Statistical Analysis Plan for Catch Up® Numeracy

University of Nottingham

INTERVENTION	**Catch Up® Numeracy**
DEVELOPER	Catch Up® [1]
EVALUATOR	University of Nottingham
TRIAL REGISTRATION NUMBER	ISRCTN: 15428227
TRIAL STATISTICIAN	Dr Michael Adkins
TRIAL CHIEF INVESTIGATOR	Professor Jeremy Hodgen
SAP AUTHOR	Dr Michael Adkins
SAP VERSION	0.96
SAP VERSION DATE	25/10/2017
EEF DATE OF APPROVAL	
DEVELOPER DATE OF APPROVAL	

Protocol and SAP changes

If any changes to the protocol impact on the SAP, these should be specified here. Changes made to the SAP after its initial publication should also be logged here.

[1] 'Catch Up® is the working name of The Caxton Trust, a not-for-profit charity registered in England and Wales (1072425) and Scotland (SC047557) as well as a company limited by guarantee (03476510). Catch Up is a registered trademark.'

Table of contents

Introduction

Catch Up 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 statistical analysis plan outlines the planned analysis of a two-arm effectiveness trial targeting primary school pupils (Years 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 the Catch Up Numeracy intervention (Rutt et al. 2014). The plan outlines the study design, randomisation process, calculation of the sample size, a mid-intervention report on recruitment and allocation and the primary and secondary outcome measures. It will also outline our primary and secondary outcome analyses, effect size calculation, missing data and non-compliance issues and finally sub-group analyses.

Study design

Catch Up Numeracy is being evaluated using a two-arm randomised controlled trial with an intervention arm comprising of 75 primary schools against an optimal business-as-usual active matched-time control arm of 75 primary schools.

Children who are part of the Catch Up Numeracy intervention receive 2x15 minute one-to-one sessions with a teaching assistant per week. To prepare them for delivering the intervention, teaching assistants are supplied with detailed session plans and receive three half-day training sessions, each of which involves a follow-up task in school. Each school also appoints a Catch Up Numeracy Coordinator to coordinate the intervention in school and to support the teaching assistants. At the end of each term, the children are reassessed and if they are deemed to have made sufficient progress they roll-off the intervention and stop receiving the additional support. Children in the optimal business-as-usual active control schools receive up to 30 minutes per week support from a teaching assistant, with the school preparing their own support plans for each child and teaching assistants supervised by a school coordinator. Like the intervention arm, the children are assessed on a termly basis and if they are deemed to have made sufficient progress they roll-off the intervention and stop receiving the additional support.

Incentives have been offered to schools allocated to the active control group as detailed in the evaluation protocol. These include a wait-list-type incentive, training in an associated intervention, Catch Up Literacy, during 2017/8. Since this intervention is not focused on numeracy, further data on implementation or attainment will not be collected from schools.

For quantitatively evaluating the impact of Catch Up, children were pre-tested in June 2016 and are due to be post-tested in November 2017 (approximately 4 months after the end of the trial) using the Progress Test in Mathematics (PTM) from GL Assessment. The children were also pre-tested on the attitudes and anxieties towards mathematics using a four-item survey, and will be post-tested in November 2017.

Randomisation

The previous trial (Rutt et al. 2014) randomised at the individual level, with the six pupils per school being equally split into three arms (business-as-usual control, active "matched-time" control, and intervention). However, this design was potentially susceptible to cross-contamination between the matched-time and intervention arms. As a result, the current trial has proceeded on the basis of school-level randomisation to avoid any potential of repeated issues of contamination.

To be entered into randomisation, all state schools were eligible as long as they had not already purchased Catch Up Numeracy and could provide a minimum of six year 4s and six year 5s who were 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 had to provide:

- 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 (UPN, Forename, Surname, Date of Birth, Sex, Free School Meal eligibility and Teaching Assistant assignment)
- Pre-test data for all eligible pupils

Randomisation took place in July 2016, with baseline testing taking place in June prior to schools being randomised. Due to issues of recruitment we randomised schools in several batches, with the first batch randomising the majority of schools (137 schools) into the two conditions.

Our procedure incorporated five core steps. Firstly, we pre-processed the school-level data – checking school names, Unique Reference Numbers and postcodes against Edubase records. Secondly, we set up a split function in R which separately randomised the three regional blocks (R code attached in the appendix) and then worked on the basis of simple randomisation to split schools equally into intervention and active control arms. Thirdly, we ran simulations that indicated that any imbalances between the groups were likely to fall within acceptable levels. We examined the level of imbalance in pupil-level data, publicly available school-level indicators 2015 data – percentage of FSM ever, average KS2 point score, and percentage of pupils with English as an additional language found that across simulations, imbalances on average KS2 score and proportion of EAL students was approximately 2%, and for FSMever was <6% (see the section on imbalance analysis for a discussion of planned sensitivity analyses). Fourthly, we randomised the schools, using a randomly generated seed. Finally, as a double-check, we conducted a balance analysis to confirm that imbalances fell within the bounds predicted by out simulations. Our initial balance analysis is presented in the analysis section, along with an outline for further analyses, and our planned model adjustments to address the imbalance.

