Catch Up® Numeracy Support: An Evaluation Protocol for an Effectiveness Trial

Professor Jeremy Hodgen, Professor Shaaron Ainsworth and Dr Michael Adkins

School of Education, University of Nottingham

Evaluation Summary	
Age range	7 to 9 year olds studying Key Stage 2
Number of Pupils	1800 (6 per TA and 12 per school)
Number of TAs	300 (2 per school)
Number of Schools	150 (in 3 counties) from state sector
Design	Randomised Controlled Trial, two-arm design: intervention vs. optimal business as usual control
Primary Outcome	Progress in Maths test measured by Progress Test in Mathematics (PTM)
Trial Registration	Registered with ISRCTN

Table 1: Catch Up Numeracy Support effectiveness evaluation design parameters

Abstract

Catch Up[1] Numeracy is a research-based intervention targeted at children who are low attaining in mathematics (Holmes and Dowker 2013). The intervention is guided by a componential approach to numeracy and is designed to address individual children's particular difficulties and weaknesses. Catch Up is aimed at children aged 6-14 who are struggling with numeracy, targeting a relatively large group: the lowest attaining 15-20% of children who are "at risk of less severe but nevertheless persistent numeracy difficulties" (Holmes and Dowker 2013, p.253). This evaluation protocol outlines the evaluation of a two-arm effectiveness trial of primary school pupils (Years 3, 4 and 5) from 150 schools based in Northumberland and Durham in the North East, Yorkshire, and Peterborough in Cambridgeshire. This trial follows an earlier Education Endowment Foundation study that showed some positive effects for Catch Up Numeracy intervention (Rutt et al, 2014). This protocol outlines the final evaluation of the project, including its rationale, objectives, and a three element evaluation plan. The evaluation will incorporate an impact evaluation, process evaluation and cost evaluation. We discuss the data collection and analysis for all three components, a plan for data release to promote independent scrutiny and overall study limitations. Overall, this evaluation is designed to assess the Catch Up Numeracy model for improving the numeracy ability of children in the bottom 5 to 15% based on a review of their progress from KS1 results and internal qualitative assessments.

[1] Catch Up[1] ® Numeracy is a registered trademark of the Catch Up charity which works to address literacy and numeracy difficulties that contribute to underachievement. See: http://www.catchup.org/

INTRODUCTION 3
BACKGROUND 4
PROJECT DESCRIPTION 4
Intervention 4
EXISTING EVIDENCE FOR THE EFFICACY OF CATCH UP NUMERACY 6
RESEARCH QUESTIONS 7
RESEARCH DESIGN 8
EXPERIMENTAL DESIGN 8
UNITS OF RANDOMISATION 9
PARTICIPANTS 10
SAMPLING FRAME 10
INCENTIVES 11
IMPACT EVALUATION 11
Outcome Measures 12
Sample size calculations 12
Analysis 14
Software 15
PROCESS EVALUATION 16
COST EVALUATION 19
DATA MANAGEMENT AND ANALYSIS 20
Backup and security of data 20
Data protection, rights and access 21
Transparency, preservation, sharing and licensing 21
Adherence 22
ETHICAL ISSUES 22
PERSONNEL 22
RISK ANALYSIS 23
EVALUATION TIMELINE 25
STUDY PROTOCOL LIMITATIONS AND FUTURE AREAS OF RESEARCH 26
REFERENCES 27

Introduction

The Education Endowment Foundation has funded the University of Nottingham to evaluate the Catch Up teaching assistant delivered one-to-one intervention for learners struggling with numeracy skills entitled “Catch Up Numeracy”. The intervention will be evaluated on the basis of a two-arm randomised controlled trial testing the evaluation against a ‘optimal business as usual’ equivalent. The trial will be flexibly delivered in schools for up to three academic terms starting in September 2016, with school recruitment from January, school randomisation in late spring and teaching assistant training in June. Post-intervention testing will take place late in the Autumn term of 2017, with the final publication of the results in the spring term of 2018.

The current trial builds on an earlier Education Endowment Foundation funded trial which showed a positive and statistically significant effect for the intervention when compared to a ‘pure’ control (Rutt et al, 2014). However, a group receiving matched time support delivered by teaching assistants made a similar gain. In the current trial, randomisation will be at school-level rather than individual-level in order to avoid the potential cross-contamination issues in the previous trial.

Together with a separately published technical appendix and analysis plan, this evaluation protocol sets out the context of the intervention in terms of attainment levels in primary mathematics, the trial objectives, the research design for both the impact evaluation– addressing the experimental design decisions, trial arms, units of randomisation, sample size calculations, outcome measures which test both the improvement in mathematics skills, as well as attitudes and anxieties, and software used – and also the process evaluation. This will outline the assessment of the fidelity of implementation, how the process evaluation will inform any future effectiveness trial, and the investigation of aspects of ‘social validity’ e.g. the acceptability, feasibility and utility of the approach. Lastly, the protocol will address issues of transparency and study ethics, outline an evaluation timetable, set out the responsibilities of the evaluation team members, and highlight any limitations of the study – including areas for future research.

Background

Low attainment in numeracy is a very significant problem in England (Marshall, 2013). Although standards in primary numeracy have risen over the past two decades (Tymms, 2004), overall performance remains below that of the highest achieving jurisdictions and primary performance is not sustained into secondary (Mullis et al, 2012). However, in contrast to this overall trend, the performance of the very lowest attaining students in England appears to have fallen, which has led to a widening of the attainment gap (Brown et al, 2008). Addressing the problem is particularly important because research has shown that early mathematics skills are associated with later educational success (Crawford & Cribb, 2013). The Catch Up: Numeracy intervention is worthy of investigation since by using teaching assistants it offers a potentially cost effective approach to addressing this problem.

Project Description

The aim of the project is to test and evaluate, within a rigorous and high quality research framework, the impact on underperforming Year 4 and Year 5 pupils, of the Catch Up Numeracy intervention on attainment and attitudes to mathematics, when it is delivered for a period of up to 3 terms by trained teaching assistants in up to 75 schools against 'in-house approaches' to supporting similar pupils when delivered by teaching assistants for a similar time, again in up to 75 schools.

