervention schools to provide information about the costs and the fit of the intervention with the school’s aims and schedules. The questionnaire included questions about previous use of IT in the school in order to describe the context in which the intervention took place. A middle management member of staff in the control schools was also asked to complete a questionnaire. They were asked to describe what interventions had been used with the pupils nominated for participation in the project, to explain the content and duration of these, and to confirm that the *onebillion* apps had not been used. Response rate was 100% in control schools and 96% in intervention schools.

Except for the session logs, which were provided by TAs to the intervention team, all implementation and process evaluation data was collected and analysed by the evaluation team. The analyses of the information extracted from the instruments were completed blindly to the outcome measure; with the exception of the data from the log, the extraction of the information from the implementation and process evaluation was carried out before the data from the post-test had been received. Because the processing of the logs in order to convert this information into numerical information matched to the pupils was very time consuming, these variables were extracted from the logs at the end of September 2018, when the evaluation team already had the post-test scores. The file with the dosage data extracted from the logs was created independently of the post-test scores by a researcher who was completely blinded to the information about the pupils' post-test scores and the TAs' questionnaires and observations.

The fidelity measures described in Table 5 were entered into the models used to assess the effectiveness of the intervention. In order to investigate whether differences in implementation of the interventions contributed to the impact of the intervention, the process variables from Table 5 were explored in the following models (see Appendix 3):

- (1) $\text{post}_{ij} = b_0 + b_1\text{pre}_{ij} + b_2\text{sessions}_j + b_4\text{sessions}_{ij} \times \text{pre}_{ij} + u_{0j} + e_{0ij}$ (where 'sessions' indicates the number of sessions each pupil attended);
- (2) $\text{post}_{ij} = b_0 + b_1\text{pre}_{ij} + b_2\text{quizz}_j + b_4\text{quizz}_{ij} \times \text{pre}_{ij} + u_{0j} + e_{0ij}$ (where 'quizz' indicates the number of quizzes each pupil passed);
- (3) $\text{post}_{ij} = b_0 + b_1\text{pre}_{ij} + b_2\text{stopp}_j + b_4\text{stopp}_{ij} \times \text{pre}_{ij} + u_{0j} + e_{0ij}$ (where 'stopp' indicates the highest activity each pupil reached by the end of the last session);
- (4) $\text{post}_{ij} = b_0 + b_1\text{pre}_{ij} + b_2\text{T_met_exp}_j + b_4\text{T_met_exp}_{ij} \times \text{pre}_{ij} + u_{0j} + e_{0ij}$ (where 'T_met_exp' indicates the degree to which the TA met the expectations set out in the Implementation Manual).

Costs

Information for the cost analysis that refers to implementation costs was obtained through the TAs' week 9 questionnaire (see Appendix 4 for the implementation and process evaluation tools). Information regarding costs of equipment and the apps was obtained from the Apple website.

Timeline

Details of the timeline for the implementation of the project are presented in Table 8. Changes to the initial timeline were necessary due to difficulties in meeting the target number of schools for recruitment by the expected date. Recruitment was extended by three months and the target number was achieved. Table 8 shows the actual implementation periods.

Table 8: Timeline for the implementation of the project

| Date | Activity |
|--|---|
| Sept 2017 – Jan 2018 | Recruitment of schools by intervention team; ethical approval; nomination of pupils for the trial and consent for participation |
| 10/01/2018 – 16/01/2018 | Tester training for pre-test delivery, by evaluation team (Nottingham, Birmingham, Warrington) |
| Jan – Feb 2018 | Pre-testing of nominated pupils; catching up pre-tests for absent pupils; quality assurance of pre-testing by evaluation team |
| 15/02/2018 | Randomisation of schools |
| 26/02/2018 – Nottingham 27/02/2018 – Birmingham 28/02/2018 – Manchester 01/03/2018 – Sheffield (cancelled due to extreme weather) | Training dates for TAs delivered by intervention team |
| 05/03/2018 – 15/06/2018 | Intervention period and collection of session logs; 30 observations and interviews with TAs by evaluation team; TA and middle management questionnaires collected; phone interviews with middle management and teachers |
| 11/06/2018 – Birmingham 12/06/2018 – Manchester 14/06/2018 – Nottingham | Tester training by evaluation team for post-test delivery; continued collection of session logs |
| 18/06/2018 – 13/07/2018 16/07/2018 – 20/07/2018 | Post-tests administered; quality assurance through observation of testers by evaluation team; catch up post-tests for absent pupils; continued collection of session logs |
| 02/07/2018 – 31/07/2018 | Second marking of 25% of post-tests by evaluation team; all tests posted to GL Assessment |
| Aug – Sept 2018 | Analysis of process evaluation data |
| Sept – Oct 2018 | Writing up process evaluation; merging of process evaluation data with files from GL Assessment |
| Nov – Dec 2018 | Data analysis and report preparation |
| Jan – July 2019 | Review of report; preparation of final report; dissemination at conferences and publication of papers |

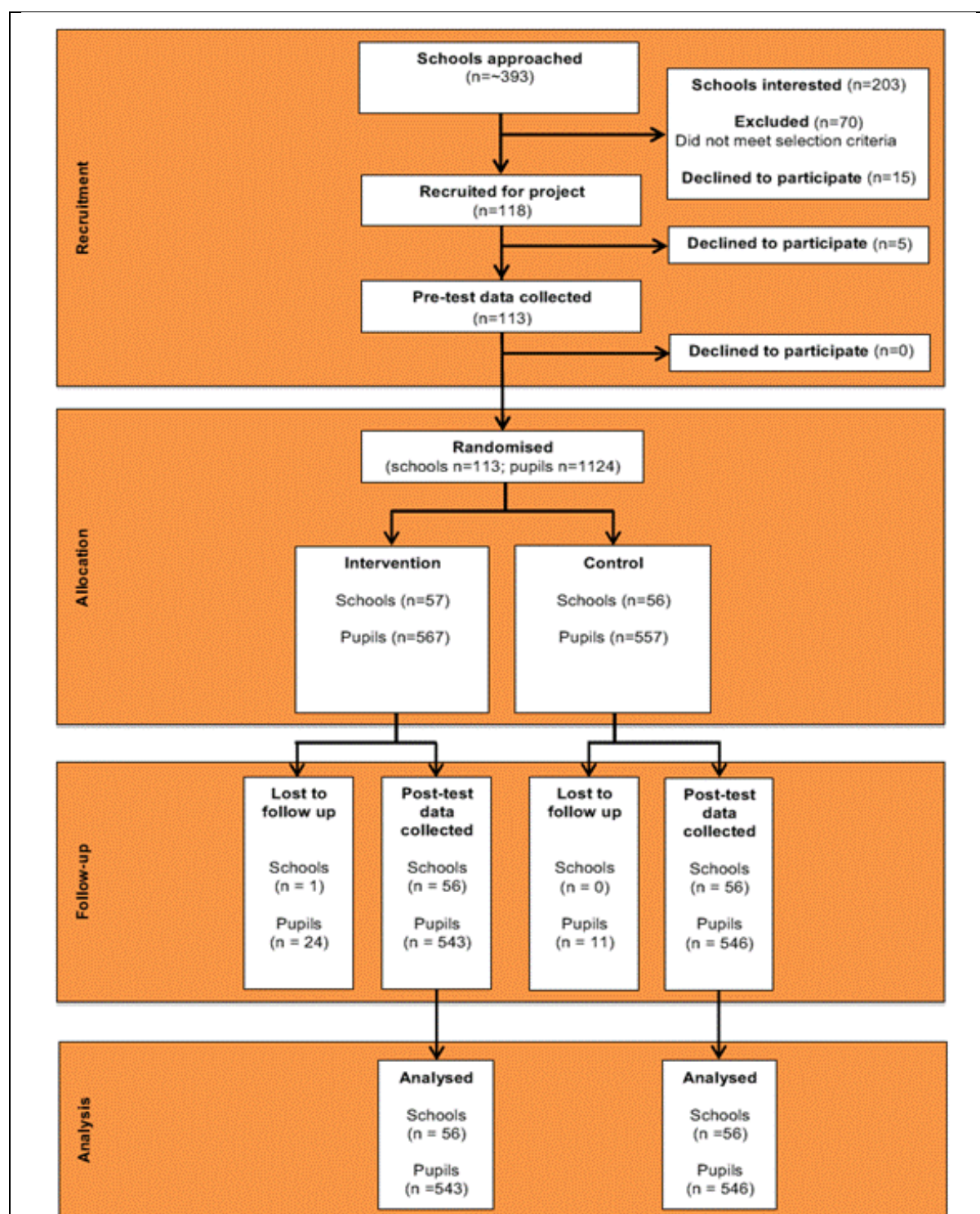
Impact evaluation

Participant flow including losses and exclusions

Recruitment was carried out by the intervention team with the support of the evaluation team. Expressions of interest were received by the intervention team. Once the school was considered eligible for participation, the intervention team informed the evaluation team of the school's involvement, and the evaluation team set up the procedures for nomination of participants and pre-test dates. As indicated previously, pupils were nominated for participation if their teachers evaluated them as showing mathematics attainment in the lower half of the class.

At the end of the recruitment period, 118 schools had returned MoUs and proceeded with the nomination of pupils; five schools withdrew before the pre-test and randomisation. No schools withdrew between pre-test and randomisation. Two schools that had been assigned to the intervention group did not implement the intervention; in one case, this was due to a change in headteacher and in the other it was due to the headteacher going on sick leave; one of these agreed to post-testing whereas the other did not, resulting in the loss of ten pupils to post-test. One pupil in one intervention school asked to stop participation because they did not like the sound used for feedback when an answer was wrong. Absences at the time of post-test were due to entirely to random reasons, mainly relocation and illness; the evaluation team made every effort to return to the schools and test absent pupils, but this was not always possible. Of the 58 pupils absent on their original post-test date, 56 pupils were revisited and successfully post-tested at a later date. The numerical information about the participants is summarised in Figure 2.

Figure 2: Participant flow diagram showing numbers from recruitment to post-test



Attrition

The total number of pupils lost to post-test was 35 out of 1124 randomised pupils, giving a 3.11% attrition rate. One school dropped out, leading to the loss of ten pupils; 23 pupils moved school; one pupil was on extended sick leave; and one pupil was absent on post-testing day. It is noted that one of the 56 intervention schools did not implement the intervention but agreed to implement the post-test.

There were no changes in the number of pupils eligible for FSM after pre-test.

The MDES at the point of analysis was 0.24 (see Table 3) for all pupils and 0.335 for pupils eligible for FSM. In line with recommendations by Schulz and Grimes (2002), the evaluation team considered the level of attrition as too small to warrant further investigation.

Pupil and school characteristics

Table 9: Pupil and school characteristics

| Variable | Intervention group (N=57) | | Control group (N=56) | |
|--|---------------------------|------------------------|----------------------|---|
| School-level (categorical) | n/N (missing) | Percentage | n/N (missing) | Percentage |
| School type ¹ | | | | |
| Academy | 19/56 (0) | 33.9% | 23/56 (0) | 41.1% |
| LA school | 37/56 (0) | 66.1% | 33/56 (0) | 58.9% |
| Ofsted rating ² | | | | |
| Outstanding | 10/56 (0) | 17.8% | 14/56 (0) | 25% |
| Good | 41/56 (0) | 73.2% | 33/56 (0) | 58.9% |
| Requires Improvement | 3/56 (0) | 5.4% | 9/56 (0) | 16.9% |
| Inadequate | 2/56 (0) | 3.6% | 0/56 (0) | 0% |
| Location ³ | | | | |
| Urban (city, town, and conurbation) | 51/56 (0) | 91.1% | 48/56 (0) | 85.7% |
| Rural (hamlet, village, town, and fringe) | 5/56 (0) | 8.9% | 8/56 (0) | 14.3% |
| School-level (continuous) | n (missing) | Mean | n (missing) | Mean |
| No. of pupils per school ¹ | 56 (0) | 296 | 56 (0) | 265 |
| No. eligible for Pupil Premium ¹ N (%) | 56 (0) | 85 (29%) | 56 (0) | 73 (27%) |
| Pupil Premium allocation 2018/19 ¹ (£) | 56 (0) | 112,327 | 56 (0) | 96,501 |
| Pupil-level (categorical) | n/N (missing) | Percentage | n/N (missing) | Percentage |
| Eligible for FSM | 134 / 543 (5) | 24.9 ⁴ | 137 / 546 (2) | 25.2 ⁴ |
| Pupil-level (continuous) | n (missing) | Mean 95% (CI) | n (missing) | Mean (95% CI) [effect size] |
| Pre-test score ⁴ | 567 (0) | 13.32 (12.96–13.69) | 557 (0) | 13.53 (13.23–13.86) –0.05 [–0.29–0.69] |
| Age in months at post-test | 543 (24) | 74.76 (74.47–75.05) | 546 (11) | 74.63 (74.34–74.92) 0.04 [–0.54–0.28] |
| ¹ Source: https://www.gov.uk/government/publications/pupil-premium-conditions-of-grant-2018-to-2019 , accessed 08/11/2018 ² Source: https://reports.ofsted.gov.uk/ , accessed 09/2018 ³ Source: https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2018 , accessed 08/11/2018 ⁴ Percentage of valid cases; seven parents did not agree to the release of this data | | | | |

Table 9 above presents a summary of the school and pupil characteristics after randomisation. It was considered that there was no need for blocking schools for randomisation because only trials with small numbers of units for randomisation require blocking (Kang *et al.*, 2008, suggest that trials with fewer than 100 units must consider blocking; Suresh, 2011, also indicates that blocking is only needed for small sample sizes). Because the use of blocks requires that blocks be taken into account in the analysis (Kahan and Morris, 2012), simple randomisation makes the analysis in this project more comparable to other analyses of interventions to improve mathematical attainment. As a check on whether there were statistically differences between schools and between pupils across the groups, the information presented in Table 9 was subjected to a comparative analysis.

The association between group allocation and other nominal variables (school type and school location) was investigated by means of Chi-square tests. These tests did not show a significant association between group allocation and school type (Chi-square=1.95; df=2; p=0.377) or between group allocation and school location (Chi-square=0.844; df=1; p=0.358).

Previous Ofsted rating is a measure that can be treated as ordinal, from inadequate to outstanding, but must be treated with caution. Schools' evaluations are implemented at different years and standards are assessed with reference to different guidelines. An analysis that treated Ofsted rating as an ordinal measure based on the Median Test did not show a significant association between group allocation and Ofsted rating (p=0.46); a more conservative analysis, which treated the categories as a nominal measure, converged with the Median Test and did not show an association between group allocation and Ofsted rating (Chi-square=6.74; df=3; p=0.08).

The number of pupils in the school was treated as a measure of school size and the amount of Pupil Premium allocation was treated as a measure of pupil deprivation in the school. Independent t-tests were used to compare the schools in the two groups on these measures. Neither comparison yielded statistically significant results; for school size, t=1.41; df=111; p=0.16; for deprivation, t=1.03; df=111; p=0.307).

Comparisons at the pupil level were based on pre-test scores and age at post-test. Although schools had been asked to nominate pupils whose achievement in mathematics was at the lower half for the class in maths, the distribution of scores in PTM 5 at pre-test is described by a normal curve, with a mean of 13.43, SD of 4.2, and z for skewness = 1.19. The median and mode were both 13; minimum score was 2 and the maximum was 24 (the highest possible score for the pre-test was 26). Cronbach's alpha was used as a measure of internal consistency. This index was equal to 0.69, which is considered adequate for a test; one item showed a negative correlation with the total and, if this item were removed, Cronbach's alpha would rise to 0.7; no other items showed a negative correlation with the total. Thus the test had good psychometric properties. A t-test for independent samples showed that the differences between the groups did not reach the conventional levels of statistical significance: t=0.18; df=1122; p=0.42. The effect size (Hedges' g) for the difference between the intervention and control group was g=-0.05 [-0.07, 0.17]. Appendix 2 presents the distribution for the pre-test scores.

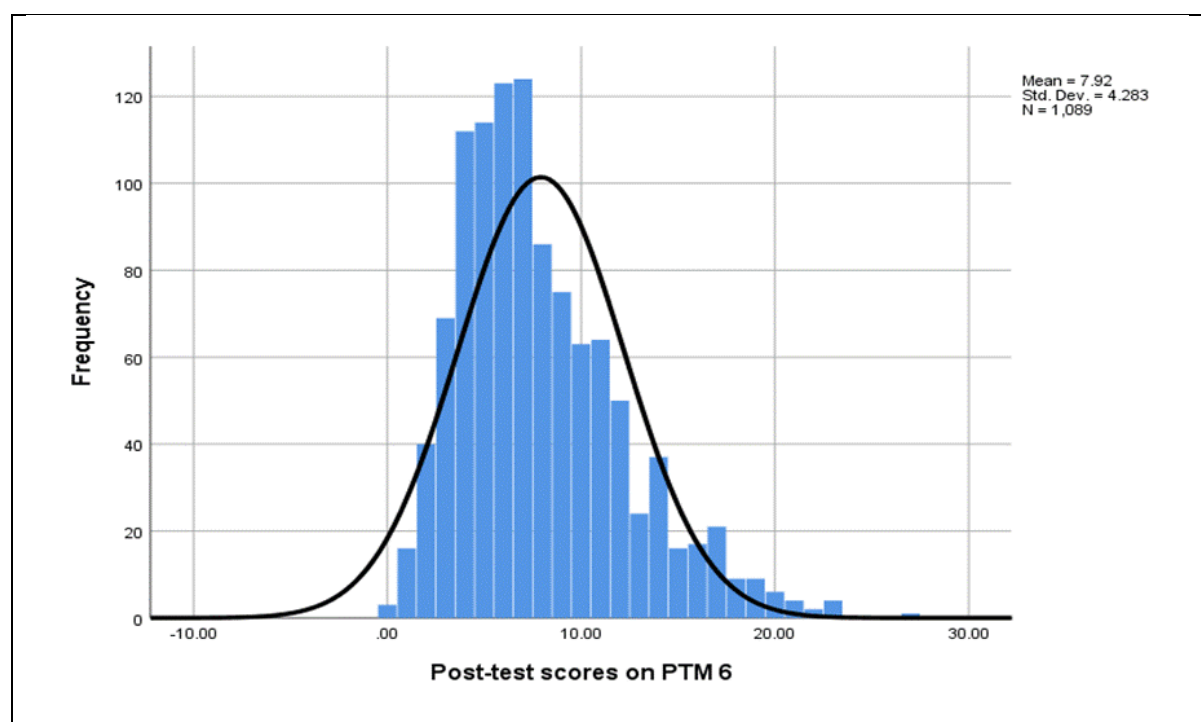
The range of ages in months at post-test was from 69 to 82 months. The mean age at post-test was 6 years 2 months (74.69 months); the SD=3.42 months. The distribution of ages was approximately normal; the z for skewness was 0.239 for age at post-test, thus not significant. The mean age for the pupils in the intervention group was 74.63 months and for the control group was 74.76; an independent samples t-test showed that this difference was not significant (t=0.63; df=1087; p=0.53; Cohen's d effect size = 0.04).

Outcomes and analysis

This evaluation used PTM 6 as a measure of attainment in mathematics. As described earlier, the items in PTM 6 are based on the recommendations for the National Curriculum in Year 1, at the start of Key Stage 1. It was indicated earlier that responses to one item were excluded from the total score; all the analyses described here used the scores based on all items except for this one (item 8a in the test). All the analyses were checked and it was found that the exclusion of this item does not affect the outcomes of any of the analyses.

Considering that the distribution of scores at pre-test on PTM 5 was normal, it was expected that PTM 6 would also show a normal distribution. However, this turned out not to be the case, and the test was positively skewed, suggesting that it was a difficult measure for the participants and thus the test may not offer good discrimination at the lower end of attainment. The mean score was 7.92; the SD was 4.28; the range was from 0–27 (the maximum possible score was 29); the z for skewness was 12.86; the mean and median were equal to 7. Cronbach's alpha was 0.75, which indicates a good level of internal consistency; no items showed a negative correlation with the total. Figure 3 presents the histogram for the distribution of post-test scores. In spite of the skewness, the three measures of central tendency converge reasonably well; parametric tests are considered robust enough to deal with this deviation from normality, so no transformation was implemented to the raw data (Field, 2009).

Figure 3: Distribution of scores at post-test on PTM 6 (N=1089)



The correlation between the pre-test (PTM 5) and the post-test (PTM 6) was 0.55; the post-test correlation with age in months was 0.13. The intra-cluster correlation for both the pre- and the post-test was 0.20. The observed correlation between pre- and post-test was lower than expected; this was taken into account in the calculation of the MDES at the point of analysis (see Table 3).

Primary analysis of impact

As specified in the 'Methods' section of this report, the effects of the intervention were tested in a series of multilevel models that used variance components analyses in linear regressions. There were 1124 pupils randomised at the school level into a control group (N=557) or an intervention group (N= 567). Of these, 1089 pupils ($n_{\text{int}}=543$, $n_{\text{cntrl}}=546$) produced scores on the outcome measure. The low level of attrition (3.11%), which was due to random factors, does not raise cause for concern regarding the validity of the trial. The details of the models are presented in Appendix 3. Tables 10a and 10b summarise the results of the main impact analyses. The observed raw mean for the intervention group was larger than the mean for the control group; this difference was statistically significant according to the multilevel model (see Appendix 3, Table 1, Model 4), after taking into account the differences between schools (Appendix 3, Table 1, Model 1) and the effect of pre-test (Appendix 3, Table 1, Model 2).

Table 10a: Raw means, confidence intervals (CI) and effect size for the outcome measure at post-test

| | Raw means | | | | Effect size | | |
|-----------|--------------------|----------------------|----------------|----------------------|----------------------------|-----------------------|-----------------------------|
| | Intervention group | | Control group | | | | |
| Outcome | N (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (int, contr) | Hedges' g (95% CI) | p value |
| Post-test | 543 (24) | 8.43 (8.06, 8.80) | 546 (11) | 7.41 (7.06, 7.76) | 1089 (543, 546) | 0.24 (0.12, 0.36) | t (1087)=3.97 p=0.000078 |

The adjusted mean difference presented in Table 10b was taken from the multilevel model (Model 4), that is estimated differences in the outcome controlling for the pre-test.

Table 10b: Effect size estimation

| Outcome | Unadjusted differences in means | Adjusted differences in means | Intervention group | | Control group | | Pooled variance |
|-----------|---------------------------------|-------------------------------|--------------------|---------------------|---------------|---------------------|-----------------|
| | | | n (missing) | Variance of outcome | n (missing) | Variance of outcome | |
| Post-test | 1.02 | 1.06 | 543 (24) | 19.05 | 546 (11) | 17.15 | 18.10 |

The slightly higher difference between the adjusted means produces a slightly higher effect size, equal to 0.25. It can be concluded from these analyses that the *onebillion* apps had a positive and statistically significant impact on the measure of the pupils' mathematical attainment, which is equivalent to three months of additional progress, when it is used in small groups of pupils with the support of an adult as additional mathematics instruction.

Subgroup analyses

Subgroups defined by FSM status: The protocol specified two subgroup analyses. In the first one, the subgroups were defined by eligibility for FSM, defined here as FSM group if the pupils were eligible for FSM and non-FSM group if the pupils were not entitled to FSM. Entitlement to FSM is used in research as a measure of socio-economic status (SES) which indicates deprivation. It is important to include a measure of SES in the assessment of interventions in order to test their potential for closing the gap between the pupils from more affluent and less affluent families. If an intervention is more effective for those pupils who come from less affluent families, it has the potential for narrowing the gap between pupils from different SES backgrounds. Schools in the intervention and in the control group provided the evaluation team with information about which of the participating pupils were entitled to FSM.

In technical terms, the analysis aims to investigate whether group membership moderates the effect of the intervention; in non-technical terms, it aims to test whether the effect of the intervention varies with subgroup membership and is higher in one subgroup than the other. The models that test this hypothesis must include first the overall effect of eligibility for FSM and next the interaction effect between group membership, intervention versus control, and eligibility for FSM. The reason for including first in the model the effect of eligibility for FSM is that the pupils in the non-FSM group might show higher levels of attainment than those in the FSM group, irrespective of whether they were assigned to the intervention or to the control group. In this statistical analysis, a significant positive interaction between FSM and intervention would indicate that the intervention pupils eligible for FSM (coded as category 1) made more progress than those not eligible for FSM (coded as category 0); this would provide evidence to support the hypothesis that the intervention has the potential for narrowing the gap between the less and the more privileged pupils in SES terms. If the interaction is not significant, there is no evidence that one group benefited more from the intervention than the other. A significant negative interaction indicates that the pupils not eligible for FSM made more progress than those eligible for FSM; instead of narrowing the gap, the intervention actually widens the gap between the two groups.

For this step of modelling, the moderation effect of FSM eligibility was tested (see Appendix 3, Table 2, Models 5 and 6). Model 3 did not converge because it included a random slope parameter which was not significant. There were 286 FSM-eligible pupils (25.6 % of the sample), 831 (74.4% of the sample) who were not FSM-eligible, and 7 for whom it was not known and who were excluded from this analysis. The percentage of pupils eligible for FSM in this sample is considerably higher than the proportion reported as entitled to and claiming FSM in primary schools across school types (local authority and academies) in January 2018, which was 13.7% (Department for Education, 2018). This means that the group of pupils eligible for FSM is over-represented in this study, as would be expected because teachers were asked to identify the pupils in the lower half attainment group of the class, and pupils eligible for FSM in general tend to show lower levels of attainment in maths.

First a model including FSM-status and the FSM \times intervention (cross-level) interaction effects were included (Model 5) and then the model was run with only the 286 FSM-eligible pupils (Model 6). It is noted that the outcomes of this latter model must be considered with caution as the number of participants in the model is lower than that in the full analysis.

However, it still includes 88 schools and Maas and Hox (2005) argue that only sample sizes smaller than 50 at the higher level in multilevel analyses lead to biased estimates; thus the results should not be dismissed due to the smaller number of participants per cluster.

The model with 1117 participants for whom the FSM eligibility status was known (Appendix 3, Table 2, Model 5) indicated that there was a significant effect of the intervention, which was moderated by a significant negative interaction between trial group, intervention versus control, and FSM eligibility ($B=-1.15$; $p \leq 0.05$). FSM eligibility *per se* was not a significant factor in the analysis. Although at first glance this may seem surprising, it should be recalled that teachers were asked to nominate pupils whom they judged to be at the lower half of attainment in their classes. Table 11 presents an overview of the post-test findings for the groups by eligibility for FSM.

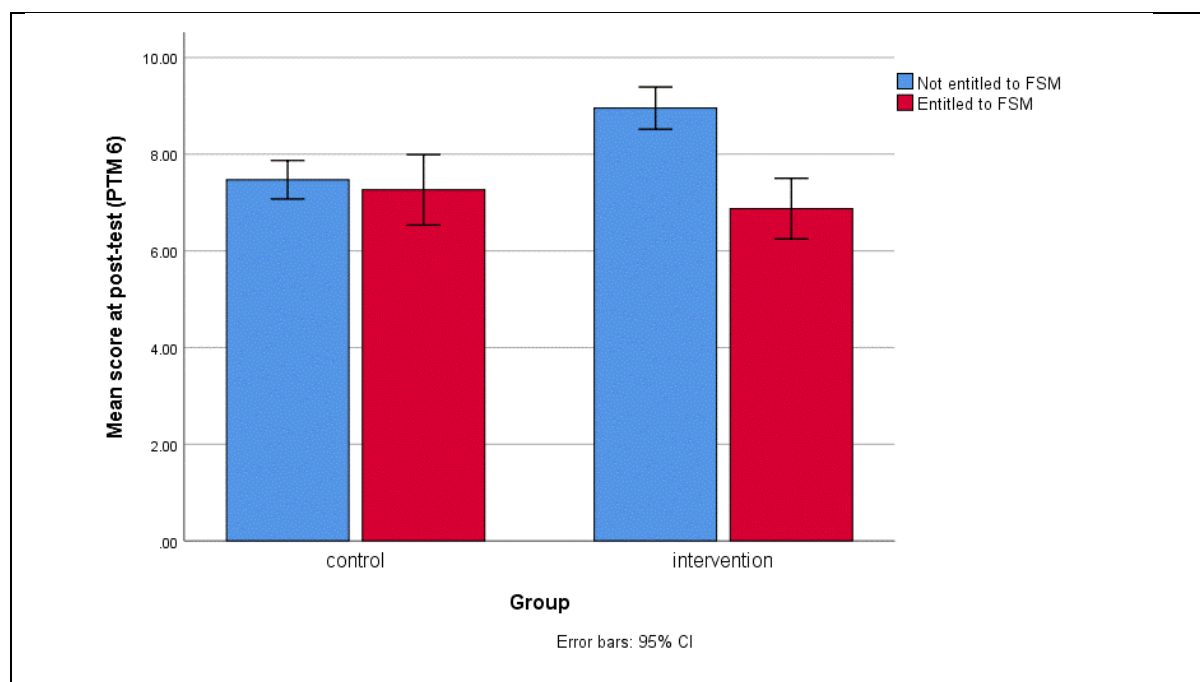
Table 11: Post-test outcome in intervention and control groups by FSM eligibility

| | Raw means | | | | Effect size | | |
|-------------------------------------|--------------------|----------------------|---------------|----------------------|----------------------------|------------------------|-------------------------|
| | Intervention group | | Control group | | | | |
| Outcome | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (int, contr) | Hedges g (95% CI) | p value |
| Post-test among non-FSM pupils only | 404 (13) | 8.95 (8.52, 9.39) | 407 (7) | 7.47 (7.08, 7.87) | 811 (404, 407) | 0.35 (−0.21, 0.48) | t(809)=4.93 p=<0.001 |
| Post-test among FSM pupils only | 134 (11) | 6.87 (6.25, 7.50) | 137 (4) | 7.26 (6.54, 7.99) | 271 (134, 137) | −0.10 (−0.33, 0.14) | t(269)=−0.79 p=0.43 |

When the intervention effects are considered only for the pupils not eligible for FSM, the intervention has a statistically significant effect as well as a higher effect size. In the group of pupils eligible for FSM, the mean for the control group is higher than the mean for the intervention group, which indicates a negative effect of the intervention. However, the difference between the intervention and the control group is not statistically significant. This lack of significance may be due to lack of power, and so one can simply conclude that there is no evidence at all of any positive impact of the intervention for pupils eligible for FSM. The effect of FSM status on the impact of the intervention is represented graphically in Figure 4.