Calculation of sample size

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.

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). At the time of the protocol development, there were few examples we could find where evaluators had fitted three-level models of pupils clustered in classes/TAs and schools. We used our professional judgement, along with information from NFER's (Rutt et al., 2014) previous evaluation of Catch Up (which assumed an ICC of 0.05 (2-level model)) and the EEF documentation of ICCs for KS2 Mathematics varying from 0.094 to 0.137.

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. Again, at the time of the protocol development, there was no information available on pre-post correlation for current or previous versions of PTM. We made a professional judgement on this, but part of the reason for equipping a trial with enough schools to detect an MDES of <0.18 was to provide flexibility should our assumptions proved to be overestimates.

This has the effect of reducing the overall variance and boosting the expected statistical power of the study. Given that few previous trials have had three levels of variation, and have used the PTM test these assumptions come with some uncertainty. The results of the calculations can be seen in figure 1.

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. Table 1 contains a further breakdown with and without the PTM covariate included in the analysis.

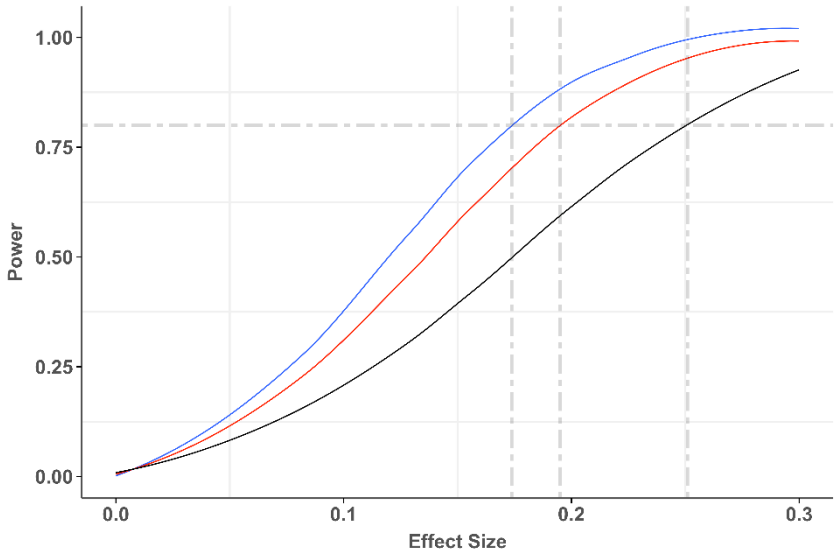
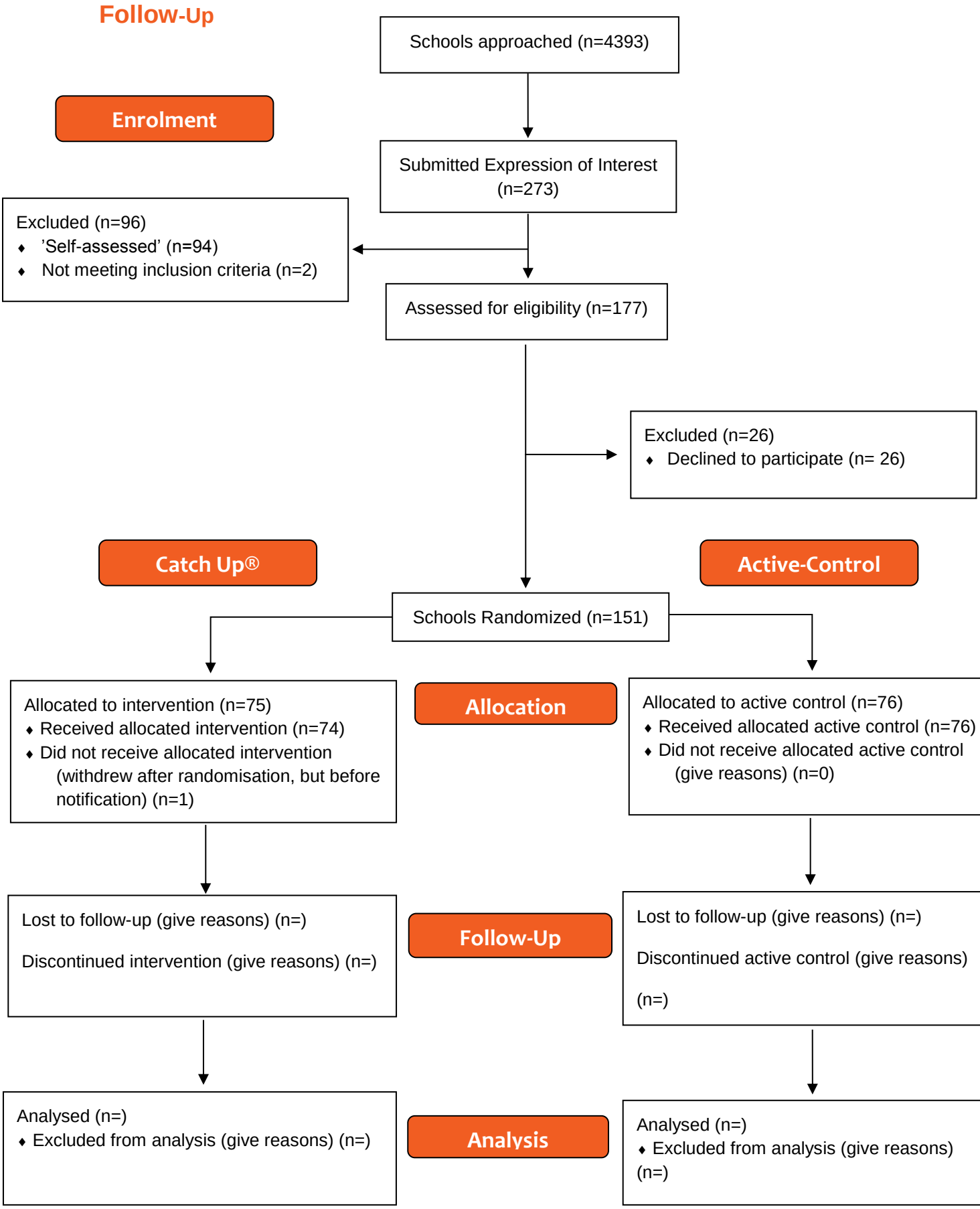