Intervention

As per the previous effectiveness trial funded by the Education Endowment Foundation (see Rutt et al, 2014), Catch Up Numeracy is a one-to-one intervention for learners who are struggling with numeracy that aims to improve attainment in numeracy and attitudes to mathematics. It consists of two 15-minute sessions per week which are delivered by teaching assistants. To prepare them for delivering the intervention, teaching assistants are supplied with detailed session plans and receive three half-day training sessions, each involving a follow-up task in school. Each school also appoints a Catch Up Numeracy Coordinator to coordinate Catch Up Numeracy in school and to support the teaching assistants.

The intervention breaks numeracy down into ten components and assesses children's ability on each. By using a checklist approach based on this assessment, it targets subsequent instruction so that the teaching assistant always addresses the child's exact area of weakness

(Dowker & Sigley, 2010). Components include counting procedures, counting principles, derived fact strategies, etc. The approach is based on research indicating that numeracy is not a single 'big' skill, but a compound of several 'little' skills that seem to be quite discrete (Dowker, 2009). Children (and adults) may be very strong in some skills but very weak in others, and indeed some brain-imaging studies suggest that different skills may be handled by different networks in the brain. By recognising and building on this finding, the Catch Up Numeracy intervention enables tutors to diagnose and treat problems precisely and effectively.

Unlike in the previous trial (Rutt et al., 2014), the intervention will be delivered flexibly and children will 'roll off' the programme. Children are assessed on a termly basis as to their eligibility for the intervention. Children who are judged to have reached an age appropriate level of numeracy then no longer receive the intervention. This is because this is how Catch Up Numeracy is usually delivered.

Additional support is provided through the Catch Up Numeracy online platform and teaching assistants may apply for accreditation. Full Catch Up Numeracy training will be provided to train replacements for any teaching assistants who leave before the start of the intervention.

The Catch Up Numeracy intervention is summarised in the logic model in Figure 1.

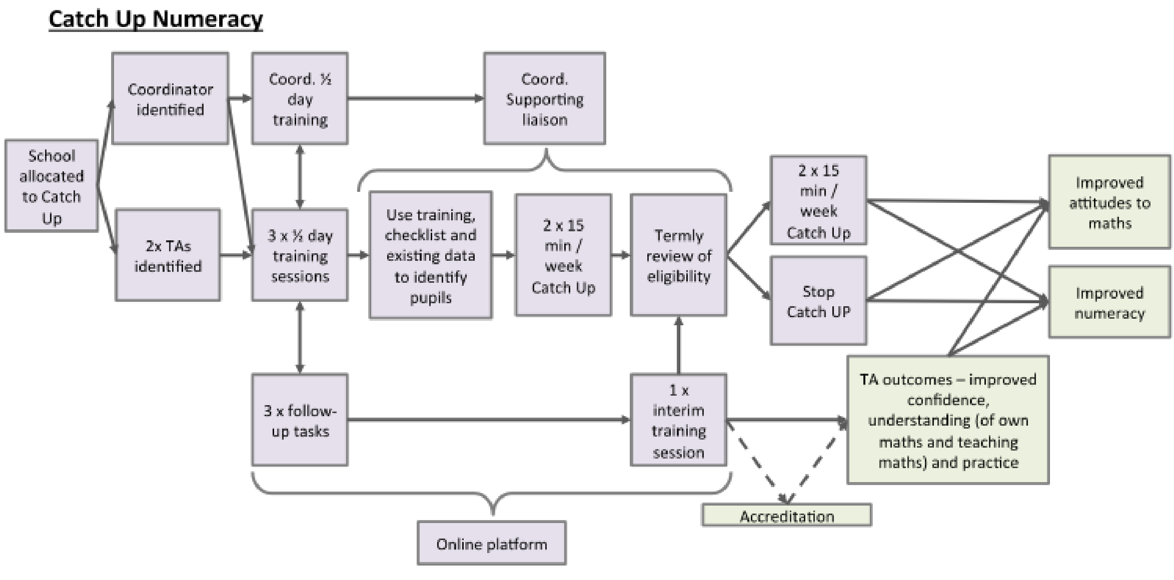

Figure 1: Catch Up Numeracy Logic Model

Existing evidence for the efficacy of Catch Up Numeracy

Evidence of the efficacy of Catch Up Numeracy comes from one independent evaluation funded by the Education Endowment Foundation and conducted by the National Foundation for Educational Research (Rutt et al., 2014) and several studies conducted by the developers. This existing evidence relates to a 'fixed-time' implementation of Catch Up Numeracy without a progress review and the opportunity for 'roll off' for individual pupils as in the current trial.

Three hundred and thirty-six children in Years 2 to 6 from 54 schools participated in a three-arm randomised controlled trial, each randomly assigned within their school to one of three groups: The Catch Up Numeracy intervention, a 'matched-time' group in which they received the same amount of one-to-one maths instruction with a TA but not using Catch Up, and a 'no-intervention' group where children received no additional TA support beyond normal classroom instruction (Rutt et al, 2014). All three groups sat the Basic Numeracy Screening Test (Gillham and Hesse, 2001) before and after the intervention. After attrition, the analysis involved 108, 102 and 108 children for each of the Catch Up, matched time and no-intervention groups, respectively. Using an intention to treat analysis, the independent evaluators found that Catch Up and the matched-time group made greater progress when compared to the business-as usual control: Effect sizes (Hedges *g*) of +0.21 (CI 0.42-0.01) for Catch Up and +0.27 (CI 0.49-0.06) for the matched-time group. These differences were statistically significant, but there was no significant difference between the Catch Up and the matched-time groups. However, there may have been some cross-contamination between these groups, because delivery was by TAs within the same schools.

An earlier study conducted by the developers found positive effects for the Catch Up Numeracy intervention. In quasi-experiment involving 440 children assigned to either Catch Up intervention (N=348), matched-time one-to-one TA support (N=50), or no-intervention (N=42) groups, Dowker and Holmes (2013) found a statistically significant difference in gain between Catch Up and both the matched-time and no-intervention groups with effect sizes (Cohen's *d*) of +0.47 and +0.55, respectively. In an earlier report of this study (Dowker and

Sigley, 2010) , the effects were compared using mean ratio gains.[2] The mean ratio gains were 2.2, 1.47 and 1.25 for the Catch Up, matched-time and no-intervention groups.