The conclusions from this analysis are that eligibility for FSM affects the impact of the intervention: the intervention has a positive impact for pupils who are not entitled to FSM but there is no evidence of impact for pupils entitled to FSM. Thus there is no indication that the intervention has the potential for closing the attainment gap between pupils from a lower SES and those from more affluent families, and it may even widen this gap.

Figure 4: Means at post-test for the intervention and control groups by FSM status



Additional analyses

As indicated earlier, one intervention school did not implement the intervention, but agreed to the administration of the post-test. It was also found during process evaluation through the middle management questionnaire that one control school had purchased the apps and nominated pupils had had access to it, although the school did not quantify how much the pupils had used the apps. Additional analyses, similar to those presented previously for investigating impact, were run with the reduced sample of 1069 pupils in 110 schools. These analyses showed that removing the intervention school that did not use the apps as well as the control school that did use them had little effect on the outcomes. The only difference between the analyses was that the intervention now explained 1% more variance than in the previous model which included these two schools and all other findings were replicated.

Cost

This section of the report presents a cost analysis of the *onebillion* intervention on a per pupil basis. There are several key steps to be completed in order to calculate the average cost per pupil. The first step considers how many pupils would receive the intervention each year. The second step presents the costs of buying the *onebillion* apps, staff time for preparation and training, and the estimated costs of the resources required to implement the intervention. In the final step this information is collated to calculate the cost per pupil.

Previous EEF-funded evaluations that have involved teacher-led interventions have analysed the cost of the programme over a three-year period. This is based on the average length of time teachers in England spend in a school before moving to a different school, retiring, or leaving teaching (Allen, Burgess and Mayo, 2010). The cost analysis for this evaluation also uses a three-year period, enabling a direct comparison between other EEF evaluations of interventions, whether they are teacher-led or TA-led.

In this evaluation, free access to the *onebillion* apps was provided; 13 schools needed to borrow one iPad from the implementation team for uploading data and one school borrowed five iPads in order to participate. The control schools received a bursary for taking part and free access to the apps at the end of the project, which could be used with pupils who had not been nominated for the project. The investment in equipment is not trivial for schools; in 24 intervention schools (43.6%) the use of the iPads for the *onebillion* intervention prevented them being used by other classes; in 17 of these schools (30.9% of the total), activities using iPads with other classes had to be cancelled, and in the others it was rescheduled (middle management questionnaire, Question 13).

The following sections describe each of the key elements required to provide an accurate analysis of the cost effectiveness of the intervention.

Number of pupils

During the evaluation, all schools nominated nine or ten pupils to receive the intervention, depending on the size of the class. Pupils needed to use the same iPad for each session so that their progress could be recorded. Although two or more pupils can use the same iPad by having different logins, they cannot use the iPad at the same time. In the evaluation, 44 of the 55 schools delivered the intervention to groups of nine or ten pupils but 11 of the 55 schools delivered to two groups of five pupils, with some schools changing group size during the intervention. In general, the number of pupils using the iPads at the same time determines the number of iPads and the time required from an adult to supervise the implementation. In this analysis, it is assumed that ten pupils will work with the apps at the same time.

After one group finishes implementing the intervention with ten pupils, schools can use the equipment with other groups. In practical terms, schools would be able to implement the intervention with two groups of ten pupils each school year assuming that they started in the Autumn term. Over three years, it would be reasonable to expect a school to implement the intervention with 60 pupils. The implementation costs are presented in Table 12.

Table 12: 'One-off' and 'ongoing' costs (£) to implement with two groups each year over a three-year period

| <i>onebillion costs</i> | Year 1 | Year 2 | Year 3 |
|----------------------------|--------|--------|--------|
| iPad x10* | 3190 | 0 | 0 |
| Headphones x10 | 50 | 0 | 0 |
| Apps | | | |
| 3–5 licences x10** | 220 | 0 | 0 |
| 4–6 licences x10*** | 390 | 0 | 0 |
| Total | 3850 | 0 | 0 |
| Cumulative total | 3850 | 3850 | 3850 |

Figures may not sum to the total due to rounding. Sources accessed 20/09/2018.

*<https://www.apple.com/uk/shop/buy-ipad/ipad-9-7>

**<https://itunes.apple.com/gb/app/maths-age-3-5-for-schools/id688143717?mt=8>

***<https://itunes.apple.com/gb/app/maths-age-4-6-for-schools/id968679884?mt=8>

In this report only the cost of the resources needed to run the intervention (that is, sufficient iPads, headphones, and licences for the apps) will be included in the analysis. The cost of ten iPads and ten headphones can be considered by schools as 'one-off' investments. These also benefit the school by increasing their IT hardware and the up-front costs would not apply in subsequent years. App licences do not have to be renewed so they too can be considered one-off costs. The estimated costs are correct as of 20/09/2018 but do not include any educational or quantity based discounts. All prices are shown inclusive of VAT. The cost of the headphones was estimated based on a unit cost of £5 per set of headphones. The current entry level iPad was used in this cost analysis although more expensive versions are available. At the moment the Maths 3–5 and the Maths 4–6 apps run only on iPads, and so there is no possibility of using cheaper tablets to carry out this intervention.

Staff time

The largest cost to schools is often the staff time to deliver an intervention. There is a financial cost, regardless of whether a school chooses to pay a TA for their extra time, hire another TA, or reallocate a TA's time from other duties. This is in addition to any cost incurred by giving staff time to do online training and to familiarise themselves with the apps.

Training/familiarisation time: TAs will require three hours (half a day) to complete the online training and familiarise themselves with the apps. In this evaluation, face-to-face training sessions were offered but online training has been developed and this was used by the substantial number of TAs who did not attend one of the face-to-face training sessions (one session was cancelled due to extreme weather conditions).

There would also be some administration time spent liaising with the class teacher or school IT technician, which will remain unaccounted for in the calculation of staff time.

Preparation time: Information about preparation time was obtained through a TA questionnaire via which 54 out of 55 TAs reported the amount of time spent planning and preparing to deliver the intervention. The preparations involved installing the apps and collecting and charging the iPads. Table 13 shows that these preparations took about 5 minutes per session and thus 20 minutes each week over the 12 weeks of the intervention.

Implementation time

The intervention is designed to be implemented in 48 half-hour sessions; 24 hours in total, to each group of up to ten pupils.

Table 13: Staff time resources required to implement intervention with two groups each year

| Staff time resources (hours) | Year 1 | Year 2 | Year 3 |
|------------------------------------|---|---|---|
| Training time (TA) | 3 hours | 0 | 0 |
| Preparation time (TA) | 8 hours (5 minutes per group per session) | 8 hours (5 minutes per group per session) | 8 hours (5 minutes per group per session) |
| Delivery time (TA) | 48 hours (24 hours per group) | 48 hours (24 hours per group) | 48 hours (24 hours per group) |
| Total (TA) hours each year | 59 | 56 | 56 |
| Cumulative total (TA) hours | 59 | 115 | 171 |

Source: Length of delivery taken from evaluation team questionnaires to TAs and completed session logs.

Schools need to set more time aside in the first year of delivering the intervention compared to subsequent years. In the first year each school would need to set aside three hours for the TA for online training, including familiarisation with the apps and organising resources (for example, downloading the apps onto the iPads) and five minutes per session to prepare the material requirements of the intervention (collecting, charging, and returning the iPads) per group. The TA would spend approximately two hours per group per week implementing the intervention.

In the second and third years the time required decreases and remains constant. The total cumulative time required over three years, assuming two groups per year, is 171 hours for the TA. This information is presented in Table 13. This does not include any time that may be spent by either the TA or the class teacher selecting, assessing or reporting on pupils' progress.

Cost per pupil

As previously stated, the following calculations are based on the assumption that the intervention would be implemented with 20 pupils (two groups) each year over a three-year period. The costs per pupil, which do not include staff costs, are shown cumulatively per pupil in Table 14. The total estimated cost by Year 3 is £3850 per school or £64 per pupil per year; according to the EEF cost rating scale (Appendix A), this is considered 'very low'.

Table 14: 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 pupils worked with | 20 | 40 | 60 |
| Cumulative cost per school (£) for materials | 3850 | 3850 | 3850 |
| Average cost per pupil per year (£) | 193 | 96 | 64 |

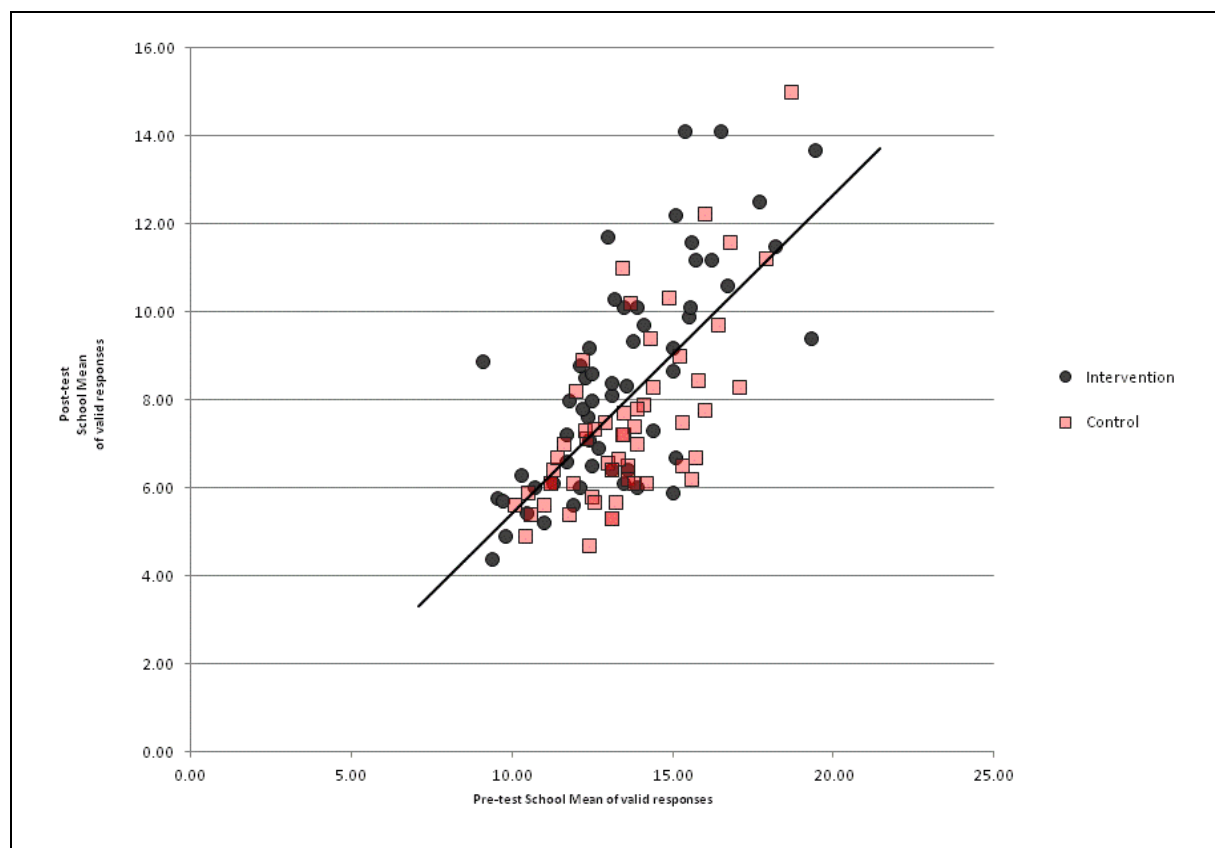
Figures may not sum to the total due to rounding.

However, should *onebillion* be implemented for fewer than the estimated three years or to only one group each year, then the pupil cost is higher. The cost per pupil increase would increase from 'very low' to 'low' if it is only implemented with two to three groups (20–30 pupils) in total over three years and to 'moderate' if it were only implemented with the first group of ten pupils.

Implementation and process evaluation

One of the main aims of the process evaluation is to understand how the intervention outcomes differ across schools and whether it is possible to identify factors that explain why the pupils might benefit from the *onebillion* intervention more in some schools than in others. Our first step was to plot a graph that shows the regression line of post-test school scores on pre-test school scores for both groups, control and intervention. Each school is identified in this graph by its mean score in the post-test plotted as a function of its mean in the pre-test. Figure 5 presents this graph.

Figure 5: School means at post-test by school means for the pre-test; the regression line is plotted for all schools independently of group



If a school's post-test score falls on the regression line, then it can be said to have done as well in the post-test as would be expected given its mean pre-test score. If a school's mean post-test score falls above the regression line, this score can be taken as better than expected, again given the school's mean pre-test score. If the post-test score for any school falls below the regression line, it can fairly be judged as having performed worse than expected, given its performance in the pre-test.

The graph shows clearly that the majority of the intervention schools had mean post-test scores above the regression line, whereas the majority of the control schools' means fall below the regression line. This provides powerful support for the conclusion presented earlier in this report of a definite impact of the *onebillion* intervention.

However, it is also clear that there are a few intervention schools whose mean post-test scores fell below the regression line. Thus, whenever possible, regression analyses were carried out in order to see whether it was possible to identify which factors within the intervention group differentiated schools that did reap the positive effects of the intervention from those that did not.

As indicated in the 'Methods' section, the factors we examined were classified into enabling factors, fidelity factors identified by the intervention team, and contextual factors which related to the differences observed in implementation identified through observations, TAs' responses to questionnaires and interviews, or middle management interviews.

The evaluation team had originally planned to collect session observations in ten schools. When confronted with a very wide variation in the conditions of session implementations, the team decided to increase the number of schools observed to 30 and collect more detailed observations than originally expected. The information collected from these

observations is described as ‘contextual factors’ and classified under different headings. Because these observations only cover one session out of 48 expected sessions, they cannot be treated as intervening variables that could be used to account for the intervention success in quantitative models. These factors are described quantitatively in order to offer some insight into the variations in circumstances of implementation.

Enabling factors

The logic model provided by the intervention team placed attendance to training and being prepared for it by bringing an iPad that could be used at the training as an enabling factor. This section describes how training was implemented and the TAs’ reactions to their training. Of the 57 schools that were assigned to the intervention group, 55 implemented the intervention and 54 answered the TA questionnaire about the training; this response rate of 99.2% can be considered very high.

How were the instructors trained?

Of the 55 intervention schools, 38 arranged for their instructors to attend one of the three courses. In 16 other intervention schools, the instructors’ training took the form of an online course which was devised by the intervention team. In one school the instructor received one-to-one training from the Programme Manager who was in charge of the day-to-day management of trial implementation.

How successful was the TA training?

Face-to-face training: The face-to-face training was planned to be very similar to the online course with respect to explanations about the implementation procedures; in fact, the intervention team basically used the videos prepared for the online course as the basis for the training. However, the face-to-face training also included further information about previous cohorts that had used the apps in the UK and Malawi and the evaluation of these previous implementations.

There were a few relatively minor technical difficulties in the face-to-face courses. At two of the courses some of the attendees were unable to access iTunesU from their iPads and were therefore unable to practise uploading the session logs. The trouble was that these particular iPads were older models, and the intervention team eventually solved the problem in the next face-to-face course by providing up-to-date iPads where these were needed. Of the 42 attendees at the face-to-face training, 15 did not think there were any challenging aspects of the intervention. Others thought learning how to use iTunes U would be challenging.

The online course: Members of staff from 16 of the schools assigned to the intervention group were unable to attend the training for a variety of reasons. These schools were sent the training materials for the online *onebillion* training and were asked to watch the online videos about the programme. After they had completed the online course, they were also asked to answer a post-training questionnaire which contained the same questions as in the questionnaire given to those who had attended one of the face-to-face courses, as well as some additional questions about the online training. Those who took the online course also judged the training that they had received and the Implementation Manual to be clear and effective. One of the TAs who took the online course reported difficulties in setting up the iPads with the appropriate codes. Another of the TAs who took the online course could not access iTunesU.

The main information that the evaluation team gathered about the success of the face-to-face training came from a questionnaire that was completed online by people who attended the courses, either at the event or soon after they were held. This questionnaire contained direct questions about the attendees’ opinions of the training that they had been given, and the answers to these were clearly on the favourable side. Of the 39 attendees at a face-to-face course who completed and returned the questionnaire on the course, 37 (94.9%) affirmed the clarity of the course that they had attended and felt that they had understood the part that they would play in implementing the instructions given in the course and also in the *onebillion* Implementation Manual.

Table 15 provides quantitative information about the answers given by people attending the face-to-face courses and by those who took the online course about the intervention. The questions are listed in the first column; TAs were asked to assign numbers indicating whether they disagreed with the affirmative (1), somewhat disagreed (2), neither agreed nor disagreed (3), somewhat agreed (4), or agreed (5). Such five-point scales are commonly used with people who have not received instruction on how to use the scales and are considered easy to answer.

Table 15: TAs' evaluation of the training they received for implementation of the intervention: mean agreement scores

| Question | Face-to-face courses (N=38) | Online training (N=16) |
|--|--------------------------------|---------------------------|
| The aims of the programme were clear | 4.89 | 4.63 |
| I understand the structure of the onebillion project | 4.95 | 4.44 |
| I understand the content of the onebillion apps | 4.92 | 4.69 |
| I feel confident to support children using the apps | 4.87 | 4.69 |
| I am clear about my daily tasks | 4.89 | 4.68 |
| I feel the Implementation Manual tells me all I need to know | 4.92 | 4.69 |
| Overall means | 4.91 | 4.63 |

With respect to the comparison of reactions to the two kinds of training, the main conclusion to be drawn from Table 15 is the near universal approval of the training given in both kinds of training course. The judgements made by those attending the face-to-face courses were more positive than those made by the online trainees but, although this difference was consistent across all six questions, it was too slight to warrant any definite conclusion about the relative effectiveness of the two different kinds of course. TAs were confident that they understood their tasks and were clear about all they needed to know in order to implement the intervention. This quantitative description was corroborated in comments by TAs under an open question, two typical examples of which are given here.

'Information was presented clearly and was easy to follow and understand.'

'All of it was very informative and useful.'

Considering that there was so little variability with respect to how prepared the TAs felt for implementing the intervention, there was no reason to analyse whether the differences in mode of training could explain the differences in success in the implementation across schools.

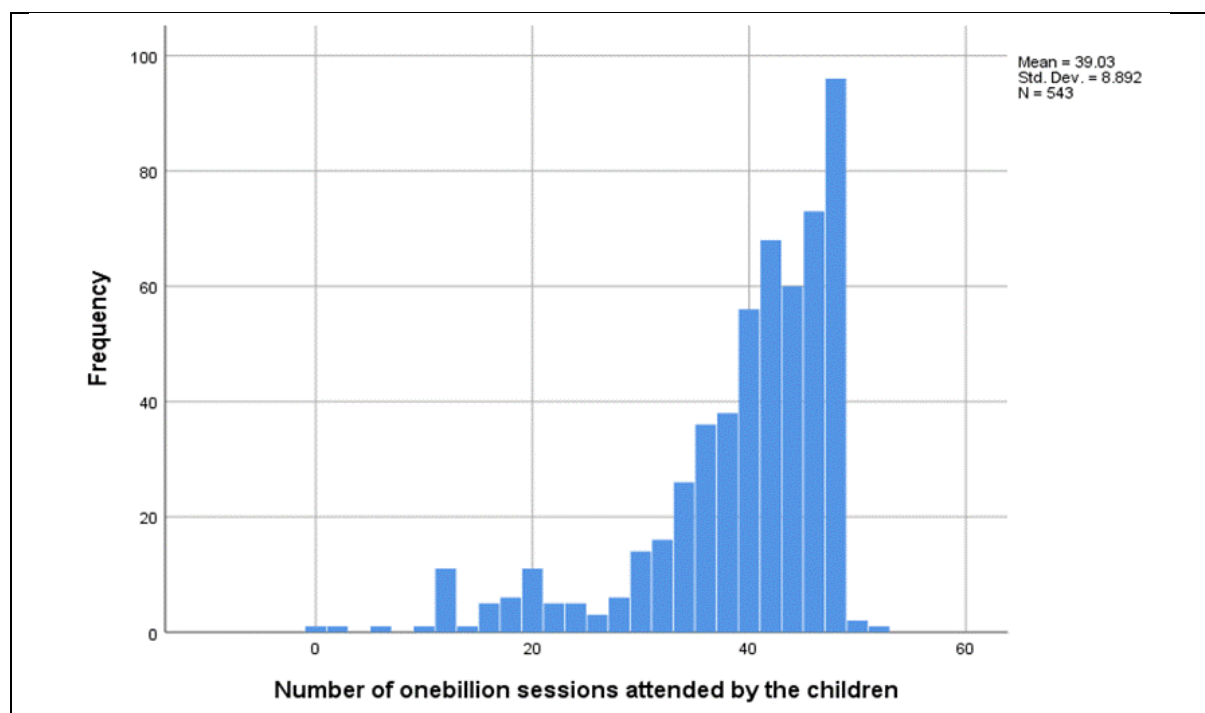
Fidelity

Fidelity measure at the pupil level: intervention dosage

The logic model provided by the intervention team specified time on the app and carrying out the activities as crucial to the success of the intervention. The logs provided by the TAs (which were to be sent digitally using the Numbers app or posted to the intervention team at the end of each week if the TAs did not succeed in filing these digitally) provided information about the number of sessions that each pupil attended, the highest certificate attained by the pupil and the most advanced activity that the pupil did on the last session. We also obtained from the session logs the number of sessions actually offered by the schools; variation in this number resulted either from starting the implementation late or from terminating it earlier than expected.

In spite of the expectation that logs would be provided weekly, many schools did not follow this procedure. The evaluation team followed up all schools until the end of the Summer and was able to obtain the logs for all 55 intervention schools that had remained in the study, including logs from the school that stopped delivering the intervention. The number of sessions attended by pupils is displayed in Figure 6. There is clearly great variability between pupils in the number of sessions attended.

Figure 6: Frequency of pupils who attended different numbers of sessions



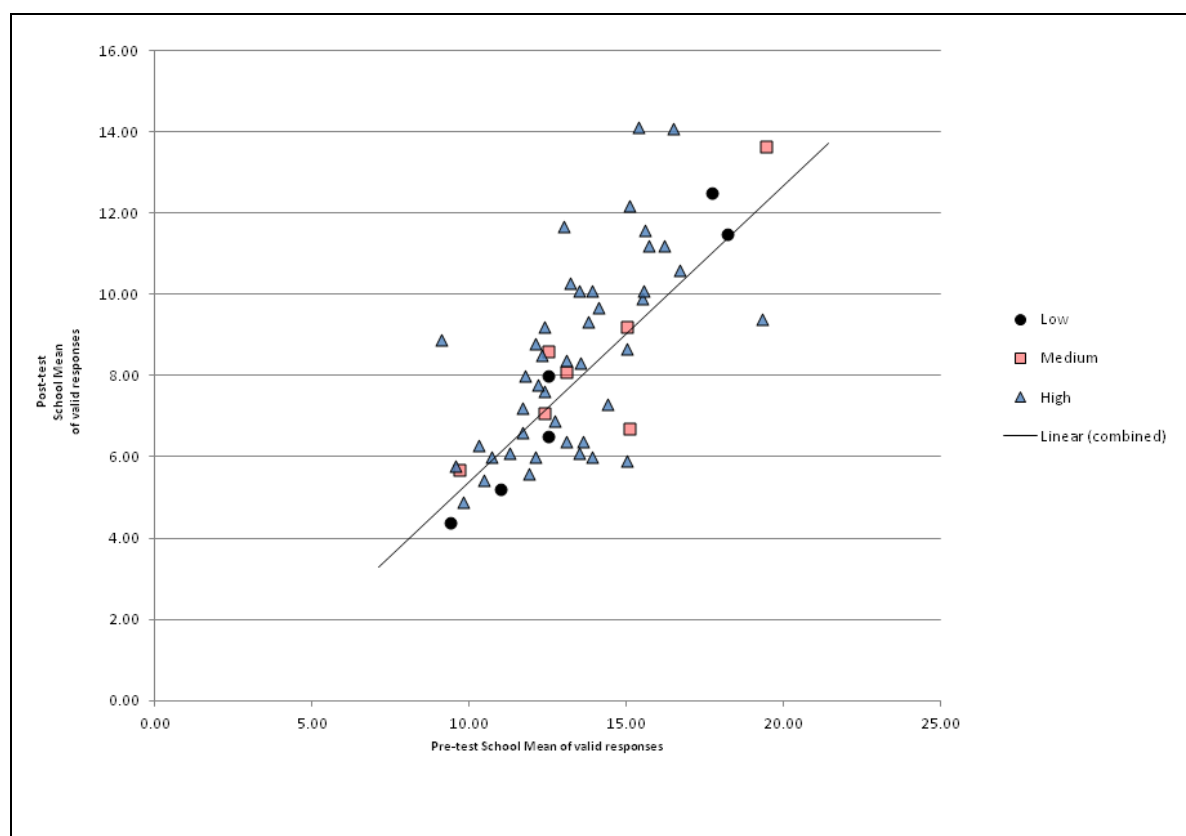
The intervention team (see Protocol, page 8) identified three levels of compliance for this trial: (1) low compliance, defined by participation in up to 30 sessions (62.5% of the sessions in this trial) which is equivalent to six full weeks of intervention delivered every day; (2) medium compliance, defined by attendance at between 31 and 40 sessions; and (3) high compliance, defined by attendance to at least 40 sessions (83.3% of sessions). The mean number of sessions attended by the pupils was 39, which is just below high compliance; the median was 41 and the mode 47. These figures indicate a high overall level of compliance with the dosage requirements as set by the intervention team. Of the 543 pupils included in the analysis, 13.1% would be classified as low compliers, 25.8% as medium compliers, and the remaining 61.1% showed high levels of compliance.

Two of the measures of dosage obtained for this project – highest level activity completed and number of quizzes successfully solved – were very highly correlated ($r=0.939$) and can be viewed as providing basically the same information; this high correlation can be seen as indicating that the information obtained from the logs is reliable. The number of sessions attended by the pupils correlated significantly but moderately with either of the first two measures: 0.58 ($p<0.05$) and 0.52 ($p<0.05$), respectively. Thus the number of sessions and the progress in the activities do not provide exactly the same information and it is worth analysing whether these two pieces of information can be seen as intervening variables that predict the pupils' post-test scores, after controlling for their pre-test scores.

The number of sessions attended by the pupils was not independent of the number of sessions offered by the school: thus a multilevel model was used to assess the effect of this measure of compliance on the impact of the intervention. A multilevel model (see Appendix 3, Table 6, Model 9) was used to test whether the number of sessions that the pupils attended explained further variance in the outcome measure, after controlling for pre-test and taking into account the nesting of pupils in schools. Only the intervention group was included in this model because it aimed to test whether the number of sessions was a moderator of the intervention effects. The measure of the number of sessions did not explain a significant portion of the variance either by itself or in an interaction term with pre-test ($B=0.01$; $df=497$; $p=0.894$).

The importance of compliance was explored further by applying the intervention team's definition of levels of compliance at the school level. A small number of schools (6 out of 56 or 10.7%) offered up to 30 intervention sessions or between 31 and 40 sessions (7 out of 56 or 12.5%); the majority of the schools (43 out of 56 or 76.8%) offered more than 40 sessions to the pupils. Using exactly the same rationale as the one outlined in Figure 6, a new graph was plotted showing the intervention group's school means at post-test as a function of the means at pre-test; however, this time, schools were differentiated by their classification as low, medium, or high compliers. This graph is presented in Figure 7. The regression line in the graph is based on the full sample of intervention and control schools.

Figure 7: School means at post-test by school means at pre-test with schools classified by their compliance level; the regression line is based on the full sample



The graph shows an equal number of low-compliance schools above and below the regression line as well as a substantial number of high-compliance schools below the regression line. This overview does not provide support for the conjecture that the main factor accounting for the success of the intervention is the number of sessions offered to the pupils.

A similar model was run with the number of quizzes successfully answered by the pupil; this measure of the pupil's use of the app did not produce a statistically significant result either.³

The correlation between the number of quizzes taken by the pupil and the number of sessions offered by the school was 0.44, which was lower than the correlation between the number of sessions offered by the school and the number of sessions attended by the pupil, which was 0.52. So it was decided to run an exploratory hierarchical linear regression analysis in which the number of sessions offered by the school would be entered in the model as the second step, after the pre-test results; the number of quizzes passed by the pupil was then entered as a third step. In this regression, the number of certificates made a significant contribution to the prediction of post-test scores. Although the amount of variance that it explained after pre-test and number of sessions offered by the school was small (3%), this finding suggests that it is not the number of sessions but success in the quizzes that matter for the pupils' progress. Appendix 6 presents the results of this analysis.

³Although the number of sessions was not a significant predictor of the outcomes in this analysis, it was thought of interest to compare the number of sessions attended by pupils eligible for FSM with the number of sessions attended by those not eligible for FSM. The mean number of sessions attended by pupils eligible for FSM was 38.5 (SD=9.7); the mean number of sessions attended by pupils not eligible for FSM was 39.3 (SD=8.3). An independent samples t-test showed that this difference was not statistically significant. Thus a difference in the number of sessions attended by the two groups could not account for the finding that the intervention had a significant effect for pupils not eligible for FSM and not for those eligible for FSM.

Fidelity measured at TA level: did the TA's delivery meet the guidelines in the Implementation Manual?

The Implementation Manual and Training Video 3 give 'Top tips' for implementing the intervention. As indicated in the 'Methods' section, the evaluation team produced a list of conditions to be met by the TA so that the pupils could work with the apps productively.

For the 30 sessions observed by the evaluation team, it was possible to give ratings of fidelity regarding whether the list of behaviours from the Implementation Manual were taking place during the session. Three ratings were produced: (1) low fidelity; (2) medium fidelity; (3) high fidelity in running the sessions. Eight sessions were classified as low fidelity as the TAs did not implement the tips on what to do before or during the sessions; 18 were classified as medium fidelity; and four as high fidelity sessions. Thus there was variability in the implementation with respect to the expectations laid out in the Implementation Manual. Some of the ways in which TAs fell below expectation were:

- Some TAs did not monitor whether all pupils were able to successfully work through the activities independently. Sometimes this was due to the limitations of space and free movement, sometimes due to the TA doing other tasks.
- Some pupils did not ask the TA for help and, if the TA did not routinely check, these pupils were not supported with their learning.
- Some sessions were very noisy, especially when headphones were not used and pupils became distracted and off task.
- Some TAs did not have time-efficient organisation for iPad identification and distribution at the start of sessions.
- Some TAs did not follow the 'Top tips' procedure in the Implementation Manual for learning support procedures.