Figure 1: Three power curves highlighting the change in the minimum detectable effect size when the number of pupils per TA is varied (Black=2, Red=4, Blue=6)

Catch Up Power Analysis - All Pupils (including FSM)			
	2 pupils per TA	**4 pupils per TA**	**6 pupils per TA**
Outcome Only	0.27	0.22	0.20
PTM Covariate	0.25	0.20	0.17

Table 1: Power analysis results for outcome only and with PTM covariate. Estimates are subject to rounding.


Follow-Up
Schools approached (n=4393)
Enrolment
Submitted Expression of Interest (n=273)
Excluded (n=96)
♦ 'Self-assessed' (n=94)
♦ Not meeting inclusion criteria (n=2)
Assessed for eligibility (n=177)
Excluded (n=26)
♦ Declined to participate (n= 26)
Catch Up®
Active-Control
Schools Randomized (n=151)
Allocated to intervention (n=75)
♦ Received allocated intervention (n=74)
♦ Did not receive allocated intervention (withdrew after randomisation, but before notification) (n=1)
Allocation
Allocated to active control (n=76)
♦ Received allocated active control (n=76)
♦ Did not receive allocated active control (give reasons) (n=0)
Lost to follow-up (give reasons) (n=)
Discontinued intervention (give reasons) (n=)
Follow-Up
Lost to follow-up (give reasons) (n=)
Discontinued active control (give reasons) (n=)
Analysed (n=)
♦ Excluded from analysis (give reasons) (n=)
Analysis
Analysed (n=)
♦ Excluded from analysis (give reasons) (n=)

Outcome measures

Primary outcome

Our primary outcome of interest – the average difference in the post-test scores between the Catch Up Numeracy treatment group and the active control condition group, will be measured by using the standardised post-test score on the Progress Test in Mathematics (PTM) provided by GL Assessment. This is because PTM has specific age appropriate tests designed for each of the three year groups involved in the study. It will be administered independently in November 2017 using paid invigilators in each region. The score is an amalgam of several measures addressing measurement, geometry, statistics, fluency in facts and procedures, conceptual understanding, problem solving and mathematical reasoning. Children in year 3, 4 and 5 will take levels 8, 9 and 10 respectively.