Research Questions

The evaluation will address this primary research question:

1. Does Catch Up Numeracy have significant effect on children's attainment in mathematics when compared to an optimal business as usual active control?

In addition, the evaluation will address the following secondary research questions:

2. Does Catch Up Numeracy have significant effect on children's attitudes towards mathematics when compared to an optimal business as usual active control?
3. Are the effects on attainment and attitudes different for children eligible for free school meals?
4. Are the effects on attainment and attitudes different for girls and boys?
5. To what extent are any effects on attainment and attitudes mediated by the treatment time?

In the process evaluation, we will address the following research questions:

6. To what extent do schools, coordinators and teaching assistants perceive the Catch Up Numeracy professional development for teaching assistants and coordinators to be effective?
7. To what extent do the Catch Up Numeracy intervention schools, teaching assistants and coordinators adhere to the guidance and materials?
    - How variable is the quality of implementation in the intervention schools?
    - To what extent are the Catch Up Numeracy resources sufficient, appropriate and easy to access, and is the assessment guidance sufficiently flexible and child-friendly?
    - What school and contextual factors afford or constrain the quality of implementation?
    - In what ways do schools manage and support the teaching assistants?

[2] The ratio gain is calculated by the months gained in Mathematics Age divided by duration of intervention in months. The mean ratio gains reported here approximate to Cohen's *d* of around 0.3 and 0.4 for Catch Up compared to the matched-time and non-intervention groups.

- To what extent, and how, do schools enable dialogue to take place between the teaching assistants and the relevant class teachers?

8. Are children assessed for eligibility on a termly basis in both intervention and control schools and do children judged to have reached an age appropriate level of numeracy 'roll off' the intervention?
9. What does usual practice in control schools look like?
    - How, and to what extent, do the active control schools implement the Education Endowment Foundation's 'Making best use of teaching assistants' guidance?
    - To what extent does the half-day planning session enable schools to implement the guidance?
10. To what extent does Catch Up Numeracy have an effect on teaching assistants' attitudes towards mathematics and mathematics teaching when compared to an optimal business as usual active control?

Research Design

The intervention has been designed to be a whole school randomised controlled trial for approximately 1800 pupils in 150 schools. Our aim is to recruit 12 children per school, 2 TAs, and for the intervention arm an additional coordinator with school recruitment focused on three large counties – Yorkshire, Northumberland and Cambridgeshire.

Experimental Design

This experiment will utilise a two armed trial: Catch Up intervention and an optimal business as usual active control. All schools enrolled in the trial will identify pupil participants, with these pupils sitting a pre-test prior to school randomisation. Before randomisation, schools will assign at least two teaching assistants who will take part in the trial whether the school is allocated to the intervention or the control and identify a member of staff who will act as a coordinator if the school should be allocated to the intervention. In the optimal business as usual active control arm, schools will appoint at least two teaching assistants and will be provided with EEF guidance (Sharples et al., 2015) and will be advised to organise an information session on this guidance. All pupils will begin the one-to-one sessions at the start of the school year. It will be recommended that all schools review progress each term, rolling off approximately one third of the participating pupils at each time point – end of autumn, spring and summer term – in both intervention and control arms. However, it may not work out exactly and there will be a process for recording the dosage. All participants will then have a delayed post-test towards the end of the autumn term of 2017. with a delay of between 3 and 9 months depending on when children are assessed as having reached the required level in numeracy.

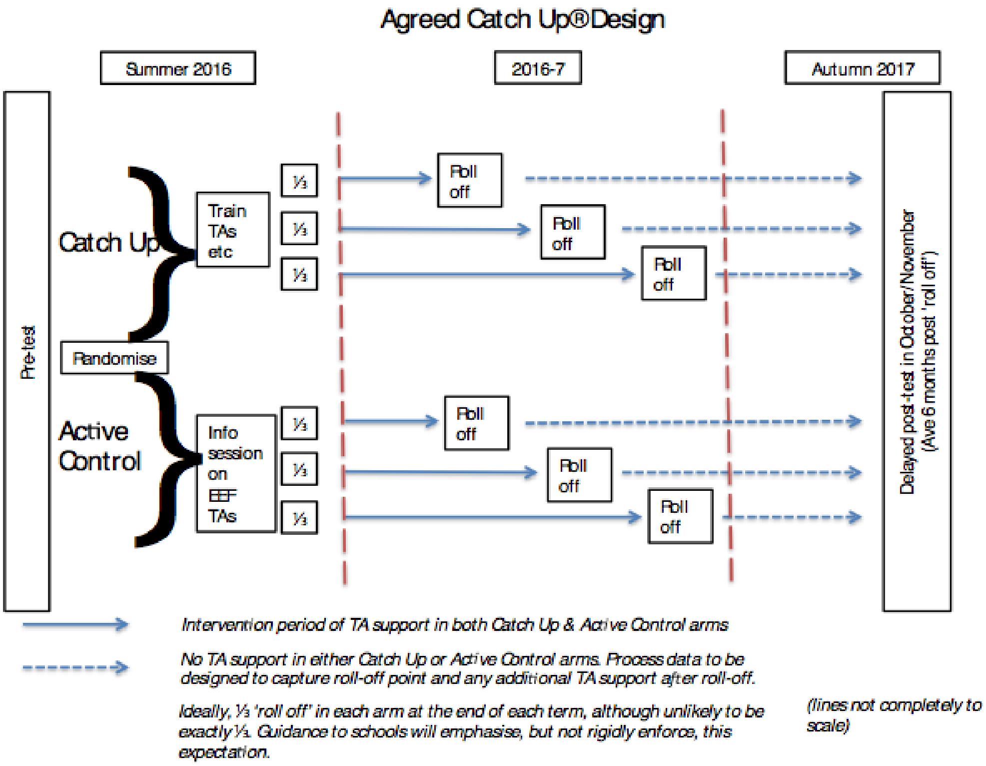

Figure 2: Experimental Design Model

Units of Randomisation

Given that in the original EEF funded evaluation of Catch Up Numeracy (Rutt et al., 2014) there was an identified potential contamination effect at the level of pupils and TAs as participants in all three arms were present in the same schools, we have chosen to ameliorate this by randomising at school level and stratifying at the regional level by targeting 100 schools in Yorkshire, 25 in the North East and 25 in Peterborough/Cambridgeshire. This has some additional advantages in making delivery easier for schools and the developer as well as being more 'real world' for schools.

