Department of Psychology, University of Nottingham
Department of Education, University of Oxford
Prof Terezinha Nunes

Statistical Analysis Plan for Evaluation of the onebillion maths apps 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 May 2018

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'. 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. In this trial, they will be used as a sequence: all children will start with the 3-5 app and move on to the 4-6 app.

Although pupils work individually and progress at their own pace, in this trial 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, who can be a teacher or a teaching assistant (TA); for brevity, the member of staff will be referred to as TA. Children are encouraged to work through the activities in each topic 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.

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. Some TAs received face-to-face training before starting the intervention; those who could not attend this training were encouraged to do the training online. All TAs received a handbook which explained their role. However, the TAs' role was not listed as a fidelity factor by the intervention team in their logic model. The only fidelity factor indicated by the intervention team was time on the apps. In order to collect information on time on the apps, TAs were requested to use a specifically designed register and a chart of received certificates.

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 were asked by the intervention team to 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 question:

- 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?

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 spoken English. 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 (who were supply teachers) 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, at both pre- and post-test.

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 and given £1000.

TAs in schools assigned to the intervention group were invited to participate in the training for implementation of the intervention. The training covered: 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 in 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 app.

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.

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 the intervention group and the schools with the highest random numbers were allocated to the control group. 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 points above boys in PTM5 and 0.3 points 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 at 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. GL assessments do not describe in the technical information the correlation between PTM 5 and PTM 6. 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 assessments[1]); (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, in agreement with the project teams, 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. Recruitment could continue beyond this number, in case schools withdrew.

At the deadline for recruitment, 118 schools met these criteria but 5 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

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

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 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.

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 will be used to replicate the results) 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 the 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.

Preliminary analyses will be carried out to describe the shape of the distribution of scores at pre- and post-test. Preliminary analyses will also be used to compare schools and pupils at pre-test. Any significant differences at pre-test will be discussed when results are interpreted.

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. Hedges’ 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). Further measures of compliance will investigate whether the TAs followed the guidelines in the Implementation Handbook (i.e. played the role they were expected to play). Although this is not mentioned by the intervention team as a fidelity factor, the evaluation team considered that adherence to these instructions could be a moderator of the impact.

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 using multilevel models (as for the analysis with all participants), comparing the intervention and the control groups, and including an interaction term between FSM status and impact of the intervention. A subgroup analysis with results considered separately for pupils eligible for FSM and not eligible for FSM will be conducted, 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.

1. **Stopping point**: The definition of stopping point is the highest number of topics 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.
2. **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.
3. **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 from 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 dosage as a mediator of the impact, in view of the significance of this factor in the intervention team's logic model. 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. If number of sessions attended is a significant factor, it will be investigated whether different effect sizes are obtained with different dosages defined by the intervention team.

Finally, it is also possible that children are present at the session, but spend their time doing other activities. The stopping point and the number of quizzes passed can provide information on how far the children progressed on the apps and will be analysed to test whether these measures of engagement can be seen as mediators of impact.

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.

The first observations of sessions revealed very large differences across TAs in their approach to offering pedagogical support, if any was offered. The evaluation team, in discussion with the EEF, decided to include a larger sample of observation sessions in order to obtain a more nuanced description of this variation.

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 out 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 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 with 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 interactions 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.

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).

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.

List of Appendices

Appendix 1: Local authorities in the target regions for recruitment and selected Local Authories within those regions

Appendix 2: Power calculation for MDES for the recruited sample after withdrawal of two schools

Appendix 3: Power calculation for the subgroup analysis including only pupils eligible for FSM in the recruited sample