Pupils missed sessions for a number of reasons. Some of the reasons that schools gave for some of the pupils having to miss sessions were:

- INSET days;
- class trips;
- sports day and sports day practice;
- snow days;
- bank holidays;
- World Book Day;
- staff sickness;
- staff shortages;
- assessment week;
- apps accidentally removed from iPads; and
- individual pupils also missed sessions when they were absent from school or if one of the iPads was broken or not charged. Some schools were able to 'catch up' missed sessions but others were not.

In order to see whether this fidelity rating was an intervening variable that predicted the pupils' outcomes, a multilevel analysis was run in which this factor was entered in the model after accounting for the pre-test and school effects (see Appendix 3, Table 6, Model 10). It should be noted that only 30 schools were included in this analysis as observations were not carried out in all schools. The analysis did not show an effect of the differences in TAs' fidelity of implementation on the post-test, after controlling for pre-test scores. Thus fidelity as measured by behaviours listed as 'Top tips' was not found to mediate success in the intervention.

Contextual factors

Kale *et al.* (2018) suggested that some differences between factors in the environment where technology is implemented, such as the motivation and ICT skill of the teachers and the accessibility of the resources, can affect the outcomes of an intervention using technology. These factors are not viewed as part of the intervention, but they can have a significant effect on impact. In this trial, we collected information on factors that were not viewed as part of the intervention, but were likely to influence the implementation process and outcomes. The information collected about contextual factors came from the observations, interviews and questionnaires answered by the TAs, and phone

interviews with middle management staff. Some of this information can be presented only in descriptive terms as it was not possible to create variables that could be used in quantitative analyses. In the first section that focuses on contextual factors, information about the context of implementation that was quantified and translated into a predictor of outcomes is presented; the more descriptive information is presented subsequently.

Implementation – the TAs' identification of their role in the intervention

In the intervention team's logic model (see Figure 1), the TAs' role in the intervention sessions is depicted mainly as technical and administrative. It is the TA's job, in cooperation with colleagues, to arrange a schedule for four intervention sessions per week, to ensure that the iPads are switched on at the beginning of each session and that the app is working in each of the iPads. No pedagogical role in the intervention sessions was assigned to the TAs in the logic model. As indicated in the 'Methods' section, there is a reference in the Implementation Manual to TAs' role also as to 'give learning support, by making sure children are attending to the learning'. The evaluation team conjectured that TAs might interpret this sentence differently. Early on in the observation of sessions this was found to be the case: some TAs brought additional materials to work with pupils who were finding a quiz challenging, others followed the guidance of sitting with the pupil and repeating the instructions from the app, others expressed frustration at not being able to teach the pupils because this was to be done entirely by the apps, and others seemed entirely unaware that a pupil had failed a quiz successive times and was off task or seemed to be on task but was pressing buttons randomly (that is, without stopping between the pressing of different buttons) or, in one case, crying. The evaluation team decided to increase the number of observations and to include questions regarding how the TAs viewed their own role in the implementation of the intervention in the questionnaire. This increase in the number of observations was approved by the EEF.

In the observations of a single intervention session in 30 of the schools in the intervention team, the evaluation team categorised each of the TAs' behaviour during implementation: (1) as giving minimal support; (2) as reacting only when an individual pupil asked for help and then providing the support suggested in the Implementation Manual; (3) as watching the pupils and giving proactive pedagogical support to all the pupils in the group; and (4) as proactively using additional materials such as blocks or counters or Numicon when the pupil had failed a quiz more than once. There was good variability in TAs' behaviour across the observation sessions. Three members of the team visited ten schools each and took detailed notes using the observation schedule; the classifications were implemented on the basis of the observation notes independently by three members of the team, who then discussed the classifications. These classifications were done blindly to the responses to the questionnaires, which were collected later, in week 9.

Responses by the TAs to the week 9 questionnaire also varied a great deal. The TAs' responses were anonymised and printed for analysis by the evaluation team, who were therefore blinded to the respondents' identity. When TAs were asked how they perceived their role in the intervention, their answers fell into four categories: (1) seven described themselves as giving 'behaviour support' or as being an 'observer'; (2) 34 used terms such as 'guide', 'facilitate', or 'support'; (3) seven referred to their role as 'teaching'; and (4) six gave answers to this question that were irrelevant (for example, 'I feel very important'). Box 1 presents a sample of answers from the questionnaires and interviews that were classified as describing the TA in the role of educator in the left column and the role of observer in the right column. The clear contrast between these answers illustrates the differences noted by observers across different schools in the implementation process.

Box 1: Illustrative answers by TAs who were classified as educators or observers by the evaluation team

| TA as educator | TA as observer |
|--|--|
| <ol style="list-style-type: none"> 1. It will need direct teaching; we were told we can use manipulatives, I use Numicon, so I have a box of manipulatives, whiteboards and pens. 2. If there is a new concept that they haven't covered before in class, then I can quickly think of different ways to explain and teach them. 3. I sit and see what's going on with them, then use practical resources to support their understanding. 4. I am a facilitator, now becoming an educator as it gets harder in 4–6 app. 5. I support on journey, I support and teach. I do it in a group of 5, I couldn't do the support otherwise. 6. I pre-teach if they are stuck on things. | <ol style="list-style-type: none"> 1. If you have a child stuck, it's tricky, so actually it's hard. I say, 'try again'. I just praise them. If they don't get it, they move on to another app. I say 'doesn't matter' but it gets frustrating for them. We can't help them though. 2. I oversee it, control behaviour. 3. I observe from a distance. 4. I'm a bystander who is just there to sort problems. 5. I sit and watch them. 6. I crowd manage. 7. I keep crowd control. |

The role described by the educator and guider/supporter groups was pedagogical whereas the TAs who considered themselves observers or managers of behaviour did not take on a pedagogical role. Table 16 shows a cross tabulation of the evaluation team's classification of the TAs' behaviour during the observed sessions (the scoring of this variable is described earlier in this section) and their own responses to the TA questionnaire (excluding irrelevant answers – see section on 'Fidelity' to see how this variable was scored).

Table 16: Cross tabulation of how TAs viewed their role (from TA questionnaire) and how they delivered the intervention (from the observation)

| | | How the TA delivered the intervention (from the observation) | | | | |
|-------------------------------|--------------------------|--|--------------|---------------|--------------------------|-------|
| | | Minimal interaction (1) | Reactive (2) | Proactive (3) | Additional materials (4) | Total |
| How the TAs viewed their role | Observer (1) | 2 | 1 | 0 | 1 | 4 |
| | Guider/ supporter (2) | 0 | 13 | 3 | 1 | 17 |
| | Educator (3) | 0 | 2 | 2 | 1 | 5 |
| Total | | 2 | 16 | 5 | 3 | 26 |

The evaluation team expected:

- TAs observed to have minimal interaction in delivering pedagogical assistance to align with the role of an observer;
- TAs observed as either reactive or proactive in providing support to align with the role of a guider/supporter (responses provided to the questionnaire did not indicate whether a TA was reactive or proactive); and
- TAs who used additional materials when giving pedagogical help to align more closely with the role of an educator.

Correlation indices of ordinal by ordinal variables between the observations and the TAs' own descriptions of their role and the classifications provided on the basis of observations (see Table 16) were significant: Kendall's tau = 0.352 and Spearman's $r=0.357$. It was concluded that the observations provided some validation of the answers that TAs gave to the questionnaire regarding how they perceived their role; this answer, which was obtained for over 90% of the sample, could be used as a variable to investigate whether the TAs' perception of their role was a mediator of outcomes.

A regression model that included the TAs' perception of their role after the pre-test scores (see Appendix 6) and the number of sessions run by the school (to represent a school factor) showed that the TAs' perception of their role accounted for a statistically significant amount of extra variance in pupils' post-test scores. The TAs' perceived role was, in fact, more important than the number of sessions implemented by the school. Figure 8 plots the school means at post-test and a function of the pre-test means for the intervention schools. Schools are differentiated by the role that the TAs thought they were expected to play in the implementation of the intervention. The regression line included in the graph is the regression for both the intervention and the control schools, in order to show the impact of the intervention and to facilitate the comparison between this graph and the one presented in Figure 8.

Figure 8: School means at post-test by school means at pre-test with differentiation of schools by the TAs' perceived role in the intervention delivery

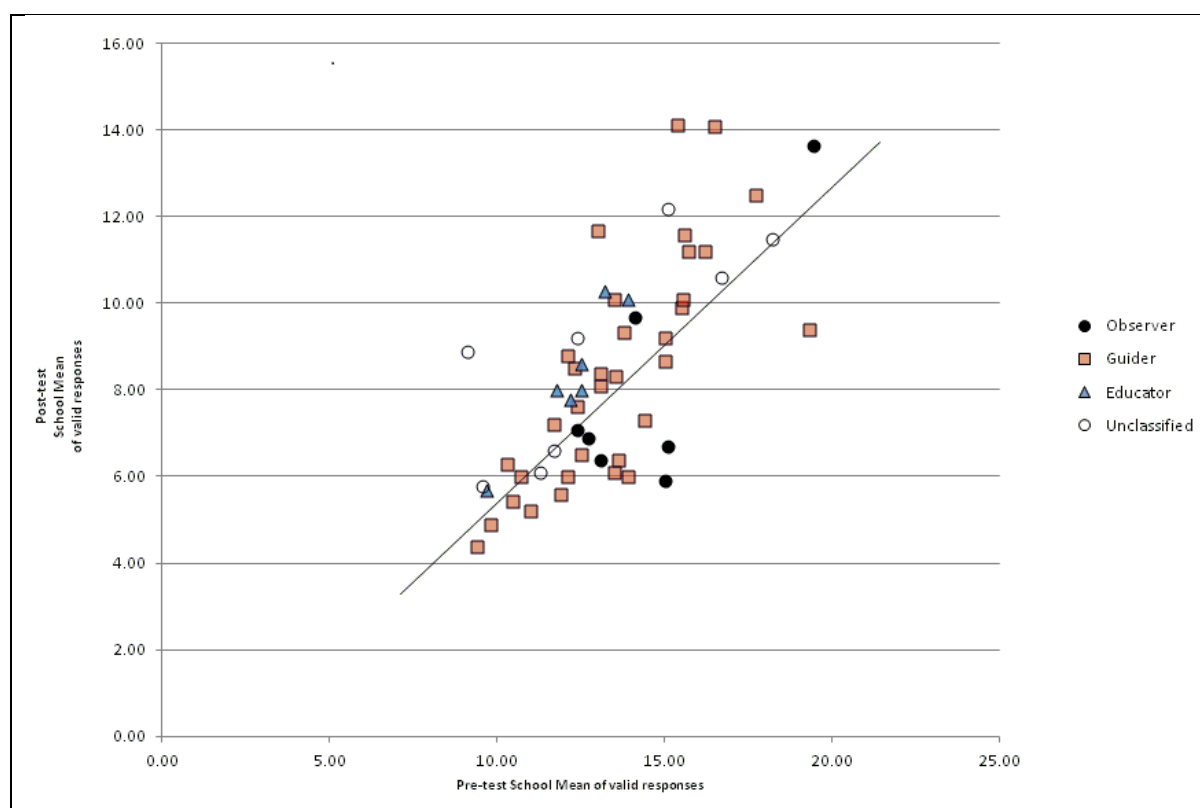


Figure 8 shows that, when the TAs perceived their role to be that of an educator, the school means at post-test were consistently above what was predicted from the regression of post-test on pre-test scores (7 out of 7 or 100%). In contrast, when the TAs perceived their role as observers, the majority of the school means (5 out of 7 or 71%) was below the prediction from the regression line in schools. Where the TAs perceived their role as guider/supporter, the majority of the means was above the regression line (19 out of 34 or 55%), two (6%) were placed on the regression line,

and 13 (38%) were below the regression line. It can be tentatively concluded that the TAs' involvement with the pupils' learning was associated with the TAs' success in the use of this intervention. However, this is a post-hoc conclusion from the analysis of the implementation and process evaluation. Only testing this conjecture in a systematic experimental design would provide unambiguous evidence for the role of the TA in the implementation of the intervention as a moderator of its effect.

Need for TAs' involvement during implementation

The findings just reported suggest that there is a real need for the TAs' engagement with the pupils during implementation. Information about this need was collected through observations of sessions as well as through the questionnaire answered by TAs.

The team was able to observe a whole intervention session in 30 of the 55 schools in the intervention group which delivered the intervention. The observer recorded actions by the TA and by the pupils taking part in the intervention session. The records included, for example, the TA providing technical (such as restarting the app) or pedagogical (such as explaining a concept verbally or with the help of other materials, listening to the instructions with a pupil, talking through the activity but letting the pupil touch the screen) assistance and the maximum length of time it took the TA to help a pupil who was requesting assistance. The observer also recorded the pupils' reaction to passing a quiz, how often the pupils were off task and whether they repeated any of the instructions.

In the first 15 minutes of each observation session, the researcher closely observed one pupil who had been identified by the TA as working well, another identified as struggling with the apps, and a third identified as somewhere in the middle; each pupil was observed for five minutes. During the second 15 minutes the researcher observed the TA's interactions with the group. Table 17 presents actions by the TAs during the 15-minutes observation slot when the observer focused on a single pupil as well as the pupil's behavioural events during this period. Table 18 presents the information collected when the focus was on the TA's interactions with the group. A rough estimate of frequency of the event per minute can be obtained by considering the total time of observation to be 450 minutes for each table. It is noted that this is a rough estimate of the need for support because some TAs showed a minimal level of interaction even when a pupil was finding a quiz very challenging; other TAs were proactive insofar as they observed the pupils and intervened when a pupil was seen to have failed a quiz repeatedly.

Table 17: Number of instances that a particular TA action or a particular pupil action was observed during the first 15 minutes of observation

| Observation of <i>onebillion</i> sessions (first 15 minutes) with three identified children, 5 minutes per pupil | No. of instances recorded across all observations |
|--|--|
| TA gave technical support | 11 |
| TA gave pedagogical support | 49 |
| Pupil had to listen to instructions more than one time | 67 (out of 105 times when pupils were observed to start a new activity) |
| Pupil was off task | 30 |
| Pupil won certificate | 38 |

Table 17 shows an imbalance between when, according to the TAs' judgement, a pupil needed technical support or pedagogical support. The number is higher for pedagogical than for technical help: pedagogical support was required approximately once every four minutes (the observations covered approximately 15 × 30 minutes). The need for technical support was relatively low. This suggests that, in the TAs' judgements, there was a far greater need for pedagogical than for technical assistance in *onebillion* sessions. Table 18 corroborates this observation and shows (as

expected from the nature of the observation which focused on the TAs and thus covered all the interactions in the group) a greater incidence of need for pedagogical support (150 instances) than for technological support (35 instances); the need for pedagogical support was noted approximately once every three minutes. The number of times that the TA gave technical support increased from 11 to 35 times. Table 18 also shows that there were few instances of one pupil attempting to help another by touching the other's screen.

Table 18: Number of instances that a particular TA action was observed during the second 15 minutes of observation

| Observation of <i>onebillion</i> session (second 15 minutes) – all pupils | No. of instances recorded across all observations |
|---|---|
| TA gave technical support | 35 |
| TA gave pedagogical support | 150 |
| TA intervened when another pupil touched the screen of a pupil | 4 |

The responses to the TA questionnaire indicated that the need for pedagogical support was considerably higher when the pupils were using the Maths 4–6 app than when they were using the Maths 3–5 app. Table 19 shows how often the TAs in their answer to the questionnaire judged that the pupils needed pedagogical support when they were using the Maths 3–5 app or the Maths 4–6 app. A Wilcoxon Signed Rank Test for correlated samples showed that, according to the TAs, support was required much more often when the pupils were using the Maths 4–6 app than when they were using the Maths 3–5 app.

Table 19: Frequency of responses by TAs in the TAs' questionnaire indicating how often they thought the pupils needed support when using each of the apps

| | How often have you needed to give pupils pedagogical support when they used the apps? | | | | |
|---------------|---|-------|--------|-------|--------------|
| | Very often | Often | Rarely | Never | <i>Total</i> |
| Maths 3–5 app | 0 | 19 | 34 | 1 | 54 |
| Maths 4–6 app | 8 | 26 | 19 | 1 | 54 |

In summary, the session observations indicate that there is a need for the TAs to play a more active pedagogical role than is indicated in the intervention team's logic model. This is particularly important when pupils use the Maths 4–6 app.

Material conditions

The information in this section is mostly based on the responses to the TA questionnaire; percentages are based on information available from 54 schools. Further notes are based on the session observations.

Number of available iPads: Owing to demand in the rest of the school, 13 schools (24%) in the intervention group reported problems in securing sufficient iPads to run the intervention; we note that this percentage is a slight under-estimation of the difficulty as some schools received iPads from the intervention team in order to participate in the intervention and we expect that, in their case, the iPads were not being used by other classes. Ten schools (18%) reported problems in securing sufficient headphones to run the intervention.

Using headphones: Eight schools were observed where the majority or all of the pupils were not wearing headphones by choice. In some schools, one or two pupils from the group did not wish to wear headphones as they disliked the noise made by the app when an answer was wrong; one pupil asked to be withdrawn from the project for this reason. In the observed sessions the use of headphones helped the pupils to concentrate for the duration of the session; in sessions run without headphones the level of noise was generally high. The majority of schools observed did not have spare headphones readily available; if one set did not function, the pupil had to work without headphones. When just one or two pupils did not have headphones, the level of noise was not high and this did not seem to interfere with the other pupils' concentration as much as it did when the majority of the pupils did not wear headphones.

Dedicated space: Eleven of the 55 schools (20%) did not have a dedicated space for the intervention. They used available spaces on a daily *ad hoc* basis. The remaining 43 schools (80%) reported having a regular space: nine used the staff room, seven used a corridor, seven used the classroom, seven used an intervention room, six schools used IT suite, three used a spare classroom, two used the library, one used the school hall, and one used a 'shared space'. Session observations indicated that the space *per se* did not affect the running of the session as much as how the space was used. Two aspects of the use of space were noted in the observations and are described below.

Seating organisation: Seating organisation was an aspect of implementation raised during two of the face-to-face training sessions; as mentioned previously, the training video showed the pupils seated at a table with the TA facing them, so that the TA was able to see the pupil's screen that was lying flat on the table. The evaluation team considered this an implicit message in the videos. At the face-to-face training sessions, the intervention team said that this was a good way to organise the seating but that if people wanted to organise seating differently they could do so. A range of seating arrangements was observed. Some rooms had a large table to seat all pupils, others had several small tables, some were spread around a classroom, some did not have tables or chairs so the pupils worked on the floor without access to tables or sometimes chairs. It was the impression during the observations that those pupils with a chair and a table on which to rest the iPad were generally better able to sustain concentration for the duration of the session.

TAs' ability to circulate the group or observe screens: Some spaces were such that the TA could not sit close to the pupils or move behind them because the space was limited or the pupils were sitting against a wall. The TA was not able to see the screens easily, so it was more difficult for the TA to monitor the pupils' progress during the session. It was easier for TAs to be proactive in supporting the pupils when they were able to see the pupils' screens. Most timely responses to requests for support were in situations where the TA could see the screens easily from where he/she was sitting or was able to move around to monitor the pupils' activity throughout the session.

Group size: Of the 30 schools observed, nine were running the sessions in two smaller groups rather than one group for all the pupils each day. Four of these schools indicated during the interview that this was due to not having enough working iPads to run the larger group.

During the observations, TAs' response time to pupils' requests for help were monitored; the size of group seemed to have less effect on response time than did the seating arrangements. Questions were answered quickly in both the larger groups of ten and the smaller groups of five.

TAs' need for technical support for implementing the intervention: TAs themselves reported that they had needed technological help to solve some of the problems and had to ask the school IT technician for assistance; 17 TAs (31.4%) reported that they needed help, with a range of difficulties listed, including setting up the apps at the start, apps freezing, codes not working, reinstalling apps, and Wi-Fi connections. Five TAs reported that they were unable to solve the technical problems satisfactorily even after they had sought technical help.

Fit with school planning

ICT resources already available: Teachers in both the control and intervention schools were asked about the resources already available to them in the domain of ICT: 92% of intervention schools and 96% of control schools have an IT Coordinator; 52.8% of intervention schools and 64.2% of control schools have an induction procedure for staff which includes IT; 18% of the control schools would have needed to buy additional iPads if they had been randomised to the intervention group and 70% would have needed to buy additional headphones; 5.5% of the intervention schools had to buy iPads and 38% had to buy additional headphones to implement the intervention. Five intervention schools reported that parents of children participating in the intervention had purchased the app for their use at home; one control school reported using the *onebillion* maths apps with nominated pupils.

These figures suggest that schools are not entirely prepared to use an app-based intervention in groups of ten but that the investment required may not be too large to prevent schools from considering such interventions.

Were schools using or planning to use app-based interventions already? When an intervention is adopted by a school in the context of a research project, it may fit with the school planning because the school already does something similar or because the school is planning to introduce something similar. In order to ascertain whether the intervention fit with the school planning, telephone interviews were carried out with middle management staff in both control and intervention schools (ten in each group). In all the intervention schools, iPads or maths software were already used with Year 1 pupils but, in three of the control schools, this was not the case. The software mentioned by middle management staff in control schools was Mathletics, Espresso, Busy Things, Under the Sea, 2Simple, Education City, ICT games website, Beebots, Sonic Slugs, IWB games (not detailed). The software mentioned in the interviews with staff from intervention schools was Mathletics, Purple Mash, RM Maths, Easy Maths, Pop Maths, Times Tables Rock Star, Maths apps on iPads (not detailed), Maths games (not detailed). Intervention schools' staff affirmed that *onebillion* was used in addition to normal maths lessons.

Schools were asked to detail the main priorities in their School Improvement Plan for maths for 2017/18. The evaluation team categorised the targets into four groups: raising attainment; trying to become a Mastery school; improving reasoning skills; and improving staff skills. Some schools had more than one target. There were no differences between control and intervention schools. Schools were also asked about their School Improvement Plan for ICT. Three schools (one control and two intervention) did not have a plan. The other schools mentioned e-safety, acquiring more hardware, developing a scheme of work, and improving staff skills. Thus no school mentioned the aim to introduce app-based learning as one of their priorities either in the maths or in the ICT plans.

Scheduling: Of the 54 schools responding, 15 (27.7%) did not have a fixed schedule for delivering the intervention. The timing of the sessions in these schools was varied to fit with timetabling of lessons, particularly on foundation subjects, or with other interventions. Of the 54 schools, 21 (38.8%) said it was not easy to fit the intervention into the timetable.

Who delivered the intervention? As indicated earlier, for brevity this report refers to the intervention deliverer as 'TA'. Information in this paragraph was provided by schools delivering the intervention. In 39 schools only one person delivered the intervention. In the other 16 schools, between two and five people delivered the intervention over the 12 weeks. In some schools, there seemed to be a regular pattern (for example, TA1 one day a week, TA2 three days a week) but in other schools there was no clear pattern. In total 76 TAs/Higher Level TAs (HLTAs)/teachers were involved with the intervention across the 55 schools. The role of the main person delivering the intervention was TA in 39 schools (78%), a HLTA in four schools (7%), a teacher in five schools (11%); in seven schools (4%) there was equal sharing between a teacher and a TA. When there was a systematic pattern of sharing the supervision of the sessions, TAs were asked to answer the questionnaires together.

Outcomes

The impact on pupils' attainment has been reported in the previous section. This section reports on the pupils' and the TAs' reactions to the intervention.

Pupils' enjoyment of the apps

The pupils' reactions were reported by TAs; no interview was carried out directly with the pupils.

TAs reported that pupils generally enjoyed the apps; in only one school one pupil was reported to dislike so much the sound that the apps made when a wrong answer was given that the pupil's parent withdrew the pupil from the project; none of the schools reported any pupils not enjoying the intervention overall. TAs were asked to indicate whether all pupils, most pupils, some pupils or no pupils in their group enjoyed using the apps. Their responses are reported in Table 20, separately for each of the apps.

Table 20: Frequency of responses by TAs indicating how often they thought the pupils enjoyed the apps

| | Did the pupils generally enjoy doing the apps? | | | | |
|---------------|--|--------------|--------------|--------------|--------------|
| | All of them | Most of them | Some of them | None of them | <i>Total</i> |
| Maths 3–5 app | 32 | 19 | 3 | 0 | 54 |
| Maths 4–6 app | 15 | 32 | 7 | 0 | 54 |

A Wilcoxon Signed Rank Test for correlated samples showed that, according to the TAs, the pupils' difference in enjoyment of the two apps was significantly different.

Most pupils were able to use the iPads without difficulties, although 13% of TAs reported having at least one pupil who had not used an iPad before and required some TA instruction. Box 2 presents some of the TAs' comments.

Box 2: A sample of TAs' comments in response to open questions included in the TA questionnaire

| |
|---|
| TAs' comments about the apps included: |
| <ol style="list-style-type: none"> 1. It's fun, bright colours and the children enjoy it. 2. The children get very excited when they get a star or complete a quiz and get a certificate. 3. Children are engaged, they think it's fun. 4. They are very keen and eager to use the apps and become absorbed. |
| TAs' comments about differences in enjoyment between the Maths 3–5 app and the Maths 4–6 app included: |
| <ol style="list-style-type: none"> 1. Keeping the children focused on the Maths 4–6 app was one of the most challenging things about the intervention. 2. Some areas in the Maths 4–6 app have not been covered in the curriculum so children have found some areas challenging. 3. For five of my children of lower ability, they really struggled going from the Maths 3–5 to the Maths 4–6 app. 4. Lowers now struggling with the Maths 4–6 app. Our lowers were not ready for this. 5. The Maths 4–6 app is harder for the less able. 6. Maths 3–5 they flew through but the less able struggle now (with the Maths 4–6 app). |

TAs' perceptions of what worked and what was challenging

TAs were asked what they thought were the best and the most challenging aspects of the apps. A sample of the positive comments is presented here with their frequency in brackets:

- children's enjoyment (21);
- children able to go at own pace (20);
- children's confidence and or perseverance (7);
- apps contents related to the curriculum (5);
- improved counting skills (4);
- children's progress (3); and
- more practice with maths (1).

Comments regarding the most challenging thing about the programme and their frequency (in brackets) were:

- time required for the programme (12);
- having sufficient iPads (6);
- amount of TA support required as the apps got harder (6);
- children off task linked to length of sessions (5);
- difficulties unlocking apps (3);
- having sufficient working headphones (3);
- codes not working for apps (2);
- children upset when failing quizzes (2);
- children upset with the sound app made when answer incorrect and then not wanting to wear headphones (2); and
- iPads breaking (1).

These responses to open-ended questions suggest that the overall reactions were positive, but that there are aspects of implementation that could be improved. The focus on technical difficulties among the challenges obscures the fact that the time required for implementation (mentioned by 18 TAs) is directly related to the TAs' view that the pupils needed support (mentioned by six TAs), which means that the pupils could not be left to work on their own.

Formative findings

As indicated earlier on, pupils generally enjoyed the apps and celebrated success in the quizzes. However, even though this was not mentioned often, some disliked the noise that corresponded to feedback when they had produced a wrong answer. To the best of our knowledge, only one pupil refused to continue with the intervention because of the noise of the apps and only in one school did pupils stop using the headphones for that reason, although several schools commented on how the noise of the apps was unpleasant. These comments were informal and we cannot quantify them. However, it is possible that this is an easy change to make on the apps.

The differences between the two apps as perceived by the TAs with respect to the need for support and level of enjoyment of the pupils suggest that the step that the pupils need to take to move from one to the other may be too large. The *onebillion* designers could consider whether it is possible to create a smoother transition between the two apps.

TAs were asked to offer their own 'top tips' for other TAs implementing the intervention in the future. They were asked both for technical and pedagogical tips. A summary of responses from the interviews and the questionnaires is presented in Box 3. Some, but not all, of their technical tips were mentioned in the Implementation Manual. However, there are no pedagogical tips in the Implementation Manual and their suggestions might be useful for future implementations.

Box 3: TAs' pedagogical and technological tips

| Pedagogical tips | Technological tips |
|--|--|
| <ol style="list-style-type: none"> 1. Be aware that you may not have covered some of the topics yet so the children will require extra support. I have used this situation as pre-tutoring. 2. Don't be scared to use other resources to support the apps. The children are used to holding resources and this helped. 3. Encourage them to use number tracks for addition or subtraction. Counters to find half or quarters. 100 square to find odd and even, etc. 4. We support the children by rewording the question and explaining the concept or method to enable them to access the question. 5. To have items available to support children when having difficulties understanding the particular subject they are learning. 6. Encourage practice of counting in fives, twos and tens in class. 7. If a child is struggling, give them something physical to count with as this usually helps. 8. (It) will need direct teaching, we were told we can use manipulatives, I use Numicon, so I have a box of manipulatives, whiteboard and pens. 9. Be aware that children often don't listen to instructions properly. It's not always an issue with their understanding of the objective. Often repeating the instruction slowly or in a different way is all they need. 10. For sums, help the children to understand that the equals sign is like a balance on a weighing scale. 11. Do it in a group of five, I couldn't do the support otherwise. I have cut out materials for the fractions work. | <ol style="list-style-type: none"> 1. Make sure the iPads are charged regularly. 2. Make sure the iPads have enough battery life when you start each morning. I checked most evenings before I went home. 3. Always charge the iPads as freezes seem to happen when the battery is less than 75%. 4. If app freezes, come out of it, go into any other app then return to <i>onebillion</i>. This usually restarts the app. Closing the app and restarting it. 5. Because we do have a lot of headphones, I actually kept the ten I needed in a box with the admin book and named each headphone. 6. Children at the beginning made quite a bit of fuss about them (headphones) but once I named them they were great. I also numbered each iPad and the children knew straightaway which one was theirs. 7. Make sure you have enough good quality headphones and spare ones if possible. 8. Expect there to be some technology issues early on, especially if a child is not used to using an iPad. 9. Make sure the unlock codes all work beforehand and have a few spare ones just in case – allow downloading time for the apps as they can take quite a long time out of a session. 10. Teach the children to turn up the volume, etc. before starting the apps. 11. Train children to pick headphones and their labelled iPad. Make sure you have enough headphones and remind children to take care of iPads. |

The most important finding from the implementation and process evaluation is, in our view, that the intervention team should consider making changes to their logic model. The pedagogical role that TAs had in the implementation of this intervention calls for a revision of the role of the TA in the logic model. The expectation that TAs who are simply observers (or 'bystanders' or simply in charge of 'crowd control', as some of the TAs put it) and that the apps teach by themselves does not seem realistic for most pupils. The description of a clear pedagogical role and its inclusion in the training materials is expected to bring substantial improvement to the intervention.