Participants

Pupils from participating schools will be identified by their teaching staff as eligible prior to randomisation based on a review of their progress from KS1 results and internal qualitative assessments. Catch Up specifically targets those pupils struggling in numeracy who are in the bottom 5-20% of attainment in mathematics as most likely to benefit from the additional one-to-one intensive teaching that the intervention provides.

Eligibility Criteria

All state schools are eligible as long as they have not already purchased Catch Up Numeracy and can provide a minimum of six year 4s and six year 5s who are eligible for the intervention (or alternatively four eligible pupils in each of year 3, 4 and 5). Junior schools could be included if they agreed to get the pre-test data from the eligible year 3s (current year 2s) from their partner infant schools.

In addition, in order to be entered into the randomisation the schools will have to have provided:

- Signed Memorandum of Understanding
- Confirmation that consent forms have been sent out and any opt-outs
- Provision of pupil data for those identified as eligible (UPNs etc.)
- Pre-test data for all eligible pupils

Sampling Frame

The intervention targets schools in three key regions of the country - Yorkshire and Humberside, the North East, and Peterborough in the East of England. The priority is to maximise recruitment of schools and these three regions have relatively low uptake of Catch Up interventions. In addition, these areas have a mix of rural and urban areas to ensure a balanced population of schools.

In terms of average Key Stage 1 Mathematics mathematics achievement, in 2015, 89% of eligible pupils achieved the age appropriate level (level 2) within Yorkshire and Humberside. For the East of England this was 91% and for the North East this was 90%. Nationally, for all schools the average for those achieving age appropriate level was 93%. However, there is variation within these large geographic regions, and there is a considerable difference when the averages for all pupils are compared with disadvantaged pupil (Free School Meal recipient) results. Figure 3 highlights the within region variation for all children and for disadvantaged pupils only. For disadvantaged pupils, the range of results are 80-88%. For all pupils, the range of results are 89-94%.

Figure 3: Spatial analysis of Key Stage 1 Mathematics Level 2 achievement by area. The upper row highlights the average achievement for all pupils and the lower row displays results for free school meal recipients only. Peterborough is situated in the north-west of the third column maps, with the remainder of the region provided for comparative purposes.

Incentives

All schools will be provided with a payment of £450 once they have completed the post-tests. Intervention schools will be provided with Catch Up Numeracy training, but in order to avoid contamination of the sample for long-term analysis, all control schools will be provided with Catch Up Literacy training in the form of a wait-list in September 2017. They will be requested not to use the training on children identified as eligible for Catch Up Numeracy.

Impact Evaluation

The impact evaluation investigates the quantitative evidence for whether there is a consistent positive difference in the post-test score between participants who have received the intervention as part of the treatment group and those that were recipients of the control condition, conditional on their pre-treatment test score. This subsection will discuss outcome measures, sample size calculations, the main and secondary outcome analyses along with additional sub-group investigations. This section is supplemented by a technical appendix that contains details on model specifications, estimation procedures and software to be used.

Outcome Measures

Our primary outcome measure is pupil test scores using the Progress Test in Mathematics published by GL Assessment which addresses number, shape, data handling and algebra. For pre-testing in year 3, 4 and 5 we will use PTM 7, 8 and 9 paper and pencil versions respectively. For post-testing in year 3, 4 and 5 we will use PTM 8, 9 and 10 paper and pencil versions respectively. See GL assessment's (2015) PTM technical information for further information including reliability testing. Pre-tests will be administered by schools prior to randomisation. Post-tests will be administered by independent administrators employed by the Univeristy of Nottingham, who are as far as possible blind to whether schools are in the intervention or active control groups. Our secondary outcome measure are pupil responses to a mathematics attitudes and anxiety questionnaire.

Pupil attitudes will be measured using four single-scale items. These are quick and cost effective to administer and similar attitude scales have been shown to be valid and reliable (e.g., Núñez-Peña et al, 2014). We will measure of self-rating and liking for mathematics. In order to assess the extent to which any effects are specific to mathematics rather than relating to more general attitudes to learning, we will measure pupil attitudes to reading using parallel items.

Sample size calculations

Given the results of the previous trial, our aim has been to maximise the possibility of detecting a small effect between the active control and the intervention arm. We used Raudenbush et al.'s (2011) Optimal Design software to run a series of statistical power calculations on the basis of recruiting 2 TAs and 150 schools for a 3-level cluster randomised trial with the intervention at level 3 (i.e. the school level). The structure of the intervention is made up of 3 levels – pupils are clustered in TAs who are further clustered in schools. However, we will be flexible and should there be little discernible variation at the TA level we will fit a more traditional two level model where pupils are clustered in schools.

We varied the number of pupils per TA (2,4 and 6) to provide us with a realistic estimate of power given the expected pool of participants in schools and fixed the remaining parameters. These were as follows: α=0.05 (which refers to the probability of rejecting the hypothesis tested when it is true – 5%), and intra-cluster correlation for level 2 (TAs) =0.05 and for level 3 (schools) =0.10 (which refers to the variance between participants with the same TA and for those in the same school). We also included an additional pre-test covariate as a school level aggregate with the assumption that the post and pre-test have a correlation of .80, setting the level 3 variance explained at $.80^2$=0.64. This has the effect of reducing the overall variance and boosting the expected statistical power of the study. The results of the calculations can be seen in figure 4.

Note that for effect sizes, being able to detect smaller effect sizes results in a more reliable statistical investigation. At 80% power, where under repeated testing we would expect the interval to contain the true effect 80% of the time, the minimum detectable effect size (MDES) is 0.172 with 6 participants per TA. If the evaluation does not recruit enough participants per TA this has the impact of decreasing our ability to detect small effect sizes (with the MDES increasing to 0.19 with 4 participants per TA and increases further to 0.25 with just 2 participants per TA. Given the small difference between the active control and the intervention arm in the previous trial equipping this trial with the power to detect smaller effect sizes will be critical and we therefore favour recruiting 6 pupils per TA where possible.