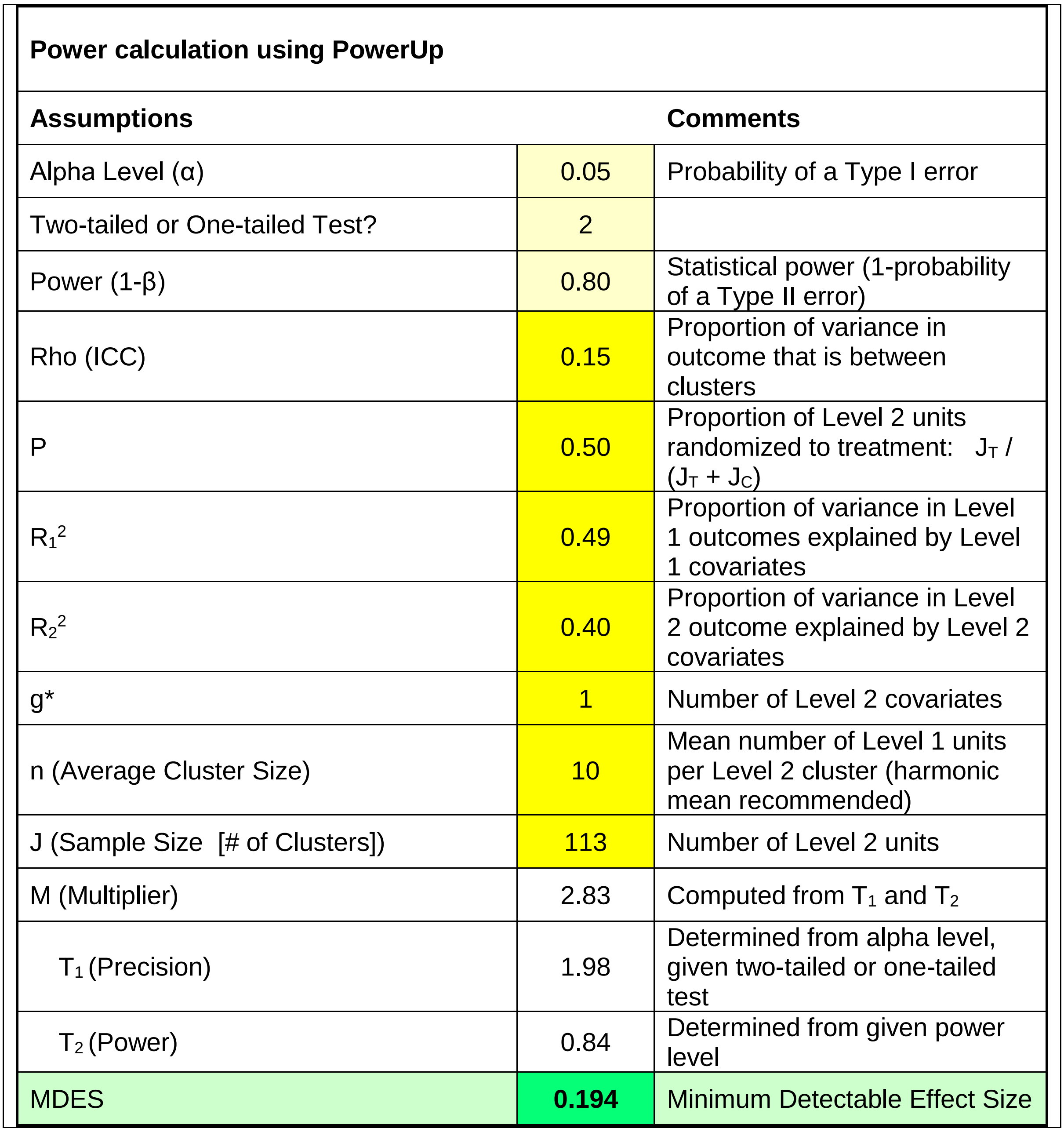
Appendix 4: MoU and parent information letters

Appendix 5: Data sharing agreement between the Oxford and Nottingham teams

Appendix 1: Local authorities in the target regions for recruitment and selected Local Authories 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
4. Yorkshire West & South	Bradford, Leeds, Wakefield, Barnsley, Calderdale, Kirklees, Doncaster, Rotherham, Sheffield

Appendix 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-β)	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_1^2	0.49	Proportion of variance in Level 1 outcomes explained by Level 1 covariates
R_2^2	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 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-β)	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_1^2	0.49	Proportion of variance in Level 1 outcomes explained by Level 1 covariates
R_2^2	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