Control group activity

Phone interviews were carried out with middle management staff in order to find out what activities were taking place in control schools with Year 1 pupils who had been nominated for the project; information was also collected by means of questionnaires answered either by middle management staff or the class teacher or TA. The points for the telephone interviews had been sent to the relevant staff prior to the interview as there would be a need to ask colleagues in the school so that answers were accurate. Middle management staff in intervention schools were interviewed using the same methodology as it was necessary to find out whether the nominated pupils had received all the expected support which they would have received if they had not been participating in the intervention.

Schools were asked whether and which maths interventions were carried out with Year 1 pupils needing support and to tell the researcher about the maths interventions that were used. Both intervention and control schools reported using pre-lesson and post-lesson teaching as well as specific interventions designed to improve number recognition, counting and number bonds; most of these were school devised but some were commercially produced interventions (for example, Numicon, 1st Class@Number). Control schools mentioned the use of these interventions more often than intervention schools; 27 of the 55 control schools (48.2%) gave additional support for maths to some or all of the nominated pupils. Intervention schools were specifically asked whether *onebillion* had replaced other maths

interventions usually offered to pupils in the school; 24 schools (43.6%) replied that this was the case and 29 schools (52.7%) that it was not; two schools did not answer this question. Thus the best conclusion from this information is that pupils' participation in the *onebillion* intervention did not give them fewer learning opportunities than those that are usually offered to pupils in Year 1.

Conclusion

Key conclusions

1. Pupils who received *onebillion* made an additional three months' progress in maths compared to the control group. This result has very high security.
2. Pupils eligible for free school meals made 2 fewer months' progress in maths if they received *onebillion* compared to those in the control group. However, this analysis involves a smaller number of pupils so we are unable to confidently claim that this negative impact is likely to occur for FSM-eligible pupils outside of this research project.
3. The process evaluation suggested that the impact of the programme might be influenced by the amount of the pedagogical support given to the pupils during the intervention sessions. Exploratory analysis suggested that pupils tended to do better when supervised by TAs who thought that their role was to teach concepts when the pupils had difficulty.
4. In this project, teachers started with Maths 3–5 and then moved to the Maths 4–6 app. TAs reported that pupils enjoyed Maths 3–5 more and required less pedagogical support to use it.
5. Further research is needed on the nature of the pedagogical support that works best in *onebillion* sessions and the effects of the programme on the mathematics attainment of pupils entitled to FSM.

Interpretation

This evaluation of the *onebillion* apps has established, through a rigorous design, that the intervention has a positive impact on Year 1 pupils' mathematical attainment. This finding is qualified by the statistically significant interaction between the intervention and eligibility for FSM: there is no evidence that the apps had a positive impact on the attainment of pupils from lower SES backgrounds as defined by eligibility for FSM. This finding is also qualified by the fact that the design investigated the impact of the apps when they are used in addition to usual mathematics teaching; thus they could result from the fact that the pupils have extra time dedicated to learning mathematics. It is stressed that this was the only possible design, given the intensity of the use of the apps – four times a week for 12 weeks. It would be unlikely that the intervention team could recruit sufficient schools if the apps were to be used instead of regular mathematics teaching during Year 1.

In the majority of the schools, the intervention was implemented by TAs and, as the project shows, the outcomes were positive. This is in line with a previous EEF review about how to engage TAs more effectively in educational settings which indicates that interventions that are focused and well-structured, such as the intervention evaluated in this project, make the best use of TAs' time. It is noted, however, that this does not mean that the TAs do not have a role in the intervention. There is strong evidence in this project to indicate that the apps may not be effective in raising attainment by themselves and that the TA plays a crucial role in its successful deployment. This is a significant lesson to be learned for future studies.

Limitations

On the whole, we are satisfied with the information and the conclusions presented in this evaluation report. The evidence provided by this project is robust. The attrition rate was low, and the differences between the intervention and control groups in attrition were small and not significant statistically: there were very few breaches of the protocol, and more than 60% of the pupils received a dosage of the intervention that was considered high by the intervention team at the protocol stage of the project. We conclude that the project offers a fair evaluation of the impact of the apps.

However, the results of the implementation and process evaluation suggest that the variation in the amount of pedagogical support offered by the adult who supervises the intervention is key to its success in improving the pupils' performance in the outcome measure. After having observed a great deal of variation in the amount of pedagogical support offered by the adults during the sessions, the evaluation team conjectured about the possible consequences of this variation. If it had been anticipated that the variation in pedagogical support might mediate the impact of the intervention, this factor could have been part of the design and could have been systematically manipulated. Thus caution must be exercised if one wishes to establish a causal link between the role played by the TA and the level of efficacy of the apps in improving the pupils' outcomes.

The consequence of the variations noted across schools in the implementation is that one must be cautious in generalising the results of this project. If a similar RCT were to be carried out and the intervention were implemented by a majority of TAs who did not think that they had a role to play in promoting the pupils' learning, which would result simply from using the apps, a different outcome would be considered likely by the evaluation team. Nevertheless, it is impressive that a definite intervention effect was found despite these variations in the implementation of the intervention.

A second limitation lies in the use of the apps in addition to, rather than instead of, regular teaching. This design, which is related to the number of hours on the app considered necessary for a significant impact, cannot exclude the possibility that the improvement results from additional maths experience in small groups. Therefore, the results should not be generalised beyond these conditions of implementation.

Finally, we should repeat two limitations already mentioned in the section on 'Outcome measures'. One was that the absence of common items in the pre- and the post-tests made it impossible to measure progress in the performance of the participants. The design does allow for a robust conclusion that the apps, as used in this project, had an impact on the pupils' performance in the outcome measure but does not allow for measuring their progress. The second, more serious problem, is that the post-test contained many items that were very close to what was taught to the group in the app-based intervention. This limits the generalisation of the findings because other measures of mathematical achievement might not have led to similar results.

Future research and publications

The preceding sections strongly indicate that future research should investigate systematically the extent to which pedagogical support is crucial for the impact of the apps to be observed. This is a school-level factor and could be manipulated in an RCT with some confidence.

Another urgent question that needs addressing is why the apps showed no sign of any positive impact on pupils eligible for FSM. Were the pupils less likely to have had previous experience with iPads and headphones and thus needed to make greater adjustment to this medium of instruction? Were they more attracted by irrelevant factors because of the novelty of iPads? Pupil-level factors are more difficult to manipulate, so more qualitative information about the pupil factors should be collected in further studies. In the present study, pupils were already being taken out of the classroom for substantial periods, which made it unwise to ask for more of their time, but future studies could benefit from a closer analysis of how pupils from different backgrounds use this sort of medium for learning.

The evaluation team intends to publish the impact analysis and the outcomes of the process evaluation in separate journal articles. A third publication will describe case studies of schools where the TAs perceived their role differently and thus played different roles during implementation. The intervention team will have the option to co-author the publications.

References

- Borenstein, M., Hedges, L.V., Higgins, J.P.T. and Rothstein, H.R. (2009) *Introduction to Meta-analysis*. Chichester: Wiley.
- Cheung, A.C. and Slavin, R.E. (2013) 'The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis.' *Educational Research Review*, 9, 88–113.
- Delgado, A.J., Wardlow, L., McKnight, K. and O'Malley, K. (2015) 'Educational technology: A review of the integration, resources, and effectiveness of technology in K-12 classrooms.' *Journal of Information Technology Education*, 14, 397–416.
- Department for Education (2018) 'Schools, pupils and their characteristics: January 2018.' Available at: https://assets.publishing.service.gov.uk/.../Schools_Pupils_and_their_Characteristics_2. (accessed November 2018)
- Field, A. (2009) *Discovering Statistics Using SPSS*. London, Sage Publications.
- GL Assessment (2015) Progress Test in Maths 5. London: GL Assessment.
- GL Assessment (2015) Progress Test in Maths 6. London: GL Assessment.
- Haßler, B. *et al.* (2016) 'Tablet use in schools: A critical review of the evidence for learning outcomes.' *Journal of Computer Assisted Learning*, 32(2), 139–156.
- Holmes, W. and Dowker, A. (2013) 'Catch Up Numeracy: a targeted intervention for children who are low-attaining in mathematics.' *Research in Mathematics Education*, 15(3), 249–265.
- Hox, J. (2002) *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Erlbaum.
- Hubber, P.J. *et al.* (2016) 'Should touch screen tablets be used to improve educational outcomes in primary school children in developing countries?' *Frontiers in Psychology*, 7: 839.
- Kahan, B.C. and Morris, T.P. (2012) 'Improper analysis of trials randomised using stratified blocks or minimisation.' *Statistics in Medicine*, 31(4), 328–340.
- Kale, U., Akcaoglu, M., Cullen, T. and Goh, D. (2018) 'Contextual factors influencing access to teaching computational thinking.' *Computers in the Schools*, 35, 69–87.
- Kang, M., Ragan, B.G. and Park, J-H. (2008) 'Issues in outcomes research: An overview of randomization techniques for clinical trials.' *Journal of athletic training*, 43(2), 215–221.
- Maas, C.J.M. and Hox, J.J. (2005) 'Sufficient sample sizes for multilevel modeling.' *Methodology*, 1(3), 86–92.
- McGee, P. and Reis, A. (2012) 'Blended course design: A synthesis of best practices.' *Journal of Asynchronous Learning Networks*, 16, 7–22.
- Miller, T. (2018) 'Developing numeracy skills using interactive technology in a play-based learning environment.' *International Journal of STEM Education*, 5(1), 39.
- Nunes, T., Bryant, P., Strand, S., Hillier, J., Barros, R. and Miller-Friedmann, J. (2017) *Review of SES and Science Learning in Formal Educational Settings*. London: Education Endowment Foundation and the Royal Society <https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/science/>
- Outhwaite, L.A. *et al.* (2017) 'Closing the gap: Efficacy of a tablet intervention to support the development of early mathematical skills in UK primary school children.' *Computers & Education*, 108, 43–58.
- Outhwaite, L.A. *et al.* (2018) 'Raising early achievement in math with interactive apps: a randomized control trial.' *Journal of Educational Psychology*. Advance online publication. <http://dx.doi.org/10.1037/edu0000286>.

Pitchford, N.J. (2015) 'Development of early mathematical skills with a tablet intervention: a randomized control trial in Malawi.' *Frontiers in Psychology*, 6, 485.

Pitchford, N.J. *et al.* (2018) 'Interactive apps promote learning of basic mathematics in children with special educational needs and disabilities.' *Frontiers in Psychology*, 9, 262.

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

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

Suresh, K. (2011) 'An overview of randomization techniques: an unbiased assessment of outcome in clinical research.' *Journal of Human Reproductive Sciences*, 4(1), 8–11.

Wechsler, D. (1992) *WISC-III UK Manual*. Sidcup, Kent: The Psychological Corporation, Harcourt Brace Jovanovich.

Williams, A. (2003) 'Informal learning in the workplace: A case study of new teachers.' *Educational Studies*, 29(2–3), 207–219.







Worth, J. *et al.* (2015) *Improving Numeracy and Literacy. Evaluation report and Executive summary*. London: Education Endowment Foundation.

Appendix A: EEF cost rating scale

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

| Cost rating | Description |
|-------------|--|
| £ £ £ £ £ | <i>Very low</i> : less than £80 per pupil per year. |
| £ £ £ £ £ | <i>Low</i> : up to about £200 per pupil per year. |
| £ £ £ £ £ | <i>Moderate</i> : up to about £700 per pupil per year. |
| £ £ £ £ £ | <i>High</i> : up to £1200 per pupil per year. |
| £ £ £ £ £ | <i>Very high</i> : over £1200 per pupil per year. |

Appendix B: Security classification of trial findings

| Rating | Criteria for rating | | | Initial score | | Adjust | | Final score |
|---|---|----------|------------|---------------|--|--|--|-------------|
| | Design | Power | Attrition* | | | Adjustment for balance [0] | | |
| 5  | Well conducted experimental design with appropriate analysis | MDES<0.2 | 0–10% | | | Adjustment for threats to internal validity [0] | | |
| 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% | | | | | |
| 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% | | | | | |
| 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 = five padlocks, well-conducted design, MDES at randomisation was 0.19 and attrition rate was low at 3.11%
- **Reason for adjustment for balance** (if made): Pre-test differences is low (mean = –0.05) and slightly in favour of the control group. No adjustments were made.
- **Reason for adjustment for threats to validity** (if made): Low attrition therefore threats to internal validity due to missing data is low. Outcomes is a standardised measure; no concurrent interventions or confounders were reported. No adjustments were made.
- **Final padlock score:** Initial score adjusted for balance and internal validity = five padlocks.

Appendix 1: Protocol

Department of Psychology, University of Nottingham Department of Education, University of Oxford

Prof Terezinha Nunes



Protocol for Evaluation of the onebillion app for improving mathematics learning in the early years

| Evaluation Summary | |
|--------------------------|---|
| Age range | Year 1 (age 5 to 6 years) |
| Number of pupils | 1124 (567 in the intervention group) |
| Number of schools | 113 (57 in the intervention group) |
| Design | Cluster randomised controlled trial |
| Primary Outcome | Progress Test in Maths (GL Assessments) |
| Protocol date | 14 March 2018 |
| Protocol version | 1 |

BACKGROUND

Intervention

This evaluation will test the impact of the *onebillion maths apps* (henceforth referred to simply as the intervention) on pupils' numeracy outcomes. The intervention is a curriculum based intervention, rather than a theoretically motivated programme, and includes two levels, one labelled as 'age 3-5 app' and the second labelled as 'age 4-6 app'. The maths 3- 5 app contains 10 modules and the maths 4-6 app contains 18 modules aligned with the English National Curriculum. Each module has several activities and a quiz. There are a total of 178 activities across both apps. Modules in the 3-5 app contain 7 activities, while those in the 4-6 app contain 6. The 28 modules cover several topics, for example, counting (with activities organised in different levels, taking counting up to 100), classification by different criteria (shape, colour), shape (geometrical shapes vocabulary, symmetry), lines and patterns

(straight or curved; repetitions of figures in a pattern), position (vocabulary about spatial relations), measures (length, time, mass and capacity), addition and subtraction (arithmetic with pictures, number bonds and number line work), sharing and fractions (half and quarter). There is no overlap in the activities in the 3-5 app and that for 4-6 year olds, and so the two apps can be viewed as one progressive sequence. For example, counting and learning numerical symbols in the 3-5 app reaches 10 and the 4-6 app starts with counting to 20 and continues to 100. In this trial, pupils will start with the 3-5 app and move on to the 4-6 app if they complete the 3-5 app. The intervention will thus be individually paced; this is taken to reflect the way that the apps would be used when schools adopt them outside a research project.

Although pupils work individually, pupils will be working in small groups in the same room at the same time and will be supervised by a nominated member of staff, which can be a teacher or a teaching assistant (TA); for brevity, the member of staff will be referred to as TA. The number of pupils in the group will be decided by the school. The displays are designed to be attractive and teaching is part of the displays, which include a voice (the teacher in the app) that provides explanations orally about what the child has to do. These instructions are repeated if the child presses the appropriate button on screen, a feature that allows the children to control the number of times they want to hear an explanation regarding what to do in the activity. Feedback is given after the child's answer by a sound if there is a mistake, and by a tick and cheers if the answer is correct. Children are encouraged to work through the activities in each module in order to master them; the activity to be attempted is indicated by its flashing on the screen. However, the apps are not closed and children can access a different activity. When they have completed the different activities, they are presented with a quiz, built into the app, that tests their knowledge of the materials covered. If they have answered all the questions correctly, they receive a certificate in the app. Otherwise, they can either do the quiz again or they can go back to the activities they failed on the quiz, repeat these, and then do the quiz again in order to receive the certificate. Once they have passed the quiz they can move on to the next module, but the app does not block their access to other activities if they have not passed the quiz. The child is instructed by the teacher voice to work on the next activity, signalled by flashes on the display, but the app does not restrict the child's access to that activity. For this reason, the intervention is described by Outhwaite et al. (2017) as "individually paced".

The role of the TA in this intervention trial is to manage and support the pupils in using the tablets and to track pupil participation and progress during the intervention. TAs were requested to use a specifically designed register and a chart of received certificates. As this is a curriculum based intervention, it targets activities designed to teach concepts and vocabulary to children aged 3 to 6 years (considering both apps as a sequence). Some activities reflect tasks taken from developmental psychology research (e.g. the give N task, placing pictures of events in logical order) and some of the activities are ordered according to results in developmental psychology (e.g. classification by a single criterion before classification by two criteria; addition and subtraction with objects before addition and subtraction with symbols). Other activities seem aligned with the curriculum but, to our knowledge, there is no research to indicate whether there is a particular order of acquisition

learning about odd and even numbers before learning how to count to 50).

There is no indication so far regarding whether the best use of the intervention is as a preparation for learning in the classroom or as a reinforcement of what has been already learned in the classroom. As it is an individually paced programme, it is likely that the synchronisation of classroom instruction and the practice with the app will vary across children. Some children might use the activities in the apps before the relevant instruction and others afterwards; this trial is not designed to test whether this affects the impact of the intervention.

The intervention aims to promote the learning of facts, vocabulary, and conceptual understanding of topics which are part of the English National Curriculum (e.g. the counting sequence; addition and

subtraction facts; labels for geometrical figures, spatial relations and comparisons) and draws on a range of learning processes, including instructional psychology's "model, lead, test" sequence. The examples at the start of an activity model what the child is expected to do; the child then works through these activities and is tested in a quiz at the end. Different ways of modelling and explaining are embedded in the activities; the trial is not designed to test whether the different types of instruction have different impacts.

Significance

The intervention aims to complement current teaching practice by offering children individually paced additional opportunities to rehearse materials that are part of the curriculum. In view of its potential to offer additional experiences with curriculum materials to a large number of children, it is important that it should be systematically evaluated using an RCT. In this trial, teachers will nominate children whom they consider to be struggling with maths in the first term in Year 1 to participate in the project. This is because these children are likely to benefit most from the additional opportunities to rehearse materials related to the curriculum. Thus, the significance of the evaluation is the assessment of its efficacy for pupils who are struggling with numeracy as they start primary school. Previous research used a 6-week, 12-week, or a 13-week training (Outhwaite et al., 2017, under review); the longer intervention produced stronger impact. In this trial, the intervention will be tested over a 12-week period, from the second half of the Spring term to the beginning of the second half of the Summer term.

RESEARCH PLAN

Research questions

The primary research question is:

- ✓ Do the children identified by their teachers as struggling with mathematics at the start of Year 1 who use the onebillion apps show better performance in Progress Test in Maths (PTM) than children also identified by their teachers as struggling with mathematics at the start of Year 1 who do not use the apps?

Secondary research questions:

- ✓ Do children, who have been entitled to FSM, benefit to the same extent as other children from using the onebillion apps as assessed by the PTM?
- ✓ Is the onebillion apps programme equally effective for girls and boys as assessed by the PTM?

Design

The design is an RCT, with two trial-arms, an intervention and a control group, and a pre- and a post-test. 113 primary schools (1124 pupils; 552 girls) were recruited to participate in the trial. The apps will be used in addition to normal classroom numeracy teaching. Schools were eligible to participate if they had at least 15 children in Year 1, had not used the apps before and have a sufficient number of iPads to implement the intervention with small groups of children.

Randomisation was implemented at school level. The school was chosen as the unit of randomisation to avoid the contamination that could take place in a within-school allocation. In schools that have more than one Year 1 class, only one class was randomly chosen to participate in the intervention.

Because the intervention is believed to be more effective for low achieving children, the teacher in the randomly selected Year 1 class nominated children for participation in the trial. The Nottingham University intervention team provided written instruction to teachers in all schools on how to nominate the children for the project: the children should be in the lower half of their class, according to the teacher's assessment, not have a statement of special educational needs, and have no difficulty in understanding English. If a class had 19-20 children, 10 children were nominated; if a class had 17-18 children, 9 children /were nominated; if a class had 15-16 children, 8 were nominated. If the school had more than 14 Year 1 children and these were distributed across different classes, all Year 1 children were treated as a single cohort and the teachers from the different classes cooperated in the nomination process. The list of nominated children was sent to the Oxford University evaluation team by the 18th January 2018, before pre-testing and randomisation; 6 schools nominated 9 children and the remaining schools nominated 10 children.

Data collected at nomination included the child's name (which will be removed from the data set and replaced with a project identifier), gender, date of birth, unique pupil identifier (UPN), school, and eligibility for FSM. Parents could allow their children to participate in the project but withhold the information on UPN and FSM eligibility status. After nomination, pre- and post-test results are added to the file as well as FSM status as recorded in the National Pupil Database (NPD).

The design includes a pre- and a post-test, using parallel forms of PTM. Although PTM is designed for administration to whole classes, in view of the children's age, the evaluation and the intervention teams agreed that individual administration would produce more valid results with such young children. The evaluation team trained testers to implement this individual administration and checked the adaptation with the test provider, GL Assessment. Quality control of this administration was based on the observation of a sample of testers (50%) during administration of the test to one group of children.

After administering the pre-test, schools were randomly assigned either to the intervention or to the control group. In order to join the project, heads of schools signed an agreement with the Nottingham team (see Memorandum of Understanding in the subsequent section) indicating that they would accept their random assignment. If assigned to the intervention group, they would provide TAs the necessary conditions for implementing the intervention. If assigned to the control group, they would continue with their usual methods of supporting children struggling with maths. As an incentive, control schools were offered the possibility of accessing the apps at the end of the project and using them with the new cohort of Year 1 pupils.

TAs in schools assigned to the intervention group were invited to participate in the training for implementation of the intervention. The training included: how to find a suitable time in the daily timetable to administer the intervention, how to prepare the tablets for use (downloading the apps, registering children, familiarisation with the apps and their interactive features, technical trouble shooting), advice on offering pedagogical support (limited in this trial), how to record the daily information on participation and quizzes passed, as well as the technical and pedagogical support offered by the intervention team. This information was given at the training events but was also made available online in a private iTunesU course that was open only to TAs delivering the intervention. The iTunesU course has seven demonstration videos; a pdf of the implementation manual was also made available to the TAs. TAs also have access to a forum where they can share best practice and ask questions to other TAs and to the Nottingham intervention team. The Nottingham team communicated to the schools that they needed to attend the training session for their region. If that was not possible, they could notify the team and attend

a training session at another region. Those schools that found it completely impossible to attend a training session were asked to arrange a phone call and follow the on-line training. The Nottingham team would then check whether they had accessed the online training. Records of attendance were provided to the evaluation team.

The intervention will be implemented for half an hour, four days per week, during 12 weeks. The intervention team recommends for this trial that all children should start with the maths 3-5 app and progress to the maths 4-6 app, once they have completed the 3-5 maths.

Pre-tests were administered in January and the first week of February 2018 by testers trained and under the supervision of the evaluation team. Randomisation was conducted before February half term by the evaluation team; notification to schools was sent in the same week by the intervention team. Schools had been notified previously about the timing of the training; after randomisation, schools assigned to the intervention group were immediately invited to participate in the training, which took place after February half-term so that the intervention could start in the subsequent week. The intervention will be completed over 12 weeks and post-test will take place immediately after the end of the intervention, in June and July.

Participants

School recruitment

School recruitment was carried out by the Nottingham intervention team, with support from the evaluation team, across four Target Regions: 1) East Midlands; 2) West Midlands, 3) Greater Manchester and North West, and 4) South and West Yorkshire. Seven schools outside these regions were also allowed to join the project (3 in Cumbria, 3 in Oxfordshire, and 1 in Milton Keynes). Local authorities in these regions are listed in Appendix 1. Schools were recruited by means of the following strategies: EEF Website; EEF Twitter; University of Nottingham Project Website; E-mails to schools through Apple distinguished educators (ADE) network and Maths Hubs Network; emails to key contacts in Local Authorities through Educational Psychology networks; School Recruitment Events. The Oxford evaluation team supported the intervention team with advice regarding all aspects of recruitment, including preparation of materials for inviting schools, the process of registration and the design of the Memorandum of Understanding (MoU), and by emailing schools that had been part of previous projects implemented by the Oxford team. Schools in regions not initially included in those indicated by the Nottingham team, which were approached by either team, were also accepted into the project.

The MoU made it explicit that schools agreed to accept their random allocation either to the intervention or to the control group, to collaborate with the evaluation team, to facilitate pre- and post-tests, to provide registers and records of children's progress through the certificates and access for process evaluation, and to answer questionnaires and phone interviews, as required. Schools allocated to the control group agreed to continue with business as usual in the support of children with lower levels of attainment in maths and those in intervention schools agreed to provide suitable conditions for the implementation of the intervention. Control schools were offered a financial incentive of £1,000 in order to cooperate with the evaluation team up to the end of the trial and the further incentive of receiving free access to the app once the project was completed. Control schools were asked to restrict the use of the apps only to children not participating in this project.

Pupil recruitment

See design section for details on how pupils were selected to participate.

Randomisation

After the pre-test of the nominated children was concluded, the schools were randomly assigned either to the intervention or to the control group by the evaluation team, with an equal allocation of schools to each group. Random numbers were generated for all schools using SPSS. Schools were ordered by these random numbers in ascending order. Schools that received the lowest random numbers were allocated to intervention group and the schools with the highest random numbers were allocated to control. The syntax used was:

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

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

Outcome measures

Primary outcome- Children's attainment

PTM was chosen for this trial because it is a test of pupils' attainment in the topics included in the National Curriculum. PTM 5 was used for the pre-test and PTM 6 for the post-test. The tests contain 20 items each and cover concepts similar to those taught in the intervention

(e.g. height, numbers – ordering and recognition - and simple arithmetic, comparisons between sets and objects, spatial relations). There is no time limit but it is estimated that

individual administration takes approximately 20-25 minutes. According to the test providers

(GL Assessment, Technical information), the tests have good internal consistency (Cronbachs' Alpha for PTM5=.87 and for PTM6=.9). Gender differences are small (girls had a raw score 2.3 point above boys in PTM5 and 0.3 point below boys in PTM6). The tests have been validated by correlations with PiM (Progress in Maths), which is the predecessor of PTM and which correlated with KS assessments; for PTM5 the correlation with PIM5 was

0.62 and for PTM6 the correlation with PIM6 was 0.78. The intervention team found previously a correlation of 0.67 between PTM5 administered individually at pre-test and post- test with 4-5-year-old children (Outhwaite, Faulder, Gulliford, & Pitchford, 2018).

Pre-tests were carried out prior to randomisation by testers, who received training for implementing the assessment from the evaluators at Oxford University. Post-intervention testing will be administered by testers trained by the evaluation team, blinded to the school's group allocation. A protocol has been developed to train the testers on how to approach the schools without identifying their group membership, how to approach the children at the start of the testing procedure, and how to provide clarification in standardised ways, if children ask questions. The training also includes ethical guidelines and instruction on how to anonymise the tests before they are posted to the evaluation team. These procedures were approved by the Oxford University Ethics Committee.

Sample size calculations

The aim at the start of the project was to have power to detect an effect size for intervention relative to control equal to 0.18 SD. This seemed reasonable given that a previous evaluation in the UK using a prior version of PTM showed a Cohen's d effect size of 0.31 (CI

= 0.06 - 0.55) after 12 weeks of implementation of the app, when it was used in addition to normal classroom practice (Outhwaite et al., 2018, in press). Considering that this design is

essentially the same used in the prior trial and a similar test will be used, the aim of detecting

an effect size which is considerably smaller than that observed in the previous study was considered a conservative estimate. It was subsequently decided to explore the number of schools required to detect an effect size of 0.2 for different correlation coefficients. Pitchford (personal communication) reported a pre- to post-test correlation in their previous study to be .67. Worth et al (2015) do not report the correlation they observed in a previous EEF supported project using the previous versions of PTM 6 and 7, but Nunes (personal communication) has calculated it and found it to be equal to 0.75. It was decided to calculate the power for this trial using two estimates of this correlation: $r=.5$ and $r=.7$.

Optimal Design software was used to explore the number of schools required for the trial in two different scenarios defined by these two levels of correlation. The calculations relied on the following assumptions: (i) Cluster Randomised Trial with person level outcomes; (ii) pupil outcomes measured at pre-test and at post-test have a correlation of $r=0.7$ at pupil level for one calculation and of $r=0.5$ for the second calculation; (iii) the same correlation for a level 2 analysis; (iv) a within-school sample of 10 pupils per school; (v) an intra-class correlation coefficient of 0.15 (estimated by the DfE as the intra-class correlation in mathematics assessments¹); (vi) power of 0.80, alpha of 0.05 and a 2-tailed significance test. Table 1 displays the results of these calculations.

Table 1: Number of schools required to detect an effect size equal to 0.2 with different levels of correlation, power of .8 and alpha= 0.05, two-tailed test

| Number of schools | Number of pupils (10 per school) | Pre and post-test correlation |
|-------------------|----------------------------------|-------------------------------|
| 128 | 1280 | 0.5 |
| 104 | 1040 | 0.7 |

The EEF decided that the target for recruitment would be 104 schools and that recruitment would be defined by the signing of MoUs and the subsequent nomination of pupils to participate in the trial.

At the deadline for recruitment, 115 schools met these criteria but 2 schools withdrew before pre-test and randomisation; 6 schools had smaller cohorts and nominated 9 pupils (as agreed in the nomination procedures) and the remaining schools nominated 10 pupils, so the total number of pupils nominated is 1124.

A new power calculation was implemented using PowerUp (Dong & Maynard, 2013) to calculate the minimum detectable effect size (MDES). Appendix 2 presents the calculation for a pre- and post-test correlation $r=0.7$ at levels 1 and 0.63 at level 2. This calculation estimated the MDES for this sample and with these assumptions as 0.19.

There are in the sample 286 pupils in 88 schools who are eligible for FSM. 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 (3.25 in this sample) to calculate the minimum detectable effect size. Appendix 3 presents the power calculation for the minimum detectable effect size for the pupils in the sample who are eligible for FSM, who can be included in the subgroup analysis. When the proportion of schools in this subgroup that was assigned to the control and the intervention group was calculated, this turned out to be almost identical as that in the complete sample (51%). The calculation using PowerUp estimated the MDES for the subgroup analysis including only pupils eligible for FSM as 0.29.

Analysis plan

Pupil performance in the PTM6 at post-test will be the primary outcome; raw scores will be used in the analyses. Analyses will be conducted in SPSS/MLwiN and R (using the EEF analytics) using 2-tailed significance tests at the 5% significance level and will include all the data available, according to an intention to treat model.

ANCOVA will be used to compare intervention and control groups on the post-test scores, controlling for pre-test scores; a multilevel model with two levels (pupils within schools) will be used to account for possible clustering at the school level.