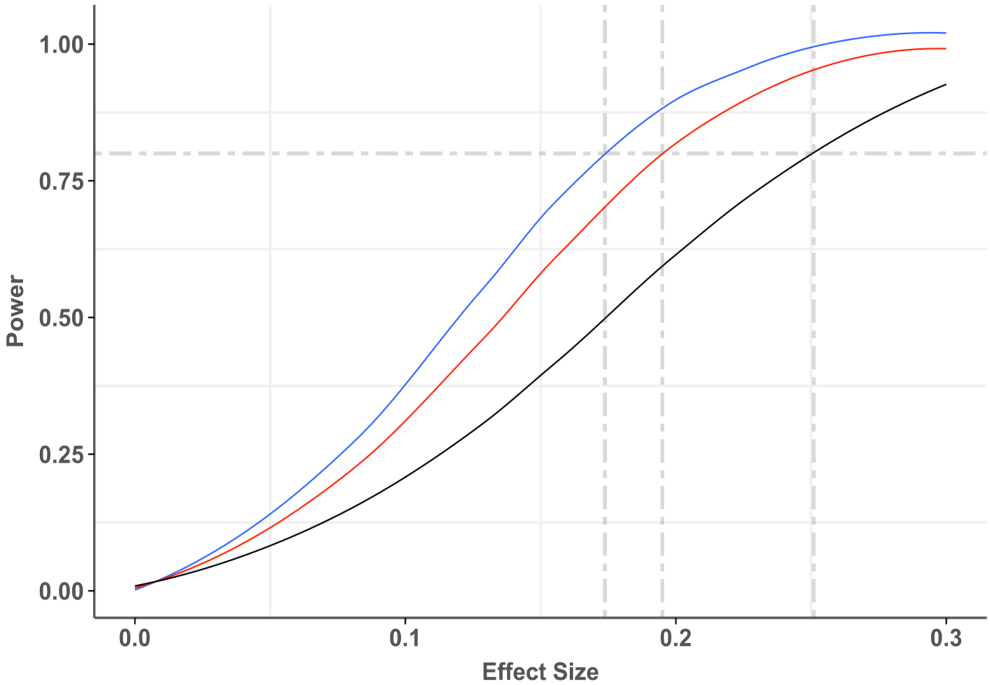

Figure 4. Three power curves highlighting the change in the minimum detectable effect size when the number of pupils per TA is varied (Black=2, Red=4, Blue=6)

Analysis

Our analysis will proceed on the basis of descriptives to explore and visualise patterns within the data before building complexity through the use of multilevel modelling to adjust our inferences to take account of the clustering at the TA and school level. There are two outcomes and both will be subject to further subgroup analyses detailed below and separately in the technical appendix which addresses a number of underlying statistical decisions.

Main outcome As discussed above, Pupil performance in maths will be measured using the PTM post-test, with the pre-test carried out prior to randomisation. Results will be modelled on the basis of intention to treat (ITT). Our analysis will proceed on the basis of starting with simple single-level models incorporating the treatment condition and gradually build complexity by incorporating the pre-test, and adding the varying intercepts for TAs and schools While fitting 2-level model can possibly be justified in terms of a more parsimonious model, ignoring the variation at the TA level would potentially have the effect of inflating the standard errors of coefficients and would, in our opinion, be an error despite the potential to reduce the complexity of the analysis. As mentioned above, if the variation at the TA level is small, then we will consider reverting to a more traditional 2-level model of pupils nested within schools.

Subgroup analysis Additional models of greater complexity will be fitted which will include sex of participant and free school meal entitlement as well as interactions between the two original data level variables of treatment and pre-test and the additional sub-group variables. Finally, we will add appropriate group-level predictors such as TA maths level for level 2 and school averages for KS1 and KS2 maths at level three.

Secondary outcome We will model this data use a multivariate multilevel model (which allows for multiple outcome variables to be tested simultaneously) using the primary outcome and the Mathematics Attitudes and Anxiety Questionnaire scale as our second dependent variable. In this model, there are four levels of nesting. Responses to the multiple outcomes are set at the first level providing the structure of the multivariate model, with level two being clusters of individuals, level three being clusters of TAs and level four being clusters of schools. In a nutshell, responses are clustered in individuals, who are clustered in TAs, who are themselves clustered in schools[3].

Despite the increased complexity of the approach, it offers four significant advantages. Importantly, this model allows for modelling correlations between dependent variables; the standard errors of specific effects tend to be smaller; it allows for the direct comparison of testing effects on the dependent variables; and helps to avoid the need for multiple comparisons adjustments such as the Bonferroni correction (Snijders and Bosker, 2012, p.283)**.** Significantly, the second and third advantage will potentially allow for stronger conclusions to be drawn, and additionally the third advantage will provide us with the opportunity to test the relationship between the attitudes scale and the PTM post-test.

Software

All final models will be estimated via Bayesian inference. While Bayesian statistics is far too large a literature to describe here, it is based on personal rather than long-run interpretations of probability. Problems are divided into immediately available quantities and those needing to be described probabilistically. We describe our prior information via a full probability distribution and then sequentially update this knowledge with new data describing the

[3] Detailed model specifications will be outlined in the technical appendix

resulting posterior quantities via quantiles or intervals rather than point estimates (see Browne, 2015; Gill, 2015). Using Bayesian inference provides several advantages for this evaluation – parameters are treated as random– an assumption far more suitable to social science applications where we would expect treatment effects not to be permanently fixed; more interpretable parameter estimates along with clear statements of uncertainty; more accurate estimates of variance parameters – particularly when the numbers of groups are small; missing data is handled seamlessly within the model; and finally results which set up further investigation via the sequential learning mechanism discussed above.

Given the potential contamination between the treatment and control groups in the previous trial, we deliberately use diffuse priors to represent our ignorance on the size and variance of the treatment effects, using diffuse priors (improper uniform priors on the data level effects (βs), inverse gamma priors on the scalar variances and inverse wishart priors on variance matrices). Therefore, we deliberately use less prior information than we have available to allow the new data to dominate the overall results – this is discussed further in technical appendix. Initially we will fit the model classically using lme4 and MLwiN (for consistency with other EEF trials) before we refit the model using linear multilevel/hierarchical regression modelling estimated by Bayesian inference using a combination of the EEFanalytics package, STAN and MLwiN (this will be to check the overall consistency in our inferences and to further test and develop the EEFanalytics package in conjunction with the University of Durham).

Process Evaluation

We will conduct a robust process evaluation that, in addition to focusing on the fidelity of implementation, aims to provide formative feedback in order to inform wider implementation and scale up and to investigate aspects of 'social validity': the acceptability, feasibility and utility of the approach. The research questions for the process evaluation are detailed above (RQs 6, 7 8 & 9). These questions address the professional development delivered by Catch Up (RQ6), the fidelity of implementation of schools in both the intervention and active control arms (RQ7), the affordances and constraints to the implementation of Catch Up Numeracy (RQ7), the 'roll off' process and the dosage received by children in both the intervention and active control arms (RQ8) and the implementation of the Education Endowment Foundation's 'Making best use of teaching assistants' guidance in the active control schools (RQ9).

Given the broad scope of the process evaluation, a variety of approaches will be needed targeted at specific aspects of the process, including existing data / records (where possible), questionnaires, observations and interviews, as follows:

Planned and actual attendance at the Catch Up Numeracy PD sessions To investigate fidelity, we will collect developer records registers of attendance by TAs and school coordinators.

Developer Professional Development Planning The developer's planning and documentation for PD sessions (presentations, handouts and agendas) will be collected in order to provide information on the aims, approach and intended content of the PD sessions. This will be used to inform subsequent collection and analysis.

Observation at PD Sessions We will observe four PD sessions, including a session for the school coordinator, in order to investigate (i) how content is delivered to TAs and how TAs respond, (ii) how content is delivered to Coordinators about the management of the initiative and how Coordinators respond, and (iii) how TAs and Coordinators interact in the PD sessions. This will enable us to provide guidance and formative feedback on the wider implementation and scale up.

Interviews with the Catch Up Numeracy team We will conduct short telephone interviews with at least four members of the development team: Dr Graham Sigley, Dr Ann Dowker, and two Catch Up accredited trainers responsible for delivering the PD sessions. This will provide evidence relating to the planned and actual content of the PD sessions and the intervention more generally.

Surveys of Intervention Participants and Schools We will survey all intervention TAs and coordinators at the end of the PD in order to investigate their perceptions of the PD and their initial views on Catch Up Numeracy. At the end of the intervention in July 2017, we will survey all participants and schools through short focused questionnaires that can produce quantifiable data: a school questionnaire to be completed by head teachers or deputies relating to school leaderships' views on the efficacy and social validity of the intervention, a questionnaire relating to management and support to be completed by the School

Coordinator, and a questionnaire aimed at TAs focused on their views of process. These will provide data about fidelity and the social validity of the intervention in general, the quality of delivery, the extent of supervision and guidance required and participants' views on the utility of the TA sessions. We will ensure that these specially-developed instruments are easy and quick to complete and will either be delivered online administered via the Bristol Online Survey or will be paper-based but, as far as possible, using a machine readable format.

Surveys of Active Control TAs and Schools We will survey TAs and schools in the active control schools in order to provide evidence about use of the Education Endowment Foundation's 'Making best use of teaching assistants' guidance as well as perceptions of the effectiveness of the guidance and will enable comparison with the intervention. These surveys will be conducted as above. We will survey TAs' attitudes to mathematics and mathematics teaching using four single-scale items similar to the measures of pupil attitudes.

In-depth Case Studies of Intervention Schools In order to explore implementation (and particularly constraints/affordances to implementation) and to further investigate aspects of social validity, we will conduct in-depth case studies in five intervention schools, chosen to represent a range of schools and across the three regions. These case studies will be conducted over 4-5 days on two occasions (once during Spring 2017 and once at the end of Summer 2017). The later visit will enable us to collect participants' reflections on the intervention and to review fidelity to the termly assessments and 'roll off' of children. During the data collection with intervention schools, we will pay particular attention to issues relating to the management of the intervention and other barriers identified in the previous evaluation, including whether Catch Up Numeracy resources sufficient, appropriate and easy to access, and whether the initial assessments are sufficiently flexible and child-friendly (Rutt et al., 2014, p.26).

Planned and actual support for children by TAs In order to investigate fidelity to the intervention, we will collect evidence on the planned and actual support received by children from TAs. For both intervention and active control schools, we will pay particular attention to schools' fidelity to the termly assessment process and the extent to which schools do 'roll off' TA support. This will enable us both to evaluate the extent to which the active control does actually offer matched TA support time and to address the secondary research question (RQ5): To what extent are any effects on attainment and attitudes mediated by the treatment time? For the intervention schools, we will make use of TA / school records kept as part of the Catch Up Numeracy intervention. For the active control schools, we will develop a light touch online tool to record the results of this assessment alongside the TA support received by each child.

Light-touch Case Studies of Active Control Schools We will conduct light-touch case studies of five schools in active control group in order to investigate how the control schools use, manage and support TAs, as well as any barriers to using TAs effectively (and the extent to which these barriers are similar or different to those encountered by the Catch Up schools). These case studies will generally require only a 1-day fieldwork visit during which we would observe at least one TA session and conduct short, focused interviews with at least one TA and one member of the school's senior leadership team with responsibility for management of, and support to, TAs.

A detailed outline of the data collection for the process evaluation is set out in Appendix 2.

Cost Evaluation

We will follow the April 2015 EEF Guidance on Cost Evaluation in estimating the costs of the delivery of the intervention. We will collect cost data from the developer via a short interview and a pro-forma. In addition, we will collect data on costs incurred by schools through the process evaluation. Using evidence collected during the process evaluation, we will estimate the costs of supervising and directing the TAs, but will report this separately to the direct, marginal costs alongside any supply cover needed as per the April 2015 EEF Guidance on Cost Evaluation. We will give particular attention to estimating the relative costs of the two approaches to delivery.