The primary analyses will be intention to treat and will include the maximum number of participants. Reasons for missing data will be investigated and, if a high number is observed, possible biases will be investigated. The effect size will be calculated using ANCOVA controlling for pre-test scores, to increase precision and power. Hedge's *g* will be used to indicate the effect size and will employ the total pupil variance; the confidence interval will be reported using the traditional 95% interval. The intra-cluster correlation will be reported for pre- and post-test. The details of the model will be included in appendices to the report.

Additional analyses will consider the impact when compliance is taken into account. Measures of pupil compliance will be based on the levels of compliance defined by the intervention team (see pupil measures in the section about implementation and process evaluation).

Subgroup analyses

Schools were asked to provide pupils' names, date of birth, gender, Unique Pupil Number (UPN) and FSM status. As these are pupils in Year 1, there is no difference between current eligibility for FSM or eligibility as defined by the National Pupil Database (NPD) everFSM variable, which takes into account eligibility in the last six years. The information on eligibility for FSM will be provided independently also by the NPD for confirmation (NPD variable EVERFSM_6). A separate analysis will be completed to test for the interaction between treatment and EverFSM. Analyses using the subgroup defined as pupils eligible for FSM (EverFSM in the NPD) will be carried out, comparing the intervention and the control groups, but the results must be taken with caution as the number of children in the analysis using pupils eligible for FSM will be reduced, which will have implications for significance levels.

Implementation and process evaluation

The focus of the process evaluation will be to assess the fidelity of the programme, to understand the conditions that make the intervention successful and to understand what business as usual in the control schools means. Prior to designing the instruments for process evaluation, the evaluation team obtained from the intervention team their logic model and their criteria for treatment fidelity. According to the intervention team, the most important fidelity measure is a measure of time using the app. In the handbook distributed at training, the intervention team asked TAs to try to make up for missed sessions by rescheduling them. The intervention team does not discriminate between consecutive missed sessions or missing sessions in different weeks, because the children work through the apps at their own pace.

Pupil measures

In order to assess compliance with number of sessions, the intervention team has asked the TAs to fill in a register on each day of the week, which indicates whether the child was present, the app with which the child worked, the number of certificates attained by the child during the session, and the time of the session. TAs were asked to note under comments if a child interrupted a session. TAs were also asked to note whether each child required technical or pedagogical assistance during the session. Each child's record will be matched to the child's identification number in the project to allow for an analysis of compliance.

Participation will be measured in three different ways.

- **Stopping point:** The definition of stopping point is the highest number of modules in sequence completed in the apps. Maths 3-5 has 10 topics and maths 4-6 has 18 topics, and so the maximum number of topics is 28.
- **Exposure:** Exposure will be measured by the number of sessions that the child attended. The planned number of sessions is 48 (4 times per week during 12 weeks). Children may occasionally miss sessions for different reasons. The intervention team identified three levels of compliance for this trial: 1) low compliance, defined by participation in up to 30 sessions (62.5% of the sessions in this trial) which is equivalent to 6 full weeks of intervention delivered every day; 2) medium compliance, defined by attendance between 31 and 40 sessions; and 3) high compliance, defined by attendance to at least 40 sessions (83.3% of sessions). Outhwaite et al. (2017) found that the level of participation equivalent to low compliance in this trial significantly decreased the impact of the intervention.
- **Success in the quizzes:** The number of certificates achieved by the child will give another measure of participation. Attendance to the sessions is a necessary step to time on task, but the children may be at the session without fully engaging with the apps or may take longer in the activities and thus complete fewer quizzes. The registers obtained by TAs will provide information on certificates of 100% correct in the quizzes obtained in each session to complement the register of attendance. It is noted that this measure might be correlated with ability: more able children may succeed in more quizzes, and this would be a source of confounding. However, the use of the pre-test as a covariate might account for this relation between number of quizzes mastered and ability, and thus avoid the confounding. Further details on how this metric will be analysed can be found below.

Analyses in the presence of non-compliance will investigate whether these different dosages of the intervention show differential effect. Multilevel models will be used with the treatment defined in four levels: no treatment (control group), low dosage, medium dosage and high dosage. The multilevel models will take into account the nesting of children in schools and will include the pre-test as covariate, as in the previous models. It will be investigated whether different effect sizes are obtained with different dosages and whether these different levels of compliance are statistically significant when compared to the no-treatment condition and to each other. This allows for including all participants in the same analysis in order to retain power.

Finally, it is also possible that children are present at the session, but spend their time doing other activities. A sample of observations will be collected in order to assess time on task for the children in the session. This analysis cannot be applied to all children, but it can potentially highlight how much of the scheduled half hour is spent on average by the children effectively using the apps. This will be reported in the implementation and process evaluation section with reference to the analysis completed.

Other measures of fidelity

In order to assess other aspects of implementation, above and beyond pupil time with the app, the evaluation is collecting data on other aspects of implementation: training of TAs and implementation during the sessions.

Due to uncontrollable circumstances (road closures due to snow), the intervention team had to cancel one of the training sessions. Some TAs were trained in face-to-face meetings and some using the iTunes videos. The face-to-face training sessions were observed in order for the evaluation team to describe this element of implementation. This description will inform the process evaluation but will not be analysed quantitatively.

TA questionnaire

TAs were asked to answer a questionnaire about the training and about their schools' use of IT with young children. TAs training using the iTunes videos were asked to answer a questionnaire that paralleled as much as possible the one used in the face-to-face training sessions. The questionnaire contains questions about the session itself (e.g. whether the TA felt that enough time was dedicated to all the elements of the training, whether they understood the structure and content of the apps) and questions about the TA's and their schools' previous use of IT.

This questionnaire allowed the evaluation team to identify three levels of previous use of IT with young children in schools (low, medium and high) and different forms of training for use of the app (face-to-face plus video based versus video based only). The combination of these two dimensions leads to six cells, which will be the initial basis for the choice of schools in which the evaluation team will carry observations of a sample of implementation sessions; 12 sessions will be observed, with two observations per cell. The schools will be purposefully selected to illustrate a variation in proportion of nominated pupils eligible for FSM, because it is possible that pupils eligible and not eligible for FSM have different levels of previous familiarity with iPads, which could influence how smoothly the sessions run. The aim of these observations is to average time on task (i.e. subtract from the half hour the amount of time required to set up and to tidy up at the end), record variations in children's need for support in the use of the iPads and in TAs' expertise in addressing the children's technical and pedagogical needs during the sessions. The observations will be followed by brief interviews with the TAs to describe their understanding and confidence in their ability to play their role in this intervention.

After 8 weeks from the start of the intervention, all the TAs will be asked to answer a questionnaire about the implementation to provide information on the material conditions effectively used, on how well the intervention fits with their schools' aims and schedules, and on how they perceived their role and how

often they felt the need to intervene and mediate the children's use of the app. The questionnaire will also include questions about the number of children in each session as there might be variation in the size of the small groups to which the implementation is delivered. Information on how the sessions are distributed during each week will be obtained from the registers. The intervention team suggested during the training sessions that it was best for sessions to be scheduled for a half hour on different days rather than to schedule two sessions on the same day, although it was agreed that two sessions on the same day would be a better option than missing out a session. This will be treated as a fidelity factor to be analysed as part of the implementation and process evaluation.

Middle-management questionnaire

The evaluation team will also use a questionnaire for a middle management member of staff to provide information about costs and the fit of the intervention with the school's aims and schedules. The appropriate person to answer the questionnaire will be identified by the link teacher nominated for the project. This will be presented to the schools from week 8 of implementation onwards in order to obtain information based on what has taken place rather than before the school has experienced the intervention. The questionnaire will also include questions about the previous use of IT in the school in order to describe the context in which the intervention took place.

A middle management member of staff in intervention and in control schools will be asked to describe what interventions have been used with the children nominated for participation in the project, the content and duration of these interventions, if any, and who was responsible for the implementation. For intervention schools, these questions will be part of the same questionnaire used to collect data on costs.

Research questions to be addressed by the process evaluation

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

- ✓ Does fidelity to treatment moderate the effectiveness of the onebillion intervention? A

secondary research question to be addressed is:

- ✓ To what extent do control schools use alternative treatments that involve the same contents and the same amount of resources as in the intervention schools?

Analysis of factors described in the implementation and process evaluation

The different sources of data described in the previous section will be used in these analyses. Information on exposure, the stopping point at the end of the intervention, and the number of quizzes answered successfully will be analysed as variables that possibly moderate the impact of the intervention. Registers of children's participation and the stopping point for each child will be noted to provide a measure of participation. Each of these measures will be analysed separately because the apps are individually paced, so that children who had the same number of hours of exposure might have achieved different levels in the apps and reached different stopping points. The measures of pupil participation will be used as indicators of dosage and can be entered in the multilevel models as predictors of the primary outcome.

The TA questionnaires will provide an indicator of the context in which the interventions took place and we will seek to investigate whether the context of the intervention affects its implementation. These analyses will be carried out with the intervention group only, and will assess whether TAs' responses can be seen as mediators of the outcomes (e.g. TAs' knowledge and confidence in the intervention; TAs' perceived efficiency in managing time; the material conditions of delivery – e.g. whether there was a dedicated space, or sufficient iPads for the delivery in a single group). The TA questionnaire will also obtain information on the number of children that participated in the intervention at the same time, because the specification by the intervention team was that in this trial the apps would be used in small groups. Although the intervention team does not include any material conditions in their logic model of fidelity of implementation, it is possible that the intervention works best if the children are in an environment relatively free of distraction and the number of iPads available in the school allows for efficient planning of the sessions.

Although the intervention is delivered through an app and does not require participation of the TA beyond monitoring the children's work, it is unlikely that the TAs will have no interaction with the children. Information on their knowledge of the app and their interaction with the children will be obtained through observations and interviews. These will be analysed qualitatively to allow for learning lessons about implementation for the future.

Observations will provide information on TAs compliance with the guidance provided by the intervention team regarding how to set up the environment, how to deal with children's technical and pedagogical difficulties, and with the fact that children might interact with peers during the sessions. Observations will also be used to investigate whether differences in group size affect the TA's response time when the children require support.

Phone interviews with middle management staff (n=10) will be used to clarify the fit of the app with the school's aims and schedules. The fit with the school's aim and schedules has been found to be a significant aspect of implementation success in science education interventions.

Cost evaluation

The cost evaluation will be calculated as if the school had been paying the entire costs of delivering the intervention, including purchase of ten tablets and the cost for downloading it. Questions will be posed to the intervention team as well as to the schools.

The intervention team will be asked about the cost of downloading the app, the fees charged to schools for training (if any), the equipment required, and the time that staff is expected to spend in training, when and where training is normally provided, as well as time required for implementing the intervention. As the intervention team has developed an iTunes training cost, the evaluation of the effectiveness of this training will contribute to the estimates of cost for the future (i.e. access cost and time taken to watch the videos).

This information will be complemented by questionnaires with middle-management school staff to describe the cost of resources that the TAs actually required for the implementation of the intervention (e.g. time delivering the intervention; time spent on preparation of the physical environment). In the onebillion intervention, an obvious question is whether the school already had tablets that were suited for the intervention or whether they had to be acquired for the purpose of this intervention; whether this investment would be a normal part of the school's plan even in the absence of this intervention; whether there was a need for additional hours in preparation or a need to cancel other activities normally scheduled which use the same resources.

The cost estimate will be initially calculated per school and will then be divided by the average number of pupils in the school who completed the intervention. The estimate of cost per pupil will be based on costs over three years, including one-off and recurring costs; the number of pupils per year will be based on the number of iPads available and estimates of available TA time for supervision. Differences in group size for delivery of the sessions will be taken into account, if these are observed.

ETHICS AND REGISTRATION

The trial was designed and will be conducted and reported to CONSORT standards and adhering to Ethics and data protection regulations from the Oxford University Ethics Committee and the University of Nottingham. The evaluation team obtained ethical approval for the trial from the University of Oxford Central Research Ethics Committee on 16 November 2017 (Application Approval: ED-CIA-17-014). Opt-out forms were used. When uploading the pupil nomination, TAs were asked to confirm that they had not uploaded information about children whose parents had opted out of the trial or UPNs and FSM status of children whose parents opted out of providing this information.

Schools obtained parental consent for participation in the trial; heads of schools agreed to this procedure when they signed the MoU (see Appendix 4 for MoU and parent information letters). If a nominated child were to withdraw from the trial before randomisation, schools were allowed to replace the child. No replacement was allowed after randomisation. Parent consent letters and the agreement between the schools and the Nottingham intervention team (MoU) were included in an appendix in the application to the Ethics Committee.

As soon as appropriate, the trial will be registered at The International Standard *Randomised Controlled Trial* Number (ISRCTN) <http://www.isrctn.com/>

Data Protection

The University of Oxford Ethics Committee has a data protection policy that can be found at: http://researchdata.ox.ac.uk/files/2014/01/Policy_on_the_Management_of_Research_Data_and_Records.pdf

A data sharing agreement between the Oxford and Nottingham teams was prepared for this project and is included as an appendix to the protocol (see Appendix 5).

PERSONNEL

Project team: The University of Nottingham is responsible for the project implementation with a team formed by Nicola Pitchford (Project Lead), Maria Neves (Programme Manager), Marc Faulder (Educational Consultant), Anthea Gulliford (Co-Investigator) and Geoffrey Wake (Co-Investigator).

Evaluation team: The evaluation team is based at the University of Oxford and is composed by Maria Evangelou, Terezinha Nunes and Rossana Barros, who will oversee all the aspects of the evaluation and will be responsible for the data analysis and report writing. The main contacts for the project are Terezinha Nunes and Maria Evangelou. Deborah Evans, Susan Baker and David Sanders-Ellis are responsible for training the testers and implementing quality control measures, maintaining contact with schools, designing the logistics and coordinating data collection at pre- and post-test, participating in data collection at pre-test and during process evaluation, participating in data analysis, and contributing to all aspects of the project, including being critical readers of the final report.

Roles and responsibilities

Each person will carry out their duties with the assistance of teams at their respective institutions.

The University of Nottingham team: will carry out the recruitment and provide a record of recruitment steps, including number of schools contacted, number of expressions of interest received, nomination of TAs at this stage and subsequent changes, nomination of contact person; collect MoUs; supply list of eligible schools for randomisation and number of classes in the school for random selection of class, if the school has more than one Year 1 class; train the TAs and provide record of training attendance and pupil registers; oversee the delivery of intervention; supply factors for success; supply information on costs; and facilitate the access and communication of the evaluation team with the schools. They will be critical readers of the report prepared for publication about the results of the trial that will be submitted to the EEF. They will co-author other publications arising from this trial with the Oxford team, if they wish to do so.

Evaluation team (Oxford University): will be responsible for trial design and registration; for obtaining ethical approval from Oxford University; for supporting the Project team in preparing letters and information for schools; for preparation and distribution of parental consent letters; for obtaining the nomination of selected children from schools; for randomisation and providing information on the outcomes of randomisation to the intervention team; for test preparation, administration, distribution, collection, and quality control of marking of tests; for liaising with DfE for obtaining data from the National Pupil Database (NPD); for carrying out analyses and writing the report and its first publication; for carrying out process evaluation observation visits; for obtaining and analysing questionnaires and interviews for the process and cost evaluation.

The evaluation team is committed to provide the Nottingham team with information about when communications with the schools will be established, what the contact people in the schools will be asked to do, and how they will be contacted. Any unexpected reactions will be reported immediately to the Nottingham team and the ways to maximise continued cooperation in the particular school will be agreed. The evaluation team is aware of the resources restrictions in schools and of the need to minimise the burden placed on administrators and teachers.

RISKS

| Issue | Likelihood of risk | Mitigating actions |
|-------|--------------------|--------------------|
|-------|--------------------|--------------------|

| | | |
|--|----------------|---|
| Difficulty or delays in recruiting schools and consequently TAs and children | Medium to high | Friendly and clear materials that will explain the design and value of the study; follow up phone calls and visits to schools. The Oxford team is very experienced in working with large numbers of primary schools and obtaining high participation rates and will support the Nottingham team as required. |
| TA or pupil attrition | Moderate | Very clear information to be offered to schools explaining the evaluation design, their involvement and the expectations. Because the role of TAs is limited and training can be done using the iTunes resources, should a TA have to leave, schools should be able to find another TA to run the intervention with relative ease. |
| Control TAs and children exposed to elements of the intervention or to variation from 'business as usual' | Low | Heads of schools have signed an agreement with the intervention team that includes their acceptance of random assignment to either the intervention or control group. It is agreed that, if they are assigned to the control group, they will not download the app and will wait until it is made available to the school free of charge after post-testing. The evaluation team will include a question in the middle-management staff questionnaire regarding the school's adherence to the commitment made in the agreement. The design of randomisation at school level greatly reduces the risk of contamination |
| Delays in commencing the delivery of the intervention | Medium | Agree a clear timetable with the intervention team; be flexible as much as possible in revising the timings of pre- and post-testing if there are small delays in recruitment. |
| Poor completion of questionnaires by TAs | Low | The team is very experienced in working with TAs and primary school teachers and will seek to make the assessment process accessible and a valuable experience for all involved. |
| Researchers lost to project due to sickness, absence or change of employment | Low | The team is able to recruit new researchers fairly quickly. The team members have worked together previously and have expertise in different aspects required for this project. |
| Children's attendance may vary and may affect the intervention dose | Medium | Although the risk cannot be avoided, attendance can be recorded and considered in process evaluation and dosage of the interventions. This would be taken into account and, if necessary, an on treatment analysis would be carried out. |

TIMELINE

| Date | Activity |
|--------------------------------|---|
| Sep - 2017 - March 2018 | Ethical approval; Draft of MoU; Draft of instruction on pupil selection for the trial; develop and register protocol at http://www.isrctn.com/ |
| Sep - 2017- Jan 2018 | Recruitment of schools; collect MoU; nomination of pupils for the trial and collection of pupil data and consent. |
| Jan 2018 | Training of testers; designing logistics for test implementation |
| Jan - Feb 2018 | Pre-test (to be completed by end of first week of Feb); randomisation; training of TAs; start of the intervention |
| Feb - June 2018 | 12 weeks of intervention to end in 1 st week of June; process evaluation; liaise with NPD to obtain data on eligibility for FSM |
| June - July 2018 | Post-testing (second visit to test absent children; to end by 1st week of July) |
| Aug - Sept 2018 | Analysis of process evaluation data |
| Sept - Oct 2018 | Data Analysis of test data and writing up process evaluation |
| Oct - Dec 2018 | Data Analysis and report preparation |
| Jan - March 2019 | Review of report; preparation of final report |

REFERENCES

- DONG, N. & 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(1), 24-67. doi: 10.1080/19345747.2012.673143
- OUTHWAITE, L.A., GULLIFORD, A., & PITCHFORD, N. J. (2017). Closing the gap: Efficacy of a tablet intervention to support the development of early mathematical skills in UK primary school children. *Computers & Education*, 108, 43-58.
- OUTHWAITE, L.A., FAULDER, M., GULLIFORD, A., & PITCHFORD, N.J. (2018). Raising early achievement in math with interactive apps: A randomized control trial. *Journal of Educational Psychology*. In press.
- PITCHFORD, N. J. (2015). Development of early Mathematical skills with a tablet intervention: a randomised control trial in Malawi. *Frontiers in Psychology*, 6(485), doi:10.3389/fpsyg.2015.00485
- RUTTERFORD, C., COPAS, A. & ELDRIDGE, S. 2015. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 1051–1067.
- WORTH, J., SIZMUR, J., AGER, R., & STYLES, B. (2015). Improving Numeracy and Literacy. Report presented to the EEF.

List of Appendices

Appendix 1.1: Local Authorities in the target regions for recruitment and selected Local Authorities within those regions

Appendix 1.2: Power calculation for MDES for the recruited sample after withdrawal of two schools

Appendix 1.3: Power calculation for the subgroup analysis including only pupils eligible for FSM in the recruited sample

Appendix 1.4: MoU and Parent Information letters

Appendix 1.5: Data sharing agreement between the Oxford and Nottingham teams

Appendix 1.1: Local authorities in the target regions for recruitment and selected Local Authorities within those regions

| Region | Local Authorities |
|------------------------------------|---|
| 1. East Midlands | Derby City, Derbyshire, Leicester City, Leicestershire, Lincolnshire, Northamptonshire, Nottingham City, Nottinghamshire |
| 2. West Midlands | Birmingham, Coventry, Dudley, Sandwell, Solihull, Staffordshire, Walsall, Warwickshire, Wolverhampton, Worcestershire |
| 3. Greater Manchester & North West | Blackburn with Darwen, Blackpool, Bolton, Burnley, Bury, Halton, Knowsley, Liverpool, Manchester, Oldham, Rochdale, Runcorn, Salford, Stockport, Tameside, Trafford, Wigan Bradford, Leeds, Wakefield, Barnsley, Calderdale, Kirklees, Doncaster, Rotherham, Sheffield |
| 4. Yorkshire West & South | |

Appendix 1.2: Power calculation for MDES for the recruited sample after withdrawal of two schools

| Power calculation using PowerUp | | |
|---------------------------------|--------------|--|
| Assumptions | | Comments |
| Alpha Level (α) | 0.05 | Probability of a Type I error |
| Two-tailed or One-tailed Test? | 2 | |
| Power ($1-\beta$) | 0.80 | Statistical power (1-probability of a Type II error) |
| Rho (ICC) | 0.15 | Proportion of variance in outcome that is between clusters |
| P | 0.50 | Proportion of Level 2 units randomized to treatment: $J_T / (J_T + J_C)$ |
| R^2 | 0.49 | Proportion of variance in Level 1 outcomes explained by Level 1 covariates |
| R^2 | 0.40 | Proportion of variance in Level 2 outcome explained by Level 2 covariates |
| g^* | 1 | Number of Level 2 covariates |
| n (Average Cluster Size) | 10 | Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended) |
| J (Sample Size [# of Clusters]) | 113 | Number of Level 2 units |
| M (Multiplier) | 2.83 | Computed from T_1 and T_2 |
| T_1 (Precision) | 1.98 | Determined from alpha level, given two-tailed or one-tailed test |
| T_2 (Power) | 0.84 | Determined from given power level |
| MDES | 0.194 | Minimum Detectable Effect Size |

Appendix 1.3: Power calculation for the subgroup analysis including only pupils eligible for FSM in the recruited sample

| Power Calculation using PowerUp | | |
|---------------------------------|--------------|--|
| Assumptions | | Comments |
| Alpha Level (α) | 0.05 | Probability of a Type I error |
| Two-tailed or One-tailed Test? | 2 | |
| Power ($1-\beta$) | 0.80 | Statistical power (1-probability of a Type II error) |
| Rho (ICC) | 0.15 | Proportion of variance in outcome that is between clusters |
| P | 0.51 | Proportion of Level 2 units randomized to treatment: $J_T / (J_T + J_C)$ |
| R^2 | 0.49 | Proportion of variance in Level 1 outcomes explained by Level 1 covariates |
| R^2 | 0.40 | Proportion of variance in Level 2 outcome explained by Level 2 covariates |
| g^* | 1 | Number of Level 2 covariates |
| n (Average Cluster Size) | 3 | Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended) |
| J (Sample Size [# of Clusters]) | 88 | Number of Level 2 units |
| M (Multiplier) | 2.83 | Computed from T_1 and T_2 |
| T_1 (Precision) | 1.99 | Determined from alpha level, given two-tailed or one-tailed test |
| T_2 (Power) | 0.85 | Determined from given power level |
| MDES | 0.286 | Minimum Detectable Effect Size |

Appendix 1.4: Memorandum of Understanding and Parent Information letters

MEMORANDUM OF UNDERSTANDING

regarding the *onebillion: app-based maths learning trial*

Purpose of the trial

The University of Nottingham, with the support of the Education Endowment Foundation (EEF), is investigating if two new maths apps, Maths 3-5 and Maths 4-6, developed by the UK-based not-for-profit onebillion, are effective at raising mathematical standards with Year 1 pupils in need of extra support with learning maths. The University of Oxford will be responsible for the independent evaluation of this trial.

Trial implementation and evaluation overview

All schools

September-November 2017

- Schools will register their interest in the trial through the Register Your Interest (RYI) form.
- Schools that have more than one Year 1 class will provide the names of each Year 1 class when returning the RYI form; only one class is randomly selected to participate in the trial. The Oxford Evaluation team will do this random selection.
- If the School is selected for participation, they will receive the Memorandum of Understanding (MoU), which will inform them about which class will participate in the trial. The school will return it signed as soon as possible.
- Schools that have returned the MoU will receive guidance on how to select 10 pupils for participation in the trial. The Oxford evaluation team will provide the materials required to obtain parental consent for the pupils' participation. If there are parental refusals, schools return forms to the Oxford Evaluation team, nominate a replaced pupil, and obtain parental consent.
- By 30th November 2017, schools will nominate 10 pupils for participation by uploading the following pupil information in Oxford's secure site: name, Unique Pupil Number (UPN), date of birth, gender and free school meal (FSM) eligibility status for all participating children. Instructions for this procedure will be sent with the parental consent forms.
- Schools also nominate a member of staff to be responsible for the project (normally a Teacher Assistant (TA); for brevity, the MoU refers to the staff member as the TA).

- Oxford University will create an identification number for each pupil and remove the names from the data files. To comply with the data protection act, the spreadsheet that connects the child's name and to the identification number will be saved in a separate password protected file.

January-February 2018

- Oxford Evaluation team schedules and implements the pre-test assessments to the 10 participating children. Schools will respond promptly to requests to schedule the pre-tests and provide adequate space.
- Oxford University randomly assigns the schools either to the intervention or to the control group.
- Nottingham Project Team communicates the assignment to the schools and schools confirm attendance to the training day.

MEMORANDUM OF UNDERSTANDING

regarding the *onebillion: app-based maths learning trial*

| Intervention Schools | Control Schools |
|--|---|
| <p>February - June 2018</p> <ul style="list-style-type: none"> • The nominated TA will attend a full-day training event, which will take place in the week starting on 26th February 2018 (venue to be confirmed) and the Nottingham Project Team will provide training for delivering the intervention. • The onebillion maths apps will be given to schools free of charge. • The intervention starts in March 2018 and runs for 12 consecutive school weeks (excluding holidays). • The intervention will be delivered to the 10 nominated children in a small group for 30 minutes each day for 4 days each week. Children will work individually with the apps through an iPad connected to a set of headphones. • The nominated TA will manage the small group of pupils and will provide technical and pedagogical support to pupils in the use of the technology, if required. • The TA will complete a register on a daily basis using a paper log and on a weekly basis on an online system. | <p>February - June 2018</p> <ul style="list-style-type: none"> • Control schools will continue to implement their support to pupils with maths difficulties as usual. Control schools agree not to use the Maths apps with the nominated children. |
| <p>April - June 2018</p> <ul style="list-style-type: none"> • TAs, the class teacher, and a subject specialist will complete questionnaires about the use of the app. • Evaluation team carries out a phone interview with staff in some randomly selected schools and/or observes an intervention session. | <p>April - June 2018</p> <ul style="list-style-type: none"> • TAs, the class teacher, and a subject specialist will complete questionnaires about the support offered to the nominated children doing this period. • If randomly selected, the class teacher will be asked to answer a phone interview. |
| <p>June-July 2018</p> <ul style="list-style-type: none"> • The Oxford Evaluation Team will assess the 10 nominated children in the Year 1 class (post-test assessment) individually. Schools will respond promptly to requests to schedule the post-tests and provide adequate space. | <p>June-July 2018</p> <ul style="list-style-type: none"> • The Oxford Evaluation Team will assess the 10 nominated children in the Year 1 class (post-test assessment) individually. Schools will respond promptly to requests to schedule the post-tests and provide adequate space. • After all the data have been provided to the Oxford evaluation team, control schools will receive £1000 for taking part in the trial and will have access to the onebillion maths apps free of charge. Control schools will maintain their commitment not to use the app with the 10 nominated children. |

All schools

- ✓ Oxford Evaluation Team analyses the results and presents the report to the EEF.
- ✓ The report will be published on the EEF site in the Summer 2019.

MEMORANDUM OF UNDERSTANDING

regarding the *onebillion: app-based maths learning trial*

Benefits of taking part in the trial include:

- ✓ All schools will receive free access to the Maths apps; intervention schools will gain access in February 2018 and control schools will gain access in July 2018. Control schools will be able to use the apps with the next cohort of Year 1 pupils, but not with the pupils nominated for this trial.
- ✓ Teaching Assistants at Intervention Schools will be trained on how to use the maths apps during a mandatory full-day of training in February 2018.
- ✓ Intervention schools will have access to online training to support them on how to use the maths apps once the intervention starts in March 2018 and control schools will gain access to the online training once the intervention finishes in July 2018.
- ✓ Control schools will receive £1000 financial incentive, after all the data have been received by the Oxford Evaluation team.
- ✓ You will be helping to build evidence for maths interventions and learn more about educational research.

Use of Data

All data, including children's test responses and any other pupil data, will be treated as strictly confidential. All data collected will remain confidential using password protected files and will remain anonymous. Only group data will be reported. Children's assessments will be analysed by the Oxford Evaluation Team. No individual school, parent, or child will be identified in any report arising from the research.

We will be asking the school to provide us with information on the children's Unique Pupil Numbers (UPNs) and free school meal (FSM) eligibility status to complement the assessment of the Maths apps. The UPN is part of the Department for Education records. The UPN will allow us to link project results to the National Pupil Database and to share data with the EEF, the EEF's data contractor FFT Education, the Department for Education, and the UK data archive for research purposes.

MEMORANDUM OF UNDERSTANDING

regarding the *onebillion: app-based maths learning trial*

Agreement to participate in the onebillion: app-based maths learning trial

This agreement is between the school named below (henceforth referred to as the School) and the University of Nottingham about a randomised control trial of the intervention using the onebillion apps and evaluated by the University of Oxford.

| | |
|---------------------------------|-----------------|
| Name of School: | |
| Address: | Postcode |
| Head Teacher: | |
| Telephone: | e-mail: |
| Randomly selected class: | |

The school

- The School understands that it will be randomly allocated in February 2018 either to the intervention group (using the apps) or to the control group (receiving the apps at the end of the trial); it commits to full participation in either group until July 2018 according to the schedule included with this agreement. Schools assigned to the control group agree not to use the apps with the 10 children nominated for participation in this project but are free to use the apps with the next Year 1 cohort.
- The School will identify 10 children in the randomly selected Year 1 class to participate in the project according to the criteria provided by the University of Nottingham. The School will seek permission from the parents of the 10 Year 1 pupils nominated for the project in November 2017 for participating in the trial and for the data to be shared with the evaluation team. The consent form seeks permission separately for sharing the pupil's unique pupil number and eligibility for free school meals. Pupils can still participate in the trial even if their parents did not agree to releasing this information. If parents refuse permission for their children to participate in the trial, the School can nominate another child as long as this is done before schools have been randomised to the groups.
- The School will also identify a teaching assistant (or another staff member) who will attend training and support the intervention, if the school is allocated to the intervention group. The School will provide the details of the nominated pupils and the TA by the 30th of November of 2017 to the Oxford University evaluation team by uploading the information in Oxford's secure site.
- In January/early February 2018, the School will provide access to the evaluation team to administer a mathematics test to the 10 nominated children in Year 1. This assessment is implemented individually and the School will provide a quiet space for the assessment.
- The School will communicate fully and promptly with the University of Nottingham and the evaluation team, share appropriate data and ensure that questionnaires and surveys are completed and returned. The School will ensure that, if selected for an interview, the Year 1 teacher, the TA implementing the intervention, and the subject leader participate in the interview.
- The School will facilitate visits to the school by the University of Oxford to administer a post-test to the nominated Year 1 pupils in June –July 2018. This assessment is implemented individually and the School will provide a quiet space for the

assessment.

- If the School is allocated to the intervention group, the School will ensure that 10 iPads with internet access and 10 headphones will be available for half an hour for four days a week during the 12 weeks of the intervention (March-June 2018) to be used by the nominated children. The School will support the TA to attend one full-day training in February 2018.
- If the School is allocated to the intervention group, it will commit time and space for the nominated TA to deliver a 30-minute slot, for 4 days each week during 12 weeks to the 10 nominated children (excluding holidays). The TA will keep register of children's attendance, if there was any need for technical assistance, if there was any need for pedagogical assistance and records of the certificates attained by the pupils on paper on a daily basis and at an online system on a weekly basis. If randomly selected for observation, the School will facilitate visits by the Oxford Evaluation team to observe the intervention sessions.
- Irrespective of its group membership in the trial, the School will deliver maths lessons as usual, including any maths interventions in place.

MEMORANDUM OF UNDERSTANDING

regarding the *onebillion: app-based maths learning trial*

The University of Nottingham (UoN)

- During the week starting on 12th February 2018, UoN will inform the School of its group allocation and of the training date and venue. UoN will provide to intervention schools all necessary items for the training, including free access to the apps.
- The training will cover many aspects of the intervention, such as how to get started using the equipment and the apps, how to navigate in the apps, how to troubleshoot technical aspects, and how to monitor pupil progress on a daily and weekly basis.
- If the School is allocated to the control group, UoN will provide the School with £1,000 financial incentive and access to the onebillion maths apps free of charge at the end of the trial (July 2018), after the Oxford Evaluation Team has received all the necessary data.
- The evaluation team will train the testers who will implement the pre- and post-tests and verify that they have a DBS before they are sent to the school.

We commit to the evaluation of the *onebillion: app-based maths learning*

Please sign both copies, retaining one and returning the second copy by post to Dr Maria Neves, School of Psychology, University of Nottingham, University Park, Nottingham, NG7 2RD

or e-mail a signed copy to maria.neves@nottingham.ac.uk

Signed for and on behalf of:

School _____

University of Nottingham

Signed _____

Signed _____

Name (print) _____

Name _____

Position _____

Position_____

Date _____

Date _____

Parent Information and Consent Forms

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 Maths app on improving mathematics learning in Yr 1. The Maths app offers children self-paced opportunities to rehearse materials that cover many aspects of the National Curriculum. The presentation is attractive and children enjoy using it. The project is funded by the Education Endowment Foundation (EEF). The University of Nottingham project team will work with a teaching assistant who will implement the programme and a team from Oxford University will be evaluating the impact of the programme on the children's mathematical attainment. The EEF decided to fund this project because the University of Nottingham team has conducted previous evaluations of the Maths apps, which provided evidence of promise.

Your child's school has kindly agreed to be part of the study and we are writing to you to ask for permission for your child to participate. The school has identified the children who would benefit from using the 'App'. Before the start of the project, all the children taking part in the project will complete an assessment carried out individually by a researcher trained by Oxford University. The Maths intervention is presented on a tablet; children work independently for 4 weekly half-hour sessions over 12 weeks, supervised by a teaching assistant. These sessions are in addition to usual, daily classes of mathematics. At the end of the programme, the children will participate in a second assessment. These assessments aim to evaluate the programme, not the children. These assessments do not influence your child's placement in school. They are necessary only for the research.

Schools will be allocated on a lottery basis by the Oxford evaluation team either to an intervention or to a control group. All the schools in the project will have the opportunity to use the apps, but not all the children. The intervention schools will use the apps during this school year and others will use them in the next academic year. 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. 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 will be asking the school to provide us with information on the children's Unique Pupil Numbers (UPNs) and free school meal (FSM) eligibility status to complement the assessment of the 'Maths app' programme. The UPN is part of the Department for Education records. The UPN will allow us to link test results to the National Pupil Database and share data with the EEF, the EEF's data contractor FFT Education, the Department for Education, and 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 take part in the research and for their UPN and FSM eligibility status to be used, **then you do not need to do anything**. If you **do not** wish your child to participate in the project or the research team to have access to your child's UPN and/or FSM eligibility status, please complete and return the attached form to your child's class teacher by 30th November.

We expect that your child will enjoy taking part in 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

UNIVERSITY OF OXFORD
DEPARTMENT OF EDUCATION



Professor Terezinha Nunes
15 Norham Gardens Oxford
OX2 6PY

Title: Maths app evaluation project

If you agree to your child taking part in the research and their UPN and Free School Meals eligibility to be used, then **you do not need to sign and return this form.**

If you would like your child to take part in the research but do not agree to releasing further information, tick the relevant box below.

☐

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

☐

I DO NOT consent for my child's free School Meals eligibility status to be released to the research team.

If you **do not agree** to your child taking part in the project, please tick the box below.

☐

I **DO NOT** agree to my child taking part in the research.

Child's name: Date of birth:

Child's class teacher:

School:.....

Parent/carer name (BLOCK CAPITALS)

Parent/carer signature:

Date

Appendix 1.5: Data Sharing Agreement

DATA SHARING AGREEMENT FOR ONEBILLION PROJECT UNIVERSITY OF NOTTINGHAM



EVALUATED BY UNIVERSITY OF OXFORD

22/1/2018

Page 1 of 3

BACKGROUND

In order to evaluate the Onebillion randomised controlled trial it will be necessary for the University of Oxford evaluation team (Oxford), participating schools and the University of Nottingham (Nottingham) project team to share data. This includes:

- ✓ outcome data obtained through assessment (mathematics assessments)
- ✓ teaching assistants' notes on session and app usage
- ✓ information on pupils' activities, including use of the app.

These data will only be used for the purposes of the evaluation and will be treated with great care to achieve high levels of security. Further information on this process is provided below.

DATA SECURITY

Data transfer between schools, Oxford & Nottingham

The project will involve transferring potentially sensitive pupil data between the schools, evaluation team (Oxford) and project team (Nottingham). Such data must be transferred securely, meaning that the following process will be followed carefully.

Secure data may be transmitted via email, with the following standards applied. It will be stored using an encrypted Microsoft Excel (.XLSX) spreadsheet. The password to open, edit and re-save this encrypted file will be agreed in advance of transfer and will not be reused across different parties. These passwords will conform to the following standards and will never be shared via email:

- ✓ Minimum length: 8 characters
- ✓ Contains at least one uppercase letter
- ✓ Contains at least one lowercase letter
- ✓ Contains at least one number

Participating schools will supply the required data on pupil characteristics directly to Oxford using a secure portal. This process can be reflected in the MOUs between participating schools and the project partners. All files are saved on a secure server. Each project has its own unique folder on the server, with access restricted to authorised users who are working on the project. This process is managed by the IT Department and all

requests must be formally approved by the project leader (Research Group convener) before access is granted. A record of access rights is maintained through the IT Department. Data transfer to Nottingham will follow the same process set out above.

Any potentially sensitive data will be stored on an encrypted USB flash drive when in transit. Oxford requires that confidential data must be encrypted, using AES 256bit encryption or stronger, when stored on mobiles devices or removable media.

Data may also be transferred physically, for example if collected as part of visits to schools for testing. Oxford will collect data on the attainment measures directly from the schools. Data collected from the schools by Oxford research assistants will be anonymised (and individual pupils will be identified through the use of a unique case identifier), the hard copies (completed paper assessments) will be collected by courier from Oxford research assistants and delivered to Oxford for counting. All data will be stored in locked rooms. Oxford will send the complete set of assessments to GL Assessments offices for scoring and production of data base. Data transfer between Oxford and GL Assessment will use GL Assessment's and Oxford's standard processes.



DATA SHARING AGREEMENT FOR ONEBILLION
PROJECT UNIVERSITY OF NOTTINGHAM
EVALUATED BY UNIVERSITY OF OXFORD

22/1/2018



After final completion of the project the data will be passed to EEF and/or its contractors for the purposes of contributing to the cross-project database. This data transfer will be carried out in line with the security procedures outlined above.

Data storage at Oxford

The Department of Education has an up-to-date information security policy which will be adhered to in relation to this project's data. All staff, students, visitors and collaborators using the Department's IT systems, data or any other information asset should follow the Department's Information Security Policy, the security policy references to ISO27002 and ICO (University's Registration Number is Z575783X, Registration Expires: 12 September 2018) (see document enclosed).

All confidential data will be stored securely; if a hard copy is kept, it will be stored in a locked cupboard in a locked room or, if stored electronically, data will be stored on departmental file servers attached to a corporate network and not on local hard drives. The server is backed up remotely to a server at the University IT Services. The PC where data will be processed is located in a locked room. The server where data will be stored is contained in a locked room. All of these locked rooms are within buildings only accessible via authorised swipe cards, and all access is logged on the access system. Also all external doors are video monitored 24 hours a day.

After completion of the project and publications, all media will be shredded and data held on the server will be deleted. We will securely erase using a utility called File Shredder. When the server is retired from service the hard drives it contains will be physically destroyed so no data can be recovered from them. This is a service periodically provided by University IT Services who have access to a disk crushing device.

Data storage at Nottingham

1. School recruitment and participation database – password protected, as specified by the procedures described above, stored on local PC backed up on School/University servers accessible only with by key or card.
2. Paper copies of the session log booklet sent to us by schools at the end of the trial. Identifying information will include school name, name of TA, and initials of pupils. We will ask schools to send this to us using the Royal Mail registered mail service and will issue schools with a prepaid envelop to do this. When this arrives in our School office it will be left in a pigeon hole for Maria to collect. The room where the pigeon hole is allocated is locked and only accessible by people (staff) with a digital code. Maria will collect these daily and lock them in a cabinet in her office which is only accessible by key allocated to her and the School Operations Manager and Head of School. At the end of the project we will shred this data.
3. Digital copies of the session logs uploaded weekly into iTunes U and entries on discussion fora. The iTunes

U course will be private for this trial and only intervention schools enrolled on the iTunes U course per region will be able to access the course, upload their weekly session log data, and engage with the discussion posts. Any posts made will be accessible to all people enrolled on the course for that region. iTunes U is a secure system for sending in data. Further details about iTunes U security and privacy can be found here: <https://support.apple.com/en-gb/HT204918>. At the end of the trial the iTunes U courses will be deleted and a

non-private course will be made available to the public so that control schools can access the training materials. In this non-private course we will remove the pupils progress monitoring information and discussion fora as this is only required for this trial. No data will be stored on the non-private course.



DATA SHARING AGREEMENT FOR ONEBILLION
PROJECT UNIVERSITY OF NOTTINGHAM
EVALUATED BY UNIVERSITY OF OXFORD

22/1/2018



USAGE OF THE DATA

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

- Checking randomisation has worked through analysing the average characteristics of individuals and schools in the treatment and control groups;
- Calculating the estimated impact of the project by comparing the outcomes of individuals in treatment and control schools.

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

MONITORING DATA/PROCESS EVALUATION

During the project the project team will be keeping in regular contact with participating schools (treatment and control groups). Where relevant, notes will be shared with the evaluation team for the purposes of the quantitative and process evaluations. The evaluation team will carry out data collection and analysis for process evaluation. Where relevant, information will be shared with the project team to support successful implementation. The project team will cooperate with the evaluation team to support the process evaluation (e.g. assisting with liaison with the schools to arrange research visits).

AGREEMENT

University of Oxford and University of Nottingham agree to work collaboratively in the sharing of data for the success of this project and will follow the procedures outlined in this document when handling the potentially sensitive data included that will be shared as part of this project. Any actual or potential breaches will be notified to the relevant data controller.

Signed:

Jo-Anne Baird

Director of the Department of Education University
of Oxford

Name
Position

University of Nottingham



Appendix 2: Pre-test and Post- test descriptions

Brief overview of the analyses of PTM5 and PTM6 used in the evaluation of the onebillion apps

Distribution of scores

In order to investigate the psychometric properties of the pre- and post-test in the evaluation of the onebillion apps, we first examined the distribution of scores and the descriptive statistics. The distribution of scores at pre-test is normal at pre-test: the z for the skewness is 1.19 (the cut-off point for normality is about 2) but the distribution for the scores at post-test is severely positively skewed ($z=10.92$), which indicates that the test was too difficult for the group. It must be noted that teachers were encouraged to select weaker pupils so a positively skewed distribution at pre- and post-test could have been expected for the sample. The discrepancy between pre- and post-test therefore cannot be explained by the selection of pupils for the study.

Table 1: Descriptive statistics of Pre and Post Test

| Descriptive Statistics | | | | | | | | |
|-------------------------|----------------|----------------------|----------------------|-------------------|--------------------------------|----------|------|-------|
| | N Statistic | Minimum Statistic | Maximum Statistic | Mean Statistic | Std. Deviation Statistic | Skewness | | z |
| Pre Total Raw Score | 1124 | 2 | 24 | 13.43 | 4.195 | -.087 | .073 | 1.19 |
| Post Total Raw Score | 1089 | 0 | 29 | 9.20 | 4.634 | .808 | .074 | 10.92 |
| Valid N (listwise) | 1089 | | | | | | | |

Figure 1: Pre-test histogram and normality curve

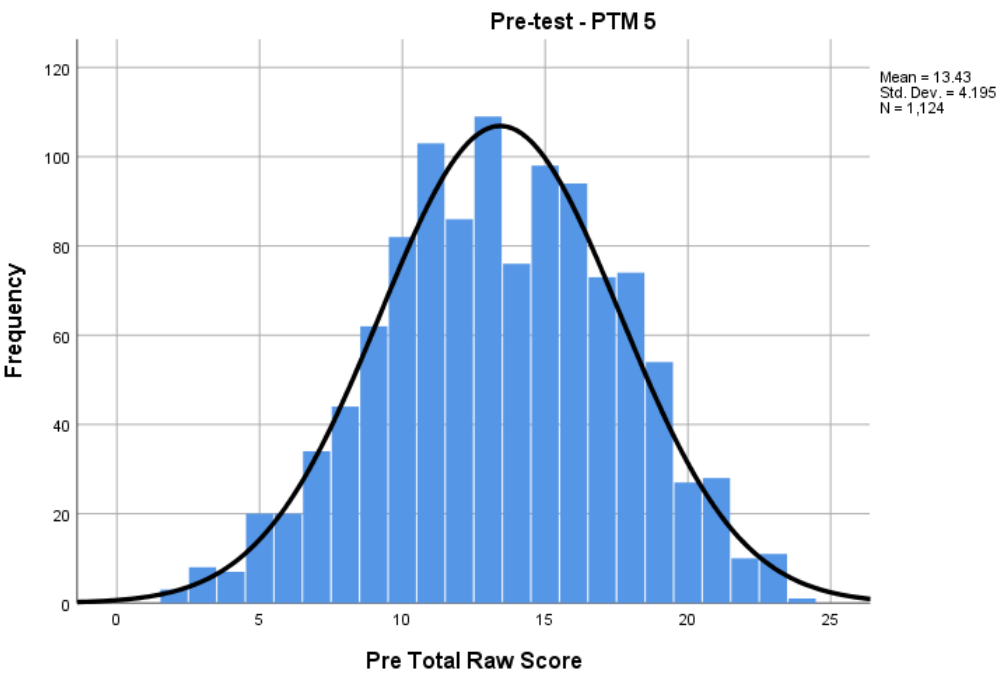
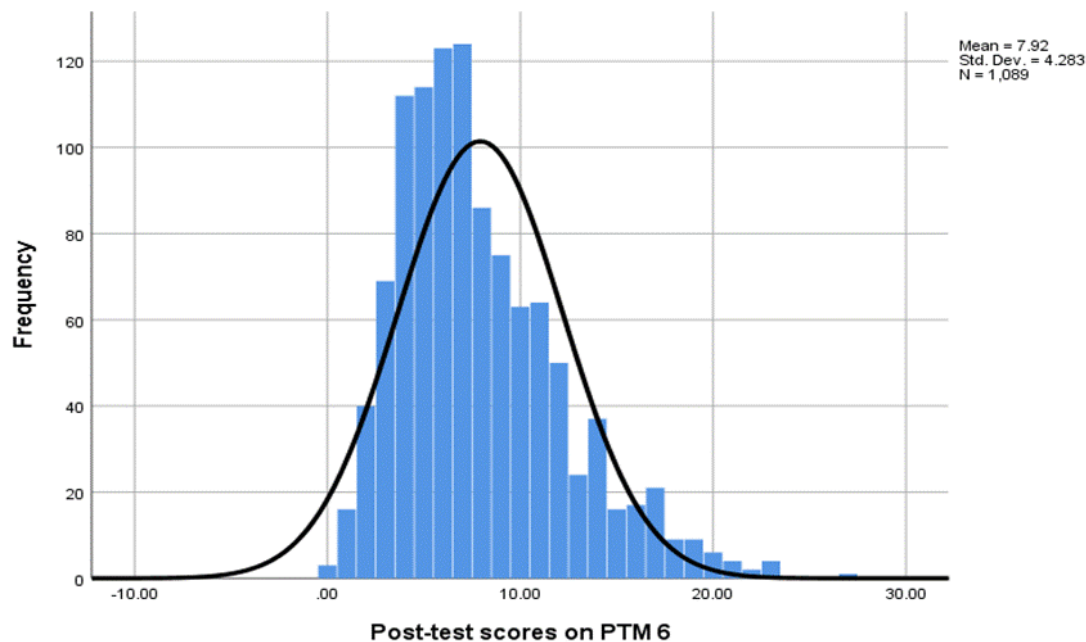


Figure 2: Post-test histogram and normality curve



The distribution of observed scores below the mean covers only 2SDs whereas at pre-test it covers 3SD, as expected in a normal curve. As expected in a normal distribution, the median and the mode at pre-test were 13 and coincide with the mean; at post-test, they were equal to 8, and one point lower than the mean.

Brief reliability analysis

The internal consistency was examined by means of Cronbach's alpha and the item-total correlations. At pre-test, one item (2c) had a negative correlation with the total. The observed reliability was .693 and would go up to .701 if this item were removed. At post-test, no items showed a negative correlation with the total. The reliability was .759. However, it was noted that item 8a is marked inconsistently with an appropriate mathematical answer (the item requires an answer for how many squares in a figure, then how many rectangles, and the answer treated as correct does not include the number of squares in the number of rectangles). The correlation between this item and the total was positive and its exclusion does not improve the reliability of the test.

Discrepancies in marking between GL assessment and the evaluation team

The Oxford team marked a sample of 280 pupils' post-test booklets from 29 different schools (25% of the sample) blindly with respect to the GL marking but implementing the same marking rules. The Oxford scoring was completed by two researchers who worked independently; when their marking differed, they sought consensus by discussing how their markings had been achieved. The consensus score was then entered into the dataset for comparison with the markings by GL assessment. No scoring was completed by the Oxford team on the Pre-test assessment. There were discrepancies in the marking, as described in the table below. The correlation between the Oxford and the GL assessment marking was .993 but the same exact score was only observed on 76.8% of the papers. These differences could be attributed mostly to the need to interpret the children's handwriting. However, because some items' scores are based on more than one answer, it is not always possible to investigate the source of discrepancy.

Table 2: Differences between Oxford marking and GL marking

| Total difference between GL and Evaluation team marking on case basis | Frequency observed | % of papers marked by Evaluation team |
|---|--------------------|---------------------------------------|
| -4 | 1 | 0.4 |
| -2 | 4 | 1.4 |
| 1 | 36 | 12.9 |
| 0 | 215 | 76.8 |
| 1 | 21 | 7.5 |
| 2 | 3 | 1.1 |

Pre- to post-test correlation

The correlation between the pre- and the post-test for the sample of children present on both occasions was .553. When the subsample of tests that was marked by the Oxford team was considered, the same correlation of .626 between pre- and post-test was observed using the GL assessment marking or the Oxford marking. Thus the discrepancies in the marking can be seen as random.

Brief interpretation

The post-test may not be sensitive to progress among lower scoring pupils but it does provide the opportunity to detect progress in general. A negatively skewed distribution would offer a more significant threat to the test's ability to detect a difference between the groups. The medium sized correlation between the pre- and post-test cannot be explained by a narrow range of scores; although the distribution is skewed, there are scores along the whole continuum at post-test.

Comparison of Post test items from PTM6 with content of onebillion apps

Table 3: Comparison of PTM6 items with content of onebillion apps.

| No of Question | Question details | Maths 3-5 - Fit or near fit | Maths 4-6 - Fit or near fit | Not covered |
|----------------|--|-----------------------------|-------------------------------|--------------|
| 1a | Smallest number | | Topics 11 and 14 | |
| 1b | Largest even number | | Topic 4, Topic 14 | |
| 1c | Odd number | | Topic 4, Topic 6, Topic 14 | |
| 1d | Number more than 5 but less than 7 | | Topic 11 | |
| 2a | 3 more than 6 | Topic 8 | Topic 2, Topic 6 , Topic 14 | |
| 2b | 4 doubled | | Topic 6 | |
| 2c | 7 less than 9 | Topic 8 | Topic 2, Topic 11, Topic 14 | |
| 3a | Time - reading digital time and changing to analogue | | Topic 5 | |
| 3b | Digital time - half an hour later | | | digital time |
| 3c | Analogue - half an hour later | | Topic 5 | |
| 4a | Half way between 2 numbers (eg 8 and 12) | | Topic 11 | |
| 4b | Counting (in 4s) up to 24 | | Topic 4, Topic 9 | |
| 4c | Counting (in 5s) up to 30 | | Topic 7, Topic 9 | |
| 5a | Counting (in 3s) up to 18 | | Topic 2, Topic 9 | |
| 5b | 2 more than | Topic 8 | Topic 11, Topic 14, Topic 17, | |
| 6a | Money questions | | | money |
| 6b | Money questions | | | money |
| 6c | Money questions | | | money |
| 6d | Money questions | | | money |
| 7a | How much taller? | | Topic 8 | |
| 7b | How much shorter? | | Topic 8 | |
| 7c | More than half full, less than half full | | Topic 18 | |
| 7d | Adding weights together | | Topic 13 | |
| 8a | Squares/rectangles | Topic 7 | | |
| 8b | Continuing patterns | Topic 7 | | |
| 8c | Counting pentagons | | Topic 10 | |

Appendix 3: Details of the models and syntax for the analyses

Table 1: Post-test outcome in intervention and control groups

| Outcome | Raw means | | Raw means | | Effect size | | |
|------------------|--------------------|----------------------|----------------|----------------------|--------------------|---------------------|---------------------------------|
| | Intervention group | | Control group | | n in model | Hedges g | p-value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | (Int; Contr) | (95% CI) | |
| Post-test | 543 (24) | 8.43 (8.06, 8.80) | 546 (11) | 7.41 (7.06, 7.76) | 1089 (543; 546) | 0.24 (0.12,0.36) | t(1087) = 3.97; p = 0.000078 |

Note: Effect sizes were calculated according to _____, where

_____. The

_____. The variance of d is given by

standard error of d is _____. Hedge's g is then calculated by $g = J \times d$, by applying the correction formula _____, where $df = (n_1 + n_2 - 2)$. The variance of g is

1 2

and _____ (Borenstein, Hedges, Higgins & Rothstein, 2009, equations 4.18 to 4.25). The p-value was taken from SPSS, assuming equal variances.

Table 1a: Effect size estimation

| Outcome | Unadjusted differences in means | Adjusted differences in means | Intervention group | | Control group | | Pooled variance | Population variance (if available) |
|------------------|---------------------------------|-------------------------------|--------------------|-------|---------------|-------|-----------------|------------------------------------|
| | | | | | | | | |
| Post-test | 1.02 | 1.06 | 543 (24) | 19.05 | 546 (11) | 17.15 | 18.10 | |

Note: The adjusted mean difference was taken from the multilevel model (model 4), i.e., estimated differences in the outcome controlling for the pre-test.

Table 2: Post-test outcome in intervention and control groups among non-FSM pupils only

| Outcome | Raw means | | | | Effect size | | |
|------------------|--------------------|----------------------|----------------|----------------------|----------------------------|----------------------|-----------------------------|
| | Intervention group | | Control group | | n in model (Int; Contr) | Hedges g (95% CI) | p-value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| Post-test | 404 (13) | 8.95 (8.52, 9.39) | 407 (7) | 7.47 (7.08, 7.87) | 811 (404; 407) | 0.35 (0.21, 0.48) | t(809) = 4.93; p < 0.001 |

Table 3: Post-test outcome in intervention and control groups among FSM pupils only

| Outcome | Raw means | | | | Effect size | | |
|------------------|--------------------|----------------------|----------------|----------------------|----------------------------|------------------------|-----------------------------|
| | Intervention group | | Control group | | n in model (Int; Contr) | Hedges g (95% CI) | p-value |
| | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | | | |
| Post-test | 134 (11) | 6.87 (6.25, 7.50) | 137 (4) | 7.26 (6.54, 7.99) | 271 (134; 137) | -0.10 (-0.14, 0.33) | t(269) = -0.79; p = 0.43 |

1. Effects of intervention on outcome

There were 1124 pupils randomised at the school level into a control group ($n = 557$), an intervention group ($n = 567$). Of these 1089 pupils ($n_{\text{int}} = 543$, $n_{\text{ctrl}} = 546$) had valid scores on the outcome. The pupils were nested in 113 schools. The effects of the intervention were tested in a series of multilevel models with post_{ij} as the outcome, where i = pupils nested in j

= schools as follows:

$$(1) \text{ Variance component model } \text{post}_{ij} = b_0 + u_{0j} + e_{0ij}$$

$$(2) \text{ Fixed effect of pre-test score } \text{post}_{ij} = b_0 + b_1 \text{pre}_{ij} + u_{0j} + e_{0ij}$$

$$(3) \text{ Fixed and random effect of pre-test score } \text{post}_{ij} = b_0 + b_1 \text{pre}_{ij} + u_{0j} + u_{1j} \text{pre}_{ij} + e_{0ij},$$

unstructured level-2 matrix

$$(4) \text{ Fixed effect of pre-test score and intervention } \text{post}_{ij} = b_0 + b_1 \text{pre}_{ij} + b_2 \text{intervention}_j + u_{0j} + e_{0ij}$$

in which post_{ij} is the outcome, b_0 is the grand intercept, b_1 - b_2 are the fixed effects, u_0 - u_1 are random effects and σ^2 is the variance of the residual terms. The maximum likelihood estimator was used to facilitate model comparisons (with equal sample sizes) using differences in the -2 log likelihood.

As shown in Table 1, Model 1 showed that 20% of the variance was between schools. Model 2 showed that the pre-test score was a significant predictor of the outcome ($B = 0.53$). The ICC of the pre-test was 0.20. Model 3 did not converge, suggesting there might not have been sufficient slope variance (i.e., differences between schools in the post-test-regressed- on-pre-test slopes). This model was not considered further and the random slope term was removed from further models. Model 4 showed that the one-billion app intervention had a positive effect on the outcome ($B = 1.06$), controlling for the pre-test (explaining 9% of the variance in the outcome).

For the next step of modelling, the moderation effect of Free-School-Meal (FSM) eligibility was tested. There were 286 FSM-eligible pupils (25.6 valid %) and 831 (74.4 valid %) who were not, and 7 for whom it was not known. First a model including FSM-status and the FSM

× intervention (cross-level) interaction effects were included:

$$(5) \text{ } \text{post}_{ij} = b_0 + b_1 \text{pre}_{ij} + b_2 \text{intervention}_j + b_3 \text{FSM}_{ij} + b_4 \text{FSM}_{ij} \times \text{intervention}_j + u_{0j} + e_{0ij} \text{ Then}$$

the model was run only for the intervention group:

$$(6) \text{ post}_{ij} = b_0 + b_1 \text{pre}_{ij} + b_2 \text{FSM}_{ij} + u_{0j} + e_{0ij}$$

Table 1. Effects of the math-app (one billion) intervention.

| | Model 1 | | | Model 2 | | | Model 3 | | | Model 4 | | |
|-------------------------------|---------|------|------|--------------------|------|------|--------------------|------|------|--------------------|------|------|
| <i>Fixed</i> | B | s.e. | sig. | B | s.e. | sig. | B | s.e. | sig. | B | s.e. | sig. |
| Intercept | 7.93 | 0.21 | | 0.85 | 0.39 | | | | | 0.27 | 0.42 | |
| Pre-test | | | | 0.53 | 0.03 | *** | | | | 0.53 | 0.03 | *** |
| Intervention | | | | | | | | | | 1.06 | 0.31 | *** |
| <i>Random</i> | | | | | | | | | | | | |
| Residual (e_{0ij}) | 14.70 | 0.67 | | 11.24 | 0.51 | | | | | 11.23 | 0.51 | |
| Intercept (σ^2_{u0}) | 3.65 | 0.70 | | 1.78 | 0.40 | | | | | 1.46 | 0.36 | |
| Slope (σ^2_{u1}) | | | | | | | | | | | | |
| Int-slope (σ_{u0u1}) | | | | | | | | | | | | |
| Variance proportions | | | | Explained variance | | | Explained variance | | | Explained variance | | |
| School | 0.20 | | | 0.51 | | | n/a | | | 0.60 | | |
| Pupil | 0.80 | | | 0.24 | | | n/a | | | 0.24 | | |
| -2LL | 6154.99 | | | 5834.75 | | | | | | 5816.16 | | |
| Δ -2LL | | | | 320.24 | | | *** | | | 18.59 | | |

Note: Significances for intercepts and variances not reported. Explained variances were calculated following Hox (2002), -2LL is -2 × likelihood. Improvement in model fit was calculated using a χ^2 -test for the Δ -2LL with the number of estimated parameters as the degree of freedom. All models were run in SPSS 25 using the maximum likelihood estimator with raw (non-centred) predictors. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 0.001$.