Given that the evaluation of the original Catch Up Numeracy identified a lack of senior management support in some schools (Rutt et al., 2014), we will give particular attention to estimating the time and marginal cost of such support. Similarly, the evaluation identified issues with the availability of suitable resources in some schools. It is unclear whether these refer to resources supplied by the school, by Catch Up Numeracy or both. Nevertheless, we will collect data in order to better estimate any pre-requisite equipment and materials that would be required.

Data management and analysis

The evaluation will make use of both primary data collected as part of the impact and process elements of the study and will endeavour to link the pupil level data to individual and school variables from the National Pupil Database. We envisage no problems with gaining access to these data sources as we have already been provided access to the NPD on multiple occasions.

The research will make extensive use of open-source software – particularly R, but also Bayesian Inference software such as MLwiN. We therefore prefer to work with non-proprietary data formats such as tab-delimited data files, as well as ASCII style files for data analysis scripts and creating research outputs. However, some analyses will no doubt be tied to closed-source software. Where appropriate, we will create additional versions of the files in non-proprietary formats to promote long-term archiving.

Backup and security of data

With the use of the primary data and NPD secondary linked datasets, these datasets will be held locally. For this purpose, we will have the use of a secured partition on the University of Nottingham's research network drive with access provided to the three evaluation team members and an additional research assistant (to be appointed) and specific IT support personnel. Data will be also securely held at times throughout the proposed research grant period on the University of Nottingham's High Performance Computing cluster which can only be accessed from specific workstations in the School of Education.

Access to the network drives is tied to researchers' workstations which are situated in private locked offices. The terminals are all password-protected and authenticated through the University's centralised Information Access Management system. Remote access, where permissions allow for it will only work through remote desktop which has the same password and authentication requirements.

The research drive partition is regularly backed-up and further stored on tape drives. In addition, we will implement version control via the commercial service Github. This allows any changes to script files to be pushed to a private external repository which can then either be rolled back or forked to allow multiple users to make changes without undermining the original file.

Version control via Github is used extensively in the private commercial sector and is now being more frequently used by the research community. Repositories can also be made public (subject to data security considerations) at a later date to allow complete transparency over the data analysis process.

Data protection, rights and access

The NPD data has strict rules in place with regards to privacy given the extensive personal details available with location, demographic and educational achievement data included. The descriptive analyses provide the greatest risk of undermining the privacy of participants and we therefore intend to avoid identifying individual schools where participants could be identifiable. Inferential statistical analyses present far less risk as using Markov Chain Monte Carlo (MCMC) based samples mean that it is impossible to identify individuals.

Transparency, preservation, sharing and licensing

The long-term responsibility for archiving the data will rest with the Education Endowment Foundation and its partners. On the completion of the evaluation, the data and supporting documentation will be released to the Foundation, as well as the developers.

Any academic publications which come from the work carried out during the evaluation will release supporting materials through Harvard's Dataverse repository. This is an open source data repository which provides the advantages of specifically organising data curation on the basis of published academic outputs. It allows the curation of replication code and documentation, provides controls over data access, and importantly provides versioning control for updates to material over time. All material which enables a successful replication (with the inclusion of synthetic versions of the data – to protect the identity of the participants, but maintain the statistical properties of the datasets) will be preserved in the repository.

Sensitive data such as from the linked National Pupil Database have a short period of access. Any sensitive data will be effectively destroyed. If the computer is being decommissioned at the same time as which the data is deleted, then the deletion will be carried out by the University's third party PC disposal partner, Stone. Otherwise it will be completed by IT

Services staff. Data that is stored on hard copy (paper) will be shredded on-site prior to disposal.

Retention of replication material will be permanent, although much of the original secondary data (linked NPD data) kept locally will need to be destroyed within a few months of the end of the project. Our analyses will use open source software such as R and so will need little preparation beyond appropriate descriptions within the script files. Data source documentation will need to be prepared, however, to enable researchers to be able to request data from EEF and the planned deposit of material UK Data Archive to enable replication and follow on analyses.

Adherence

Adherence will be reviewed at the evaluation's progress management meetings and on request from DfE or the Education Endowment Foundation.

Ethical Issues

To ensure that the study has appropriate ethical oversight, each of the participating institutions – University of Oxford, Catch Up/Caxton Trust and the University of Nottingham – will ensure that their respective ethics committees, or equivalent, have reviewed and granted approval for the intervention to go ahead. Consent will be addressed through parental opt-out consent for both the collection and use of the data for the trial and subsequent data archiving. TAs and Teacher consent will be dealt with through opt-in consent for both the observations and interviews. The trial is registered with ISRCTN (www.controlled-trials.com) reference number:

Personnel

The University of Nottingham Evaluation team

Professor Jeremy Hodgen will direct the project and will contribute to all aspects of the project.

Professor Shaaron Ainsworth will advise on aspects of experimental design and will contribute to the impact and process evaluations and report-writing. She will deputise for Hodgen when necessary.

Dr Michael Adkins will conduct the randomisation procedures, quantitative analysis and contribute to all aspects of the study, including drafting the evaluation protocol and report-writing.
Sheila Evans will be responsible for the day-to-day management of the evaluation, maintaining contact with schools and minimizing attrition in addition to undertaking the fieldwork and analysis relating to the process evaluation. S/he will assist with report-writing.
An administrator will be appointed during the spring of 2016 to assist the team, including to maintain contact with schools.

The developers: Catch Up and the University of Oxford

Dr Graham Sigley (Deputy Director, Catch Up) will direct all aspects of the project that come within the responsibility of Catch Up/the Caxton Trust the project and will contribute to all aspects of the project.
Dr Ann Dowker (Department of Experimenatal Psychology, the University of Oxford), will provide support, information advice and guidance in the maximisng the effective implementation of the project on a consultancy basis.
Ms Julie Lawes (Director, Catch Up) will provide strategic input and support where appropriate and will deputise for Dr Graham Sigley where required.
Catch Up Accredited Trainers (CUATs) will deliver all training and briefing sessions and will contribute to the production of any project briefing material that is required in addition to the training and support materials that are included within the Catch Up Numeracy integrated training, resource and support package.
An administrator (Ann Fletcher) has been appointed since the autumn 2015 to assist the development and delivery of the project, including maintaining contact with schools.
The Catch Up office team will provide input and support in the delivery of the project, including the ongoing support that project schools and their staff are entitled to as part of their membership of the Catch Up community resulting from undertaking Catch Up Numeracy training. This ongoing support will continue beyond the time period of the project.