Table 2. Differential effects of pupils eligible for free school meals (FSM)

| | Model 5 | | | Model 6 | | |
|-------------------------------|---------|------|------|-------------------------|------|------|
| n = 1117 | | | | FSM-pupils only n = 286 | | |
| <i>Fixed</i> | B | s.e. | sig. | B | s.e. | sig. |
| Intercept | 0.42 | 0.42 | | -0.30 | 0.78 | |
| Pre-test | 0.53 | 0.03 | *** | 0.55 | 0.05 | *** |
| Intervention | 1.35 | 0.33 | *** | 0.11 | 0.47 | |
| FSM | -0.41 | 0.36 | | | | |
| FSM x intervention | -1.15 | 0.51 | * | | | |
| Boy | | | | | | |
| Boy x intervention | | | | | | |
| <i>Random</i> | | | | | | |
| Residual (e_{0ij}) | 11.04 | 0.50 | | 9.73 | 0.97 | |
| Intercept (σ^2_{u0}) | 1.42 | 0.35 | | 1.25 | 0.70 | |

Note: Significances for intercepts and variances not reported. Explained variances were not calculated as sample sizes differed across the three models. All models were run in SPSS 25 using the maximum likelihood estimator with raw (non-centred) predictors. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 0.001$.

As we can see in Table 2 (model 5) there was no main effect of FSM on the outcome ($B = -0.41$), but the interaction effect was significant ($B = -1.15$; $p \leq 0.05$). Non-FSM pupils in the intervention group outperformed the non-FSM pupils in the control group, but there was no significant difference between the FSM-pupils in the intervention and control-groups. Estimated means were calculated as post-test = $7.48 + 0.53(\text{pre-test}) + 1.35(\text{intervention}) - 0.41(\text{FSM}) - 1.15(\text{intervention} \times \text{FSM})$. This means that there was an intervention effect on the non-FSM pupils, but not on the FSM-pupils. Running the model (Model 6) among the FSM-pupils only did not yield a significant effect consistent with Model 5.

2. Effects of fidelity

In order to investigate whether differences in implementation of the interventions in different classrooms (with different teachers) had an effect on the effect of two process variables were explored: (1) the number of sessions pupils were reported to use the app, (2) how teachers were observed to run the sessions, in the following models:

$$(9) \text{post}_{ij} = b_0 + b_1\text{pre}_{ij} + b_2\text{sessions}_j + b_4\text{sessions}_{ij} \times \text{pre}_{ij} + u_{0j} + e_{0ij}$$

$$(10) \text{post}_{ij} = b_0 + b_1\text{pre}_{ij} + b_2T_met_exp_j + b_4T_met_exp_j \times \text{pre}_{ij} + u_{0j} + e_{0ij}$$

Note: T_met_exp is the variable name for the measure of TA compliance with expectations of support described in the Top Tips section of the Implementation Manual (a three-point ordinal variable described in the methods section). As shown in Table 3 there were no significant effects of either implementation fidelity measure.

Table 3. Implementation fidelity effects on outcome (in intervention group only)

| | Model 5 | | | Model 6 | | |
|---|---------|------|------|---------|------|------|
| | n = 553 | | | n = 291 | | |
| <i>Fixed</i> | B | s.e. | sig. | B | s.e. | sig. |
| Intercept | 1.08 | 2.42 | | -0.14 | 2.39 | |
| Pre-test | 0.51 | 0.17 | *** | 0.64 | 0.15 | *** |
| No of app sessions | 0.01 | 0.06 | | | | |
| Pre-test x No of app sessions | 0.00 | 0.00 | | | | |
| TA's compliance with expectations of support | | | | 0.66 | 1.20 | |
| Pre-test x TA's compliance with expectations of support | | | | -0.05 | 0.08 | |
| <i>Random</i> | | | | | | |
| Residual (e_{0ij}) | 11.02 | 0.72 | | 10.53 | 0.92 | |
| Intercept (σ^2_{u0}) | 1.90 | 0.60 | | 2.07 | 0.82 | |

Note: Significances for intercepts and variances not reported. Explained variances were not calculated as sample sizes differed across the three models. All models were run in SPSS 25 using the maximum likelihood estimator with raw (non-centred) predictors. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 0.001$.

3. Sensitivity analyses

In the following two anomalous classrooms were removed from the analyses (no 4900 and 10200). The same models 1-10 were run for the reduced sample of 1,069 pupils in 111 schools, 536 valid outcome scores (11 missing) in the control group and 533 (24 missing) in the intervention group. As we can see in Table 4 the findings from models 1-4 in the main analyses hold in the reduced sample. The intervention effect was slightly stronger in reduced sample ($R^2 = 0.10$) than in the full sample ($R^2 = 0.09$). All other findings were replicated, negligible differences as compared to the findings from the full sample.

Table 4. Effects of the math-app (one billion) intervention, in reduced sample.

| | Model 1 | | | Model 2 | | | Model 3 | | | Model 4 | |
|-------------------------------|----------|------|------|--------------------|------|------|--------------------|------|------|-----------|------|
| <i>Fixed</i> | B | s.e. | sig. | B | s.e. | sig. | B | s.e. | sig. | B | s.e. |
| Intercept | 7.83 | 0.21 | | 0.92 | 0.39 | | | | | 0.30 | 0.41 |
| Pre-test | | | | 0.52 | 0.03 | *** | | | | 0.52 | 0.03 |
| Intervention | | | | | | | | | | 1.14 | 0.30 |
| <i>Random</i> | | | | | | | | | | | |
| Residual (e_{0ij}) | 14.41 | 0.66 | | 10.96 | 0.50 | | | | | 10.96 | 0.50 |
| Intercept (σ^2_{u0}) | 3.18 | 0.64 | | 1.66 | 0.38 | | | | | 1.32 | 0.34 |
| Slope (σ^2_{u1}) | | | | | | | | | | | |
| Int-slope (σ_{u0u1}) | | | | | | | | | | | |
| Variance proportions | | | | Explained variance | | | Explained variance | | | Explained | |
| variance | | | | | | | | | | | |
| School | 0.18 | | | 0.48 | | | n/a | | | 0.58 | |
| Pupil | 0.82 | | | 0.24 | | | n/a | | | 0.24 | |
| -2 LL | 6011.87 | | | 5692.28 | | | did | not | | converge | |
| | 5678.694 | | | | | | | | | | |
| Δ -2LL | | | | 319.59 | | *** | | | | 13.59 | |

Note: Significances for intercepts and variances not reported. Explained variances were not calculated as sample sizes differed across the three models. All models were run in SPSS 25 using the maximum likelihood estimator with raw (non-centred) predictors. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 0.001$.

Table 5. Differential effects of students eligible for free school meals, in reduced sample.

| | Model 5 | | | Model 6 | | |
|--------------------------|---------|------|----------|---------|------|------|
| <i>Fixed</i> | B | s.e. | sig. | B | s.e. | sig. |
| Intercept | 0.49 | 0.42 | | -0.20 | 0.74 | |
| Pre-test | 0.52 | 0.03 | *** | 0.53 | 0.05 | *** |
| Intervention | 1.40 | 0.32 | *** | 0.33 | 0.42 | |
| FSM | -0.57 | 0.36 | | | | |
| FSM x intervention | -0.99 | 0.51 | p = .051 | | | |
| Boy | | | | | | |
| Boy x intervention | | | | | | |
| <i>Random</i> | | | | | | |
| Residual (e_{0ij}) | 10.76 | 0.49 | | 9.27 | 0.95 | |
| Intercept (s^2_{u0}) | 1.27 | 0.33 | | 0.53 | 0.59 | |

Note: Significances for intercepts and variances not reported. Explained variances were not calculated as sample sizes differed across the three models. All models were run in SPSS 25 using the maximum likelihood estimator with raw (non-centred) predictors. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 0.001$.

Table 6. Implementation fidelity effects on outcome (in intervention group only) in reduced sample.

| | Model 9 | | | Model 10 | | |
|---|---------|------|------|----------|------|------|
| | B | s.e. | sig. | B | s.e. | sig. |
| <i>Fixed</i> | | | | | | |
| Intercept | 1.08 | 2.42 | | -0.14 | 2.39 | |
| Pre-test | 0.51 | 0.17 | ** | 0.64 | 0.15 | *** |
| No of app sessions | 0.01 | 0.06 | | | | |
| Pre-test x No of app sessions | 0.00 | 0.00 | | | | |
| TA's compliance with expectations of support | | | | 0.66 | 1.20 | |
| Pre-test x TA's compliance with expectations of support | | | | -0.05 | 0.08 | |
| <i>Random</i> | | | | | | |
| Residual (e_{0ij}) | 11.02 | 0.72 | | 10.53 | 0.92 | |
| Intercept (s^2_{u0}) | 1.90 | 0.60 | | 2.07 | 0.82 | |

Note: Significances for intercepts and variances not reported. Explained variances were not calculated as sample sizes differed across the three models. All models were run in SPSS 25 using the maximum likelihood estimator with raw (non-centred) predictors. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 0.001$.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Hox, J. (2002). *Multilevel analysis: techniques and applications*. Mahwah: Erlbaum.

Syntax for running the models

* Encoding: UTF-8.

* GET FILE='C:\onebillion\Pre-post dataset corrected, checked, anonymised 8Oct2018.sav'.
* DATASET NAME onebill WINDOW=FRONT.
* DATASET ACTIVATE onebill .

***** .

GET FILE='C:\onebillion\Pre-post dataset corrected, checked, anonymised 15Nov2018.sav'.
DATASET NAME new WINDOW=FRONT.

DATASET ACTIVATE new .

** this dataset has the additional 3 cases in the DV at Time 2 .
DESCR Post_minus_8a .

FREQ Post_minus_8a / STA MEA STD MIN MAX SKE KUR / HIST NORM .

EXAMINE VARIABLES=Post_minus_8a

/PLOT BOXPLOT STEMLEAF HISTOGRAM NPLOT

/COMPARE GROUPS

/STATISTICS DESCRIPTIVES EXTREME

/INTERVAL 95

/MISSING LISTWISE

/NOTOTAL.

* AND the supplementary analyses with schools 4900 and 10200 removed,

COMPUTE supp_filt = 1 .

IF (SchoolID = 4900 OR SchoolID = 10200) supp_filt = 0 .

FREQ supp_filt .

TEMP .

SELECT IF (supp_filt = 1) . FREQ

Post_minus_8a .

SORT CASES BY pupilid .

FREQ pupilid SchoolID Allocation .

**** compute valid outcome **** . COMPUTE

validoutcome = 0 .

IF (Post_minus_8a GE 0) validoutcome = 1 . FREQ

validoutcome .

TEMP .

SELECT IF (supp_filt = 1) . FREQ

validoutcome .

*** Tables 5 and 6 *** .

SORT CASES BY allocation .

SPLIT FILE BY allocation .

FREQ Post_minus_8a / STA MEA STD SEMEAN .

DESCR Post_minus_8a / STA MEA SEMEAN .

SPLIT FILE OFF .

T-TEST GROUPS=Allocation(0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

* two-level model (all) variance component .

MIXED Post_minus_8a

/PRINT = SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept

/SAVE pred (bill_pr1)

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

* ICC for pretest .

MIXED Pre_TotalRawScore

/PRINT = SOLUTION TESTCOV

/METHOD=REML

/FIXED intercept

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

* two-level model (all) fixed effect of pretest .

MIXED Post_minus_8a WITH Pre_TotalRawScore

/PRINT = SOLUTION TESTCOV

/METHOD=REML

/FIXED intercept Pre_TotalRawScore

/SAVE pred (bill_pr2)

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

```
.
```

```
* two-level model (all) fixed and random effects of pretest .
```

```
MIXED Post_minus_8a WITH Pre_TotalRawScore
```

```
/CRITERIA=CIN(95) MXITER(100000) MXSTEP(10) SCORING(1) SINGULAR(0.000000000001)  
HCONVERGE(0.01,
```

```
ABSOLUTE) LCONVERGE(0.01,, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
```

```
/PRINT = SOLUTION TESTCOV
```

```
/METHOD=ML
```

```
/FIXED intercept Pre_TotalRawScore
```

```
/RANDOM intercept Pre_TotalRawScore | SUBJECT (SchoolID) COVTYPE (UN) . EXE
```

```
.
```

```
** doesn't converge **** .
```

```
* /SAVE pred (bill_pr3)
```

```
FREQ Allocation .
```

```
* two-level model (all) fixed effect of pretest and condition .
```

```
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation
```

```
/PRINT = SOLUTION TESTCOV
```

```
/METHOD=ML
```

```
/FIXED intercept Pre_TotalRawScore allocation
```

```
/SAVE pred (bill_pr4)
```

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

```
.
```

```
FREQ FSMEligibility allocation .
```

```
CRO FSMEligibility BY allocation .
```

```
RECODE FSMEligibility ('0' = 0)('1' = 1) INTO fsm . FORM  
fsm (F4.0) .
```

```
FREQ fsm .
```

```
CRO FSMEligibility BY fsm .
```

*** Tables 2,3 and 4 *** .
SORT CASES BY fsm .
SPLIT FILE BY fsm .

FREQ Post_minus_8a / STA MEA STD SEMEAN . DESCR
Post_minus_8a / STA MEA STD SEMEAN . SPLIT FILE
OFF .

CRO allocation BY FSMeligibility BY validoutcome / MIS INC. SORT
CASES BY allocation .

SPLIT FILE BY allocation .
TEMP .

SELECT IF (fsm = 0) .

FREQ Post_minus_8a / STA MEA STD SEMEAN .
TEMP .

SELECT IF (fsm = 1) .

FREQ Post_minus_8a / STA MEA STD SEMEAN .
SPLIT FILE OFF .

TEMP .

SELECT IF (fsm = 0) .

T-TEST GROUPS=allocation (0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

TEMP .

```
SELECT IF (fsm = 1) .
```

```
T-TEST GROUPS=allocation (0 1)
```

```
/MISSING=ANALYSIS
```

```
/VARIABLES=Post_minus_8a
```

```
/CRITERIA=CI(.95).
```

```
TEMP .
```

```
SELECT IF (allocation = 0) .
```

```
DESCR Post_minus_8a / STA MEA STD SEMEAN . TEMP  
.
```

```
SELECT IF (allocation = 0) .
```

```
FREQ Post_minus_8a / STA MEA STD SEMEAN .  
SPLIT FILE OFF .
```

```
TEMP .
```

```
SELECT IF (allocation = 0) .
```

```
T-TEST GROUPS=fsm(0 1)
```

```
/MISSING=ANALYSIS
```

```
/VARIABLES=Post_minus_8a
```

```
/CRITERIA=CI(.95).
```

```
SPLIT FILE BY fsm .
```

```
TEMP .
```

```
SELECT IF (allocation = 1) .
```

```
FREQ validoutcome .
```

```
TEMP .
```

```
SELECT IF (allocation = 1) .
```

```
DESCR Post_minus_8a / STA MEA STD SEMEAN . TEMP  
.
```

```
SELECT IF (allocation = 1) .
```

```
FREQ Post_minus_8a / STA MEA STD SEMEAN .  
SPLIT FILE OFF .
```


TEMP .

SELECT IF (allocation = 1) .
T-TEST GROUPS=fsm(0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

SORT CASES BY gender .
SPLIT FILE BY gender .

DESCR Post_minus_8a / STA MEA SEMEAN . FREQ
Post_minus_8a / STA MEA STD SEMEAN . SPLIT
FILE OFF .

T-TEST GROUPS=gender(0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

* two-level model (all) fixed effect of pretest and condition and FSM .
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation fsm
/PRINT = SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept Pre_TotalRawScore allocation fsm fsm*allocation

/SAVE pred (bill_pr5)

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

```
*** grand mean centered predictor *** .  
COMPUTE sample = 1 . AGGREGATE  
  
/OUTFILE=* MODE=ADDVARIABLES  
  
/BREAK = sample  
  
/Pre_TotalRawScore_m = MEAN(Pre_TotalRawScore) .  
DESCR Pre_TotalRawScore_m .  
  
COMPUTE Pre_TotalRawScore_gmc = Pre_TotalRawScore - Pre_TotalRawScore_m . DESCR  
Pre_TotalRawScore_gmc .
```

```
MIXED Post_minus_8a WITH Pre_TotalRawScore_gmc allocation fsm  
  
/PRINT = SOLUTION TESTCOV  
  
/METHOD=ML  
  
/FIXED intercept Pre_TotalRawScore_gmc allocation fsm fsm*allocation  
  
/SAVE pred (bill_gmc_pr5)  
  
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE  
.
```

```
* two-level model (all) fixed effect of pretest and condition for FSM-eligible children .  
TEMP .  
SELECT IF (fsm = 1) .  
  
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation  
  
/PRINT = SOLUTION TESTCOV  
  
/METHOD=ML  
  
/FIXED intercept Pre_TotalRawScore allocation  
  
/SAVE pred (bill_pr6)  
  
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE  
.
```

```
*** gender models **** .  
FREQ gender .  
  
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation gender  
  
/PRINT = SOLUTION TESTCOV
```

/METHOD=ML

/FIXED intercept Pre_TotalRawScore allocation gender gender*allocation

/SAVE pred (bill_pr7)

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

** GET FILE="C:\onebillion\Process variables and matching PupilID's anonymised 08Nov2018.sav".

* DATASET NAME fidelity WINDOW=FRONT.

* DATASET ACTIVATE fidelity .

* SORT CASES BY pupilid .

FREQ Number_sessions_attended Compliance ta_sum pa_sum max_session Stopping_point
No_Cert_Missing No_Cert_Achieved Skipped What_TA_did View_Role TA_Style_observation
session_expectations TA_OR_Teaching / STA MEA STD MIN MAX SKE KUR .

** controls for fidelity *** .

FREQ Number_sessions_attended Compliance ta_sum pa_sum max_session Stopping_point
No_Cert_Missing No_Cert_Achieved Skipped What_TA_did View_Role TA_Style_observation
session_expectations TA_OR_Teaching / STA MEA STD MIN MAX SKE KUR .

* n sessions attended ** .

SELECT IF (allocation = 1) .

MIXED Post_minus_8a WITH Pre_TotalRawScore Number_sessions_attended

/PRINT = DESCR SOLUTION TESTCOV

/METHOD=ML

```
/FIXED      intercept      Pre_TotalRawScore      Number_sessions_attended  
Pre_TotalRawScore*Number_sessions_attended
```

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

```
.
```

```
FREQ TA_Style_observation session_expectations .
```

```
* n session expectations ** .
```

```
SELECT IF (allocation = 1) .
```

```
MIXED Post_minus_8a WITH Pre_TotalRawScore session_expectations
```

```
/PRINT = DESCR SOLUTION TESTCOV
```

```
/METHOD=ML
```

```
/FIXED      intercept      Pre_TotalRawScore      session_expectations  
Pre_TotalRawScore*session_expectations
```

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

```
.
```

```
GET FILE='C:\onebillion\Pre-post dataset corrected, checked, anonymised 15Nov2018.sav'. DATASET  
NAME new WINDOW=FRONT.
```

```
DATASET ACTIVATE new .
```

```
** this dataset has the additional 3 cases in the DV at Time 2 . DESCR  
Post_minus_8a .
```

```
FREQ Post_minus_8a .
```

```
* AND the supplementary analyses with schools 4900 and 10200 removed,
```

```
COMPUTE supp_filt = 1 .
```

```
IF (SchoolID = 4900 OR SchoolID = 10200) supp_filt = 0 . FREQ  
supp_filt .
```

```
TEMP .
```

```
SELECT IF (supp_filt = 1) . FREQ  
Post_minus_8a .
```

```
SELECT IF (supp_filt = 1) . SORT  
CASES BY pupilid .
```

FREQ pupilid SchoolID Allocation .

```
**** compute valid outcome **** . COMPUTE  
validoutcome = 0 .
```

```
IF (Post_minus_8a GE 0) validoutcome = 1 . FREQ  
validoutcome .
```

CRO validoutcome BY allocation .

* two-level model (all) fixed effect of pretest .

```
MIXED Post_minus_8a  
/PRINT = SOLUTION TESTCOV
```

```
/METHOD=ML
```

```
/FIXED intercept
```

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

.

* two-level model (all) fixed effect of pretest .

```
MIXED Post_minus_8a WITH Pre_TotalRawScore  
/PRINT = SOLUTION TESTCOV
```

```
/METHOD=ML
```

```
/FIXED intercept Pre_TotalRawScore
```

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

.

```
* two-level model (all) fixed and random effects of pretest .
MIXED Post_minus_8a WITH Pre_TotalRawScore
/CRITERIA=CIN(95) MXITER(100000) MXSTEP(10) SCORING(1) SINGULAR(0.000000000001)
HCONVERGE(0.01,
    ABSOLUTE) LCONVERGE(0.01,, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/PRINT = SOLUTION TESTCOV
/METHOD=ML
/FIXED intercept Pre_TotalRawScore
/RANDOM intercept Pre_TotalRawScore | SUBJECT (SchoolID) COVTYPE (UN) . EXE
.
** doesn't converge **** .
* /SAVE pred (bill_pr3)

FREQ Allocation .

* two-level model (all) fixed effect of pretest and condition .
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation
/PRINT = SOLUTION TESTCOV
/METHOD=ML
/FIXED intercept Pre_TotalRawScore allocation
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
.

FREQ FSMEligibility allocation .
CRO FSMEligibility BY allocation .

RECODE FSMEligibility ('0' = 0)('1' = 1) INTO fsm . FORM
fsm (F4.0) .

FREQ fsm .

CRO FSMEligibility BY fsm .

*** Tables 2,3 and 4 *** .
SORT CASES BY fsm .
SPLIT FILE BY fsm .

FREQ Post_minus_8a / STA MEA STD SEMEAN . DESCR
Post_minus_8a / STA MEA STD SEMEAN . SPLIT FILE
OFF .
```

CRO allocation BY FSMeligibility BY validoutcome / MIS INC. SPLIT
FILE BY fsm .

TEMP .

SELECT IF (allocation = 0) .
FREQ validoutcome .
TEMP .

SELECT IF (allocation = 0) .

DESCR Post_minus_8a / STA MEA STD SEMEAN . TEMP

.

SELECT IF (allocation = 0) .

FREQ Post_minus_8a / STA MEA STD SEMEAN .
SPLIT FILE OFF .

TEMP .

SELECT IF (allocation = 0) .
T-TEST GROUPS=fsm(0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

SPLIT FILE BY fsm .

TEMP .

SELECT IF (allocation = 1) .
FREQ validoutcome .
TEMP .

SELECT IF (allocation = 1) .

DESCR Post_minus_8a / STA MEA STD SEMEAN . TEMP
.

SELECT IF (allocation = 1) .

FREQ Post_minus_8a / STA MEA STD SEMEAN .
SPLIT FILE OFF .

TEMP .

SELECT IF (allocation = 1) .
T-TEST GROUPS=fsm(0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

SORT CASES BY gender .
SPLIT FILE BY gender .

DESCR Post_minus_8a / STA MEA SEMEAN . FREQ
Post_minus_8a / STA MEA STD SEMEAN . SPLIT
FILE OFF .

T-TEST GROUPS=gender(0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

* two-level model (all) fixed effect of pretest and condition and FSM .
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation fsm
/PRINT = SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept Pre_TotalRawScore allocation fsm fsm*allocation


```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

```
.
```

```
* two-level model (all) fixed effect of pretest and condition for FSM-eligible children .
```

```
TEMP .
```

```
SELECT IF (fsm = 1) .
```

```
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation
```

```
/PRINT = SOLUTION TESTCOV
```

```
/METHOD=ML
```

```
/FIXED intercept Pre_TotalRawScore allocation
```

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

```
.
```

```
*** gender models **** .
```

```
FREQ gender .
```

```
MIXED Post_minus_8a WITH Pre_TotalRawScore allocation gender
```

```
/PRINT = SOLUTION TESTCOV
```

```
/METHOD=ML
```

```
/FIXED intercept Pre_TotalRawScore allocation gender gender*allocation
```

```
/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE
```

```
.
```

```
** GET FILE="C:\onebillion\Process variables and matching PupilID's anonymised 08Nov2018.sav".
```