Risk Analysis

We summarise our assessment of the main risks in the table below

Risk	Likelihood	Impact	Action
Failure to recruit schools	Low/Moderate	High	Involve developers and local advocates in recruitment process Recruitment Event Establish timeline for recruitment
Attrition of schools and pupils	Moderate	Moderate	Over-recruit schools and pupils Appropriate financial incentives to control schools
Loss of staff	Low	Low	The School of Education at Nottingham has a large staff team
Fidelity to intervention	Low	Low/Moderate	Development team Monitor through process evaluation
Trustworthy and robust evaluation evidence	Low	Moderate/High	Robust process evaluation including triangulation of evidence and a focus on social validity in addition to fidelity

Evaluation Timeline

Study Protocol Limitations and Future Areas of Research

The design and analysis of the intervention have several limitations that are unavoidable, which would benefit from further investigation. These include the following:

1. Catch Up versus active control may not be reliably different.

There is currently a lack of data of several aspects of the research design. These include the impact of the roll off on effect size, the impact of the PTM test versus the previous Basic Number Screening Test (BNST) and how Catch Up with and without roll off compare.

Firstly, Catch Up with the flexible rolling off of pupils once they reach age appropriate levels has not been previously assessed and while we have designed the evaluation to be able to detect a small effect size we have little basis for an appropriate effect size estimate. Secondly, Catch Up has been built on using the BNST, but given the lack of data on how the intervention will perform with the new PTM test the predictions on effect size have been calculated on the basis of the previously used BNST. Lastly, Catch Up with roll off has not been compared to Catch Up without the flexible delivery mechanism and the impact on the effect size is unknown. While this could have been ameliorated through the use of three arms – specifically to test Catch Up delivered over three terms for all pupils, this was not broadly consistent with the operation of the intervention in practice.

2. The Rolling Off Mechanism

There have been no previous trials which investigated the roll off process. There are a number of potential issues here – forgetting, continue to treat, unreliable judgements, and different approaches between arms.

Firstly, even where roll off works perfectly, pupils may forget before the administration of the delayed post-test (in the Autumn term of the following academic year) which may dilute the effect. Secondly, schools may continue to treat after pupils roll off the study in either arm which may undermine the interpretation of the effectiveness of the flexible delivery model. Thirdly, TAs may not be experienced enough to be able to make reliable judgements on which pupils should roll of the intervention or control. Lastly, roll off may be treated differently within the intervention and control arms – so similar learners are treated differently in the two arms and therefore the dosage may differ.

The issue of forgetting and continuing to treat could have been partly addressed through the use of an immediate post-test. However, there are associated issues such as over-testing and attrition which needed to be accounted for. Ineffective roll off decisions could have been partially addressed through the use of an additional treatment arm where learners received the full dose, however, with high quality process data, it may be possible to highlight whether the roll-off decision works as expected. Lastly with different approaches to roll off between the arms, this could have been partially ameliorated by strictly removing the highest performing third after each term. However, this would have undermined the flexible and TA led approach to roll off.

References

Brown, M., Askew, M., Hodgen, J., Rhodes, V., Millett, A., Denvir, H., & Wiliam, D. (2008). Individual and cohort progression in learning numeracy ages 5-11: Results from the Leverhulme 5-year longitudinal study. In A. Dowker (Ed.), *Mathematical Difficulties: Psychology and Intervention* (pp. 85-108). Oxford: Elsevier.

Browne, W.J. (2012) *MCMC Estimation in MLwiN, v2.26*. Centre for Multilevel Modelling, University of Bristol.

Crawford, C., & Cribb, J. (2013). *Reading and maths skills at age 10 and earnings in later life: a brief analysis using the British Cohort Study*. London: Institute of Fiscal Studies.

Dowker, A. (2009). *What works for children with mathematical difficulties?* Nottingham: DCSF Publications.

Dowker, A., & Sigley, G. (2010). Targeted interventions for children with arithmetical difficulties. *Understanding Number Development and Difficulties, BJEP Monograph Series II, 7*, 65–81.

Gill, J. (2015) *Bayesian Methods: A Social and Behavioral Sciences Approach,* 3rd edition, Boca Raton, FL: CRC Press.

Gillham, B. & Hesse, K. (2001). *Basic Number Screening Test: National Numeracy Strategy Edition: Forms A & B, for Ages 7 to 12 Years. 3rd edition.* London: Hodder Education

GL Assessment (2015) “PT Maths: Technical Information” url: http://www.gl-assessment.co.uk/sites/gl/files/images/PTM%20Technical%20Information_incl%2011T_FINAL.pdf

Holmes, W., & 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. doi: 10.1080/14794802.2013.803779

Marshall, P. (Ed.). (2013). *The tail: How England's schools fail one child in five and what can be done*. London: Profile Books.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA / Amsterdam: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College / International Association for the Evaluation of Educational Achievement (IEA).

Núñez-Peña, M. I., Guilera, G., & Suárez-Pellicioni, M. (2014). The Single-Item Math Anxiety Scale: An Alternative Way of Measuring Mathematical Anxiety. *Journal of Psychoeducational Assessment, 32*(4), 306-317. doi:10.1177/0734282913508528

Raudenbush, S.W., et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from [url]: http://sitemaker.umich.edu/group---based/optimal_design_software.

Rutt, S., Easton, C. & Stacey, O. (2014) "Catch Up Numeracy: Evaluation Report and Executive Summary", London: Education Endowment Foundation.

Sharples, J., Webster, R., & Blatchford, P. (2015). *Making best use of teaching assistants: Guidance Report*. London: Education Endowment Foundation.

Snijders, T.A.B. & Bosker, R.J. (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London: Sage.

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal, 30*(4), 477-494. doi: 10.1080/0141192042000237194