```
* DATASET NAME fidelity WINDOW=FRONT.
```

```
* DATASET ACTIVATE fidelity .
```

* SORT CASES BY pupilid .

FREQ Number_sessions_attended Compliance ta_sum pa_sum max_session Stopping_point
No_Cert_Missing No_Cert_Achieved Skipped What_TA_did View_Role TA_Style_observation
session_expectations TA_OR_Teaching / STA MEA STD MIN MAX SKE KUR .

** controls for fidelity ** .

FREQ Number_sessions_attended Compliance ta_sum pa_sum max_session Stopping_point
No_Cert_Missing No_Cert_Achieved Skipped What_TA_did View_Role TA_Style_observation
session_expectations TA_OR_Teaching / STA MEA STD MIN MAX SKE KUR .

* n sessions attended ** .

SELECT IF (allocation = 1) .

MIXED Post_minus_8a WITH Pre_TotalRawScore Number_sessions_attended

/PRINT = DESCR SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept Pre_TotalRawScore Number_sessions_attended
Pre_TotalRawScore*Number_sessions_attended

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

FREQ TA_Style_observation session_expectations .

* n session expectations ** .

SELECT IF (allocation = 1) .

MIXED Post_minus_8a WITH Pre_TotalRawScore session_expectations

/PRINT = DESCR SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept Pre_TotalRawScore session_expectations
Pre_TotalRawScore*session_expectations

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

T-TEST GROUPS=gender(0 1)

/MISSING=ANALYSIS

/VARIABLES=Post_minus_8a

/CRITERIA=CI(.95).

MIXED Post_minus_8a WITH gender

/PRINT = SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept gender

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

** include school level gender (% boys) *** . AGGREGATE

/OUTFILE=* MODE=ADDVARIABLES

/BREAK=SchoolID

/Gender_fin = FIN(Gender 1 1).

FREQ Gender_fin .

CRO Gender_fin BY schoolid .

CRO Gender_fin BY allocation .

MIXED Post_minus_8a WITH gender gender_fin allocation

/PRINT = SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept gender gender_fin allocation gender_fin*allocation

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

CRO gender BY fsm .

CRO gender BY fsm BY allocation .

MIXED Post_minus_8a WITH gender fsm allocation

/PRINT = SOLUTION TESTCOV

/METHOD=ML

/FIXED intercept gender fsm allocation fsm*gender

/RANDOM intercept | SUBJECT (SchoolID) COVTYPE (ID) . EXE

.

Appendix 4: Instruments used for implementation and process evaluation

Onebillion: app-based maths. Feedback about training workshop

Name of person attending:TA/Link teacher/Head
(please circle)

Name of School :

Your feedback will help the project team think about what works best in training for the use of the apps. Please contribute to it by answering all the questions as accurately as you can. The data will be gathered into a secure website in Oxford and will be anonymised, but we need to collect information on which schools are represented in the survey. No identified data will be made available to anyone later on.

Instructions: Please tick one option for each question. Please do not tick more than one option nor tick in between boxes. Thank you for your help.

| Question | | | | | |
|--|----------|-------------------|---------------------------|----------------|-------|
| | Disagree | Somewhat disagree | Neither agree or disagree | Somewhat agree | Agree |
| 1. The aims of the programme were clear. | | | | | |
| 2. I understand the structure of the onebillion project. | | | | | |
| 3. I understand the content of the onebillion apps. | | | | | |
| 4. I feel confident to support children using the apps. | | | | | |

| | | | | | |
|--|--|--|--|--|--|
| 5. I am clear about my daily tasks. | | | | | |
| 6. I am clear about my weekly tasks. | | | | | |
| 7. I feel the implementation manual contains all I need to know. | | | | | |
| 8. I feel confident to complete session logs. | | | | | |
| 9. I feel confident to submit weekly digital progress logs. | | | | | |
| 10. I can use an iPad with confidence. | | | | | |

| | No | Yes (Please detail below) | | | |
|--|-------|---------------------------|-------|------------|--------------|
| 11. Are there any aspects of the programme about which you are not clear? | | | | | |
| 12. Is there anything about the implementation which you feel will be more challenging? | | | | | |
| 13. Do you have any questions which were not answered as part of the training? | | | | | |
| 14. Were there any parts of the training day for which there was not enough time allowed? | | | | | |
| | Yes | No (Please detail below) | | | |
| 15. Did you find the 'Overview of the Unlocking Talent programme' a useful part of the training day? | | | | | |
| 16. Did the interactive training videos provide you with sufficient practical instruction? | | | | | |
| | Never | Sometimes | Often | Very Often | All the time |
| 17. Do children in Foundation stage in your school have access to iPads? | | | | | |
| 18. Do children in Y1 in your school have access to iPads? | | | | | |
| 19. Does your school have a computer suite? | YES | | | NO | |
| 20. How many iPads does your school have? (approximate answer if unsure) | | | | | |
| 21. What was the best thing about the training day? | | | | | |

22. What was the most challenging aspect of the training day?

Thank you

Onebillion: app-based maths. Feedback on training package

Name of person training:TA/Link teacher/Head (please circle)

Name of School :

Your feedback will help the project team think about what works best in training for the use of the apps. Please contribute to it by answering all the questions as accurately as you can. The data will be gathered into a secure website in Oxford and will be anonymised, but we need to collect information on which schools are represented in the survey. No identified data will be made available to anyone later on.

Instructions: Please tick one option for each question. Please do not tick more than one option nor tick in between boxes. Thank you for your help.

| Question | | | | | |
|--|----------|-------------------|---------------------------|----------------|-------|
| | Disagree | Somewhat disagree | Neither agree or disagree | Somewhat agree | Agree |
| 1. The aims of the programme were clear. | | | | | |
| 2. I understand the structure of the onebillion project. | | | | | |
| 3. I understand the content of the onebillion apps | | | | | |
| 4. I feel confident to support children using the apps. | | | | | |
| 5. I am clear about my daily tasks. | | | | | |
| 6. I am clear about my weekly tasks. | | | | | |
| 7. I feel the implementation manual contains all I need to know. | | | | | |
| 8. I feel confident to complete session logs. | | | | | |
| 9. I feel confident to submit weekly digital progress logs. | | | | | |
| 10. I can use an iPad with confidence. | | | | | |

| | No | Yes (Please detail below) |
|---|----|---------------------------|
| 11. Are there any aspects of the programme about which you are not clear? | | |
| 12. Is there anything about the implementation which you feel will be more challenging? | | |

| | Yes | No (Please detail below) | | | |
|--|---|--------------------------|-------|------------|--------------|
| 13. Was there enough explanation in the training materials to enable you to understand what you have to do? | | | | | |
| 14. Did you have enough time to look at all the materials before starting the programme? | | | | | |
| 15. Were the questions you had after watching the videos answered to your satisfaction by the intervention team? | | | | | |
| 16. Which of the 7 training videos did you find most useful? | Tick 1 2 3 4 5 6 7 | Comments: | | | |
| | Never | Sometimes | Often | Very Often | All the time |
| 17. Do children in Foundation stage in your school have access to iPads? | | | | | |
| 18. Do children in Y1 in your school have access to iPads? | | | | | |
| 19. Does your school have a computer suite? | YES | | NO | | |
| 20. How many iPads does your school have? (approximate answer if unsure) | | | | | |
| 21. What was the best thing about the training materials? | | | | | |
| 22. What was the most challenging aspect in the training materials? | | | | | |

Thank you

Observation schedule for onebillion app visit

Name of School:

Date:

Name of observer:

Session number:

Name of TA:

Group size:

Start of session:

End of session:

☐
☐

Start of session (Ask TA if possible to identify discretely 3 children across the range for you to follow)

- | | | |
|---|---|---|
| 1. Is the intervention taking place in a suitable area? (Circle) classroom / corridor / IT room / spare classroom / Other | Y | N |
| 2. Are the iPads numbered so the children know which is theirs? | Y | N |
| 3. Are the iPads laid out on tables ready for the children to start? | Y | N |

4. Are there any children not using headphones? (number)

5. How long does it take for the children to sit down and start the games?

mins

6. Are the children able to do the following independently without TA help?

i) Get headphones: all / some / none

ii) Find iPad: all / some / none

iii) Select their name in the app: all / some / none

iv) Find their starting point in the app: all / some / none

Observe 3 children, 5 mins each playing games

| Child ID | gave technical support | gave pedagogical support | repeated instructions | went to next flashing game | off task | stopped early | completed 30 Minutes | won certificate and reaction |
|----------|------------------------|--------------------------|-----------------------|----------------------------|----------|---------------|----------------------|------------------------------|
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |

During the session, **the TA**

| Child ID | gave technical support | gave pedagogical support | restarted app for child | listened to instructions with a child having difficulty | explained concept/task to child | worked together letting child touch the screen to answer | intervened when another child touched the screen of another child | intervened when another child touched the screen of another child | TA response time to questions |
|----------|------------------------|--------------------------|-------------------------|---|---------------------------------|--|---|---|-------------------------------|
| 1 | | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | | | | | | | | | |
| 4 | | | | | | | | | |
| 5 | | | | | | | | | |
| 6 | | | | | | | | | |
| 7 | | | | | | | | | |
| 8 | | | | | | | | | |
| 9 | | | | | | | | | |
| 10 | | | | | | | | | |

End of session

- | | | |
|---|---|---|
| 7. Were all the i-Pads charged sufficiently for the duration of the session? | Y | N |
| 8. Did all the children press the back button to return to the names screen? | Y | N |
| 9. Did all the children put their headphones away for next session? | Y | N |
| 10. Did the TA fill in the session log: during the session/ after the session / both/ neither | | |

Ask the children:

11. Which games did you enjoy playing? (list games/concepts)

Extra notes:

TA post observation interview

Assure TA of confidentiality for any answers given.

Training

1. How were you chosen to be the person running this intervention in your school?

.....

.....

.....

2. Did you attend a training event? Y N Birmingham/Nottingham/Manchester

3. Does any other member of staff also deliver this intervention?

☐ Y N

4. If so, did this member of staff attend training?
N/A

☐ Y N

5. Did another member of your staff attend the training instead of you?

Y N N/A

Head/class teacher/another TA/other (detail)

Birmingham/Nottingham/Manchester(see 6a)

☐

6. a) If so, how did they cascade information of the training day to you?
(tick if appropriate)

☐

Went through the information from the day with me

Passed on the materials without discussion

Other (detail)

N/A

b) How did you cascade to other members of staff delivering the intervention but not attending training?

Went through the information from the day with them Passed

on the materials without discussion

Other (detail)

N/A

7. a) If you were not at training, did you watch the on-line videos? Y N N/A

b) If they were not at training, did they watch the on-line videos? Y N N/A

8. If you had any questions after watching the videos, or after training, did you have an opportunity for your questions to be answered? Y N

Running the intervention

9. Where does the intervention usually take place? Corridor /classroom /IT suite /spare room / other (detail)

10. What size group do you normally run? 10 / 2 x 5/ other (detail)

11. Have you changed, or do you plan to change, the group size since you started? Y N

If yes, reason for change.....

12. Do you have a fixed schedule to run the intervention? Y N

If yes, (circle) M T W TH F

If no, how do you arrange the flexible timetable?

.....

Children

13. Do the children generally enjoy doing the app and stay on task for 30 minutes?

Y N

14. Did you have the impression that some children had never used an iPad before?

Y N

15. What do you do if a child answers on screen for another child?

.....

.....

Technical support

16. Have you needed to give children technological support ?

(circle) each day/ each week/ not at all

17. What has this entailed? (circle) headphones / iPads /other

.....

.....

18. Do you feel you have been able to solve these issues satisfactorily?

Y N

19. Do you have any tips to pass on to other TAs who might do this in the future?

Y N

.....

.....

Pedagogical support

20. Have you had to give pedagogical support? (circle) each day/ each week/ not at all

21. Do you feel confident to give this support? (circle) not at all/ somewhat/ all the time

22. Do you have any tips for other TAs who might do this in the future? Y N

.....

.....

23. What do you do if a child doesn't pass a quiz?

.....

.....

24. Have you used the forum for your area?

Y N

25. Do you know how and when to use this?

Y N

Daily/weekly tasks

26. In this project you are asked to keep an attendance log; if you were not keeping this log, how would you keep track of children's attendance?

.....

27. How are you catching up absent children?

.....

.....

28. How do you view your role in the intervention?

.....

.....

29. Do you feel confident in this role? Y N

30. Any other comments you would like to make about the intervention?

.....

.....

.....

.....

.....

.....

.....

Interviewer

onebillion Middle Management Phone Interview (Control)

Introduction

Hello <Link teacher name>. My name is <researcher name> and I am calling on behalf of the University of Oxford regarding the phone interview for onebillion that was arranged via email last week.

Thank you for agreeing to speak to me today. I have some brief questions that I would like to go through with you. We are independent evaluators and any information you give is confidential.

The interview will take about 5-10 minutes and consists of 6 main questions. All of your details will be anonymised.

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

| Intervention name | How often? | Which Children? (HA, LA etc) | TA or Teacher? | Training needed? | Uses iPad or computer? | Group size |
|-------------------|------------|----------------------------------|----------------|------------------|------------------------|------------|
| | | | | | | |

1. Do your year 1 children use iPads and/or maths software in their maths and/or IT lessons?

- How often?
- What software do they use?
- How many iPads in school for the class to use?

2. Which maths interventions are carried out in your school for Year 1 children needing support?

- Summary

- For each intervention names, ask:

Is the intervention part of the curriculum or part of an extra intervention?

3. What are your main priorities in your School Improvement Plan for maths this year?

- Brief description of the school's maths improvement plan

4. What are your main priorities in your School Improvement Plan for IT this year?

- Brief description of the school's IT improvement plan

5. Do you intend to use the onebillion apps next year when they are available to you? (If so, how?)

- How many children do you think might use the apps within the next academic year?
- Do you already have the resources for this to happen next year or will you need further resources? If so, what?

6. Is there anything else that you would like to add about Year 1 maths interventions in your school, that we have not covered?

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

If you have any further questions please do not hesitate to contact me on 01865 284893

onebillion Middle Management Phone Interview (Intervention)

Introduction

Hello <Link teacher name>. My name is <researcher name> and I am calling on behalf of the University of Oxford regarding the phone interview for onebillion that was arranged via email last week.

Thank you for agreeing to speak to me today. I have some brief questions that I would like to go through with you. We are independent evaluators so any information you give is confidential.

The interview will take about 5-10 minutes and consists of 9 main questions. I will refer to all of the children that have used the onebillion maths apps as the nominated children. All your details will be anonymised.

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

1. Do your year 1 children use iPads and/or maths software in their maths lessons?

- How often?
- What software do they use?

2. Are the nominated children using the onebillion maths apps instead of regular maths lessons?

- Are they taken out of the class to do the onebillion maths intervention?

3. The nominated children were selected because the teachers thought that they could use extra support for maths. Have some or all of the nominated children received extra support beyond the apps?

4. Other than onebillion, which additional maths interventions are carried out in Year 1 for children needing support?

• Summary

| Intervention name | How often? | Which Children? HA, LA etc | TA or Teacher? | TA Training needed? | Uses iPad or computer? | Group size |
|-------------------|------------|-------------------------------|----------------|---------------------|------------------------|------------|
| | | | | | | |

For each intervention named, ask:

- Is the intervention part of the curriculum or an extra intervention?

5. What are your main priorities in your School Improvement Plan for maths this year?

- Brief description of the school's maths improvement plans

6. What are your main priorities in your School Improvement Plan for IT this year?

- Brief description of the school's IT development plan

7. Do you intend to use the onebillion apps next year?

- How many children do you think will use the apps within the next academic year?
- Do you already have the resources for this to happen next year or will you need further resources? If so, what?

8. Is there anything else that you would like to add about Year 1 maths interventions in your school, that we have not covered?

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

If you have any further questions please do not hesitate to contact me on 01865 284893.

TA intervention questionnaire week 9

Q1. Do you have a dedicated space in which to deliver the intervention?

☐ yes

☐ no

Q2. Where do you deliver the intervention?

☐ corridor

☐ classroom

☐ IT suite

☐ spare classroom

☐ library

☐ intervention room

☐ staff room

☐ other

Other (please detail)

Q3. Which size group do you normally run?

☐ 9 / 10

☐ 2 groups of 5

☐ other

Please detail what group sizes were run

Q4. Reason for choosing this grouping? (select all that apply)

- ☐ number of iPads available
- ☐ space
- ☐ class timetabling
- ☐ preference of group size
- ☐ other

If other please detail.

Q5. Do you have enough iPads to deliver to a group of 10 if desired?

- ☐ yes
- ☐ no

Q6. Have you changed the group size since you started?

☐ yes

☐ no

Please detail grouping arrangement and reason for the change.

Q7. Do you have a fixed schedule for the intervention? (i.e. same days/times)

☐ yes

☐ no

Please detail how you arrange the flexible timetable

Q8. Did you have any problems ensuring you had the iPads to run the intervention due to other classes using them?

☐ yes

☐ no

Q9. Did you have any problems ensuring you had the headphones to run the intervention due to other classes using them?

☐ yes

☐ no

Q10. How long does it take you to prepare for a session?

Q11. Was it easy to fit the sessions into the class timetable?

☐ yes

☐ no

Q12. How easy did you find it to set up the apps?

- ☐ easy
- ☐ some issues
- ☐ required support

Q13. Do the children generally enjoy doing the 3-5 app and stay on task for 30 minutes?

- ☐ all of them
- ☐ most of them
- ☐ some of them
- ☐ none of them

Q14. Do the children generally enjoy doing the 4-6 app and stay on task for 30 minutes?

- ☐ all of them
- ☐ most of them

☐ some of them

☐ none of them

Q15. Did you have the impression that some of the children had never used an iPad before?

☐ yes

☐ no

Q16. If a child answers on screen for another child, do you:

- ☐ let the child complete the question and then remind them not to do this
- ☐ stop them straight away from answering a question on the other child's iPad
- ☐ this hasn't happened in my sessions

Q17. How often have you needed to give children technological support when they used the 3-5 app?

- ☐ very often
- ☐ often
- ☐ rarely
- ☐ never

Q18. How often have you needed to give children technological support when they used the 4-6 app?

- ☐ very often

☐ often

☐ rarely

☐ never

Q19. What has this technological help entailed?

- ☐ headphone issues
- ☐ iPad issues
- ☐ apps freezing
- ☐ other

Other (please detail).

Q20. Do you feel you have been able to solve these issues satisfactorily?

- ☐ yes
- ☐ no

Q21. Did you need support at any time from the IT technician or person with responsibility for IT?

- ☐ yes
- ☐ no

Please detail the support needed/given

Q22. Do you have any technology related tips to pass on to other TAs doing this in the future?

☐ yes

☐ no

Please detail any tips you have

Q23. How often have you needed to give children pedagogical support when they used the 3-5 app?

☐ very often

☐ often

☐ rarely

☐ never

Q24. Do you feel confident to give this support?

☐ not at all

☐ somewhat

☐ all the time

Q25. How often have you needed to give children pedagogical support when they used the 4-6 app?

☐ very often

☐ often

☐ rarely

☐ never

Q26. Do you feel confident to give this support?

- ☐ not at all
- ☐ somewhat
- ☐ all the time

Q27. Did you seek support for any pedagogical issues?

- ☐ yes
- ☐ no

Q28. If you have given pedagogical support, did you usually use any materials for this?

- ☐ Yes
- ☐ No

Please detail what materials you used

Q29. Do you have any pedagogical tips for other TAs who may deliver this intervention in the future?

☐ yes

☐ no

Please detail any tips you have

Q30. What do you usually do if a child repeatedly doesn't pass a quiz

Q31. How do you view your role in the intervention?

Q32. Have you used your forum area on iTunes U?

☐ yes

☐ no

Q33. How many children have completed both apps so far?

Did these children choose to:

☐ repeat games

☐ stop playing

Q35. What is the best thing about this intervention?

Q36. What is the most challenging thing about this intervention?

Middle Management Control Questionnaire

Q1. Do any of the nominated children receive any other extra support sessions for maths, including 1 to 1 or small group, outside of regular maths lessons?

☐ Yes

☐ No

Please complete the form detailing interventions for each child.

| | Child's initials | Intervention name | Number of sessions | Duration of each session | Delivered by (TA/Teacher etc) | Did it use iPads or computers? | 1-1 / small group / large group | What materials were used? |
|---|------------------|-------------------|--------------------|--------------------------|-------------------------------|--------------------------------|---------------------------------|---------------------------|
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| 7 | | | | | | | | |
| 8 | | | | | | | | |

| | | | | | | | | |
|----|--|--|--|--|--|--|--|--|
| 9 | | | | | | | | |
| 10 | | | | | | | | |

Q2. When you were selecting children to be nominated for the onebillion intervention, did you include only children who do not have difficulties with understanding English?

☐ Yes

☐ No

Q3. When you were selecting children to be nominated for the onebillion intervention, did you exclude children with Statements for SEN or ECH plans?

☐ Yes

☐ No

Q4. Have the nominated children had previous experience in school of using iPads?

☐ Yes

☐ No

If "No", how many did not have previous experience of using iPads in school?

Q5. Do you have a dedicated IT lead teacher/co-ordinator in your school?

☐ Yes

☐ No

Q6. Do teachers have access to an IT technician in school if they have any problems?

☐ Yes

☐ No

Q7. Does the school have an induction procedure for staff which includes IT?

☐ Yes

☐ No

Q8. Can you confirm that the nominated children have not been using the onebillion apps in school?

☐ Yes

☐ No

Q9. Would you have had to buy any iPads specifically for this intervention, if you had been allocated to the intervention group?

☐ Yes

☐ No

If "Yes", how many iPads did you have to buy?

Q10. Would you have had to buy any headphones specifically for this intervention, if you had been allocated to the intervention?

☐ Yes

☐ No

If "Yes", how many headphones did you have to buy?

Q11. What IT software is used with Y1 children, if any?

Q12. What IT software is used in:

| | |
|--------|---------------|
| | |
| | Software used |
| | |
| | |
| EYFS | |
| Year 2 | |
| KS2 | |

Q13. How much did your school spend on IT hardware, approximately to the nearest thousand pounds during 2017-2018?

Q14. How much did your school spend on IT software, approximately to the nearest hundred pounds during 2017-2018?

Middle Management Intervention Questionnaire

Q1. Do any of the nominated children receive any other extra support sessions for maths, including 1 to 1 or small group, outside of regular maths lessons?

☐ Yes

☐ No

Please complete the form detailing interventions for each child.

| | Child's initials | Intervention name | Number of sessions | Duration of each session | Delivered by (TA/Teacher etc) | Did it use iPads or computers? | 1-1 / small group / large group | What materials were used? |
|---|------------------|-------------------|--------------------|--------------------------|-------------------------------|--------------------------------|---------------------------------|---------------------------|
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| 7 | | | | | | | | |
| 8 | | | | | | | | |
| 9 | | | | | | | | |

| | | | | | | | | |
|----|--|--|--|--|--|--|--|--|
| 10 | | | | | | | | |
|----|--|--|--|--|--|--|--|--|

Q2. When you were selecting children to be nominated for the onebillion intervention, did you include only children who do not have difficulties with understanding English?

☐ Yes

☐ No

Q3. When you were selecting children to be nominated for the onebillion intervention, did you exclude children with Statements for SEN or ECH plans?

☐ Yes

☐ No

Q4. Have the onebillion maths apps replaced usual maths interventions for the nominated children?

☐ Yes

☐ No

Q5. Have the nominated children had previous experience in school of using iPads?

☐ Yes

☐ No

If "No", how many did not have previous experience of using iPads in school?

Q6. Do you have a dedicated IT lead teacher/co-ordinator in your school?

☐ Yes

☐ No

Q7. Do teachers have access to an IT technician in school if they have any problems?

☐ Yes

☐ No

Q8. Does the school have an induction procedure for staff which includes IT?

☐ Yes

☐ No

Q9. Has your school bought any additional licences for the onebillion apps?

☐ Yes

☐ No

Q10. Do you know if any parents of children involved in the onebillion intervention have bought either of the apps?

☐ Yes

☐ No

☐ I don't know

Q11. Did you have to buy iPads specifically for this intervention?

☐ Yes

☐ No

If "Yes", how many iPads did you have to buy?

Q12. Did you have to buy headphones specifically for this intervention?

☐ Yes

☐ No

If "Yes", how many headphones did you have to buy?

Q13. Did the use of iPads for the onebillion intervention prevent their use by other classes?

☐ Yes

☐ No

If yes, did any activities involving iPads have to be cancelled with other classes?

☐ Yes

☐ No

Q14. What IT software is used with Y1 children, if any?

Q15. What IT software is used in:

| | Software used |
|--------|---------------|
| EYFS | |
| Year 2 | |
| KS2 | |

Q16. How much did your school spend on IT hardware, approximately to the nearest thousand pounds during 2017-2018?

Q17. How much did your school spend on IT software, approximately to the nearest hundred pounds during 2017-2018?

Q18. Do the onebillion apps fit with your School Improvement Plan for maths?

☐ Yes

☐ No

Q19. Do the onebillion apps fit with your School Improvement Plan for IT?

☐ Yes

☐ No

Q20. Does the rest of the school know that this intervention is running in your school?

☐ Yes

☐ No

☐ I don't know

Q21. What impact if any, has this intervention had in the school?

Appendix 5: Schools' location and dates of observations

The selection of schools for observation and post-observation interviews with TA.

| Anonymous school ID | School location | Date observation and interview carried out | Week of intervention | 3-5 App | Children observed working on | |
|---------------------|-----------------|--|----------------------|---------|------------------------------|---|
| | | | | | | |
| 4800 | North West | 18.04.18 | 3 | 5 | 2 | |
| 8200 | North West | 26.03.18 | 4 | 2 | 3 | |
| 9000 | North West | 26.03.18 | 4 | 6 | 3 | 1 |
| 10300 | West Midlands | 27.03.18 | 4 | 0 | 9 | |
| 10500 | North West | 17.04.18 | 4 | 2 | 3 | |
| 2300 | Yorkshire | 09.05.18 | 4 | 5 | 4 | |
| 4700 | North West | 17.05.18 | 5 | 3 | 5 | |
| 1600 | North West | 25.04.18 | 6 | 1 | 8 | |
| 5600 | West Midlands | 26.04.18 | 6 | 1 | 9 | |
| 5200 | North West | 18.04.18 | 7 | 0 | 9 | |
| 6400 | North West | 26.04.18 | 7 | 0 | 9 | |
| 4200 | North West | 01.05.18 | 7 | 0 | 10 | |
| 5800 | North West | 01.05.18 | 7 | 0 | 8 | |
| 8100 | West Midlands | 01.05.18 | 7 | 0 | 9 | |
| 10900 | West Midlands | 01.05.18 | 7 | 4 | 5 | |
| 6900 | North West | 01.05.18 | 8 | 0 | 10 | |
| 5700 | East Midlands | 08.05.18 | 8 | 1 | 6 | |
| 11100 | West Midlands | 08.05.18 | 8 | 0 | 9 | |
| 3200 | Yorkshire | 09.05.18 | 8 | 0 | 9 | |
| 3500 | Yorkshire | 09.05.18 | 8 | 0 | 9 | |
| 5900 | East Midlands | 09.05.18 | 8 | 0 | 10 | |
| 6300 | West Midlands | 09.05.18 | 8 | 0 | 10 | |
| 700 | North West | 10.05.18 | 8 | 0 | 9 | |
| 1800 | East Midlands | 16.05.18 | 9 | 0 | 10 | |
| 2700 | North West | 16.05.18 | 9 | 0 | 8 | |
| 3900 | North West | 16.05.18 | 9 | 0 | 10 | |
| 1700 | North West | 23.05.18 | 9 | 0 | 10 | |
| 4600 | West Midlands | 23.05.18 | 10 | 2 | 6 | |
| 6700 | West Midlands | 24.05.18 | 10 | 0 | 9 | |
| 5000 | North West | 23.05.18 | 11 | 0 | 8 | 1 |

Selection of schools for observations

Schools were purposely selected for observations to ensure that they were representative of the geographic regions, which type of training they attended, and the proportion of nominated children eligible for FSM.

| Training type | FSM status of nominated children | School location | Previous iPad use and TA confidence Low | Previous iPad use and TA confidence Moderate | Previous iPad use and TA confidence High |
|--------------------------------|----------------------------------|-----------------|---|--|--|
| Attended training day 21/39 | Below Median FSM 9/21 | East Midlands | | · C9 | |
| | | West Midlands | · C1 · C6 | | · C4 · C3 |
| | | North West | | · C2 | · B5 · A5 · A8 |
| | | Yorkshire | | | |
| | Above median FSM 12/21 | East Midlands | · A2 | | |
| | | West Midlands | | | · C8 |
| | | North West | · B9 · B4 | · B10 · A7 · A4 | · B3 · C5 |
| | | Yorkshire | | | |
| Online Only 9/16 | Below Median FSM 6/9 | East Midlands | | | |
| | | West Midlands | | · B8 | |
| | | North West | · B6 | | · B7 · B2 |
| | | Yorkshire | · A1 | | · A10 |
| | Above median FSM 3/9 | East Midlands | | | |
| | | West Midlands | · C7 | | |
| | | North West | | · B1 | · A3 |
| | | Yorkshire | · A6 | | |

Letters in the cells represent the observer, and the number, the order in which the observation was conducted by each observer.

- A. Observations and interview conducted by SB
- B. Observations and interview conducted by DSE
- C. Observations and interview conducted by DE

Appendix 6: Multiple regression analyses

Multiple regressions with post-test as outcome; pre-test entered as first step, number of sessions offered by the school as second step, and number of quizzes passed by the child as the third step

Descriptive Statistics

| | Mean | Std. Deviation | N |
|---|--------|----------------|-----|
| Post test score | 8.5469 | 4.32300 | 501 |
| Pre Total Raw Score | 13.45 | 4.251 | 501 |
| Number of onebillion sessions school ran (session logs) | 43.41 | 6.259 | 501 |
| Total Number of Certificates achieved Max 28 (session logs) | 19.52 | 4.718 | 501 |

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | R Square Change | Change Statistics | | | Sig. F Change |
|-------|-------------------|----------|-------------------|----------------------------|-----------------|-------------------|-----|-----|---------------|
| | | | | | | F Change | df1 | df2 | |
| 1 | .552 ^a | .304 | .303 | 3.60906 | .304 | 218.383 | 1 | 499 | .000 |
| 2 | .555 ^b | .308 | .306 | 3.60256 | .004 | 2.803 | 1 | 498 | .095 |
| 3 | .585 ^c | .342 | .338 | 3.51620 | .034 | 25.763 | 1 | 497 | .000 |

a. Predictors: (Constant), Pre Total Raw Score

b. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs)

c. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs), Total Number of Certificates achieved Max 28 (session logs)

ANOVA^a

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|-----|-------------|---------|-------------------|
| 1 | Regression | 2844.509 | 1 | 2844.509 | 218.383 | .000 ^b |
| | Residual | 6499.639 | 499 | 13.025 | | |
| | Total | 9344.148 | 500 | | | |
| 2 | Regression | 2880.887 | 2 | 1440.444 | 110.987 | .000 ^c |
| | Residual | 6463.261 | 498 | 12.978 | | |
| | Total | 9344.148 | 500 | | | |
| 3 | Regression | 3199.416 | 3 | 1066.472 | 86.259 | .000 ^d |
| | Residual | 6144.732 | 497 | 12.364 | | |
| | Total | 9344.148 | 500 | | | |

a. Dependent Variable: Post-test score

b. Predictors: (Constant), Pre Total Raw Score

c. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs)

d. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs),
Total Number of Certificates achieved Max 28 (session logs)

Coefficients^a

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|-------|---|-----------------------------|------------|---------------------------|--------|------|---------------------------------|-------------|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | .999 | .536 | | 1.865 | .063 | -.053 | 2.051 |
| | Pre Total Raw Score | .561 | .038 | .552 | 14.778 | .000 | .486 | .636 |
| 2 | (Constant) | -.872 | 1.239 | | -.704 | .482 | -3.306 | 1.562 |
| | Pre Total Raw Score | .561 | .038 | .552 | 14.806 | .000 | .487 | .636 |
| | Number of onebillion sessions school ran (session logs) | .043 | .026 | .062 | 1.674 | .095 | -.007 | .094 |
| 3 | (Constant) | -.923 | 1.209 | | -.764 | .445 | -3.299 | 1.452 |
| | Pre Total Raw Score | .494 | .039 | .486 | 12.580 | .000 | .417 | .571 |

| | | | | | | | |
|--|-------|------|-------|-------|------|-------|------|
| Number of onebillion sessions school ran (session logs) | -.025 | .029 | -.037 | -.893 | .372 | -.082 | .031 |
| Total Number of Certificates achieved Max 28 (session logs) | .201 | .040 | .220 | 5.076 | .000 | .123 | .279 |

a. Dependent Variable: Post-test score

Multiple regression with post-test as outcome; pre-test entered as first step and TAs' perception of their role entered as second step

Descriptive Statistics

| | Mean | Std. Deviation | N |
|--|--------|----------------|-----|
| Post-test sores | 8.3527 | 4.43005 | 465 |
| Pre Total Raw Score | 13.49 | 4.223 | 465 |
| Number of onebillion sessions school ran (session logs) | 41.92 | 8.029 | 465 |
| How did TA perceive their role (TA Questionnaire) categorised by evaluation team | 2.00 | .545 | 465 |

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | R Square Change | Change Statistics | | | Sig. F Change |
|-------|-------------------|----------|-------------------|----------------------------|-----------------|-------------------|-----|-----|---------------|
| | | | | | | F Change | df1 | df2 | |
| 1 | .551 ^a | .304 | .302 | 3.70002 | .304 | 202.160 | 1 | 463 | .000 |
| 2 | .556 ^b | .310 | .307 | 3.68870 | .006 | 3.846 | 1 | 462 | .050 |
| 3 | .566 ^c | .321 | .316 | 3.66344 | .011 | 7.394 | 1 | 461 | .007 |

a. Predictors: (Constant), Pre Total Raw Score

b. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs)

c. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs),
How did TA perceive their role (TA Questionnaire) categorised by evaluation team

ANOVA^a

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|-----|-------------|---------|-------------------|
| 1 | Regression | 2767.606 | 1 | 2767.606 | 202.160 | .000 ^b |
| | Residual | 6338.553 | 463 | 13.690 | | |
| | Total | 9106.159 | 464 | | | |
| 2 | Regression | 2819.940 | 2 | 1409.970 | 103.624 | .000 ^c |
| | Residual | 6286.219 | 462 | 13.607 | | |
| | Total | 9106.159 | 464 | | | |
| 3 | Regression | 2919.175 | 3 | 973.058 | 72.504 | .000 ^d |
| | Residual | 6186.985 | 461 | 13.421 | | |
| | Total | 9106.159 | 464 | | | |

a. Dependent Variable: Post-test scores

b. Predictors: (Constant), Pre Total Raw Score

c. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs)

d. Predictors: (Constant), Pre Total Raw Score, Number of onebillion sessions school ran (session logs), How did TA perceive their role (TA Questionnaire) categorised by evaluation team

Coefficients^a

| Model | | Unstandardized Coefficients | | Standardized Coefficients Beta | t | Sig. | 95.0% Confidence Interval for B | |
|-------|---|-----------------------------|------------|-----------------------------------|--------|------|---------------------------------|-------------|
| | | B | Std. Error | | | | Lower Bound | Upper Bound |
| 1 | (Constant) | .553 | .575 | | .962 | .337 | -.577 | 1.682 |
| | Pre Total Raw Score | .578 | .041 | .551 | 14.218 | .000 | .498 | .658 |
| 2 | (Constant) | -1.141 | 1.037 | | -1.101 | .272 | -3.178 | .896 |
| | Pre Total Raw Score | .574 | .041 | .547 | 14.124 | .000 | .494 | .653 |
| | Number of onebillion sessions school ran (session logs) | .042 | .021 | .076 | 1.961 | .050 | .000 | .084 |
| 3 | (Constant) | -3.078 | 1.252 | | -2.459 | .014 | -5.538 | -.618 |
| | Pre Total Raw Score | .589 | .041 | .562 | 14.464 | .000 | .509 | .670 |

| | | | | | | | |
|--|------|------|------|-------|------|------|-------|
| Number of onebillion sessions school ran (session logs) | .042 | .021 | .076 | 1.986 | .048 | .000 | .084 |
| How did TA perceive their role (TA Questionnaire) categorised by evaluation team | .857 | .315 | .105 | 2.719 | .007 | .238 | 1.476 |

a. Dependent Variable: Post-test scores

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

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk


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

This document is available for download at <https://educationendowmentfoundation.org.uk>



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

<https://educationendowmentfoundation.org.uk>

 @EducEndowFoundn

 Facebook.com/EducEndowFoundn