Previous trials suggest that the use of age-standardised scores may in certain cases inflate effect sizes due to floor and ceiling effects. We will investigate this by conducting an initial descriptive analysis of the histograms of the raw and age standardised pre- and post-test scores to check whether floor and ceiling effects are evident, and whether there are differences between the raw and age standardised scores. We will fit the model to the raw scores, adding an additional non-nested random effect for test-level. We will also explore the use of a Tobit model using the standardised scores which incorporates adjustments for floor or ceiling effects (McBee, 2010). We will report these models alongside the main model. If the models are significantly different, we will comment on the robustness of the effect sizes in the main model.

The outcome measure will be conditioned on a pre-test score which was administered before randomisation by schools in June 2016 using the Progress Test in Mathematics published by GL Assessment with children taking level 7, 8 and 9 in year 2, 3 and 4.

Secondary outcomes

Our secondary outcome of interest – the average difference in pupil attitudes and anxieties to mathematics between the Catch Up Numeracy treatment group and the active control condition group – will be measured using an amalgam of four single-scale items developed from Thomas and Dowker (2000) that can be consulted in Appendix 2. As with the primary outcome, this measure will be conditioned on a pre-test score which was administered before randomisation by schools in June 2016.

Analysis

Our analysis will investigate the effect of Catch Up Numeracy against the active control condition on the basis of intention-to-treat (ITT) using a linear multilevel model estimated by Bayesian Inference. Our primary model on which the report will be based will include pre-test measures, the group allocation and adjustments for the stratification used in the randomisation process and will be discussed further below. In addition to the full model, we will report a simple model to allow comparison between EEF trials. This model will only include the pre-test measures and the group allocation as covariates. While we expect that point estimates and intervals will remain broadly similar between classical and Bayesian approaches when using diffuse or weakly informative priors, Bayesian inference still offers advantages over classically derived estimates. Firstly, the assumption of repeated sampling is not needed, in that the

posterior estimates are based on sequential updating – we update our prior knowledge with new data. This makes estimates more straightforward to interpret. Secondly, Bayesian models average over uncertainty (between the prior information and data) leading to more conservative estimates – particularly in situations with small sample sizes. Thirdly, the posterior distribution allows for a much more straightforward interpretation of models with interaction terms as the posterior predictive distribution can be analysed using different manipulations of the model predictors. Lastly, we can make predictions for new cases – e.g. schools and fully take account of the predictive uncertainty. We will fit several models of increasing complexity analysing their fit using Leave-one-out Cross Validation (LOO-CV). However, our primary varying intercepts model (random effects) on which the impact of the intervention will be assessed is as follows.

Our notation is loosely based on the general practice of the Centre for Multilevel Modelling at the University of Bristol. The individual level of our model has a grand mean of the PTM post-test (represented by β_0), which we allow to vary by membership of TA group and School (represented by the intercept adjustments v_{0k} and u_{0jk}); an individual-level binary treatment covariate where 0 represents those pupils who received the matched-time intervention and 1 which represents those pupils who received the Catch Up Numeracy intervention; a normally distributed and mean-centred pre-test covariate, two group-level stratification dummy variables (using Yorkshire as the control category), and lastly an error term (ε_{ijk}).

$$y_{ijk} = \beta_0 + \beta_1 Treatment_i + \beta_2 Pre-test_i + \beta_3 North-East_k + \beta_4 East-Anglia_k + \underbrace{v_{0k} + u_{0jk}}_{\text{Varying intercepts}} + \epsilon_{ijk}$$

$$\begin{aligned} v_{0k} &\sim \mathcal{N}(0, \sigma^2_{School}), \text{for } k = 1 \ldots K \\ u_{0jk} &\sim \mathcal{N}(0, \sigma^2_{TeachingAssistant}), \text{for } j = 1 \ldots J \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2_{\epsilon}), \text{for } i = 1 \ldots N \end{aligned}$$

As discussed in the software section below, it will be fitted using Stan, an open source Hamiltonian Markov Chain Monte Carlo (MCMC) sampler within R using weakly informative priors. However, to ensure consistency it will also be fitted using MLwiN's Gibbs sampler using diffuse priors, as well as classically using lme4's Maximum Likelihood and MLwiN's IGLS algorithm.

We will report both classical confidence intervals to maintain comparability with other EEF trials in addition to Bayesian credible intervals. We will discuss any significant variation between the Bayesian and classical estimates in the sensitivity analysis.

Interim analyses

We are currently undertaking data simulations to test our Bayesian models, the generated effect size quantities, and the impact of varying forms of missing data on our power analyses. A separate technical report will be finalised in due course.

Imbalance at baseline for analysed groups

Our initial analysis of imbalance shows a small imbalance between treatment and control with regards to pupil-level FSM and school-level percentage of FSM ever. We will extend this analysis further to check imbalance once the NPD sweep is complete to check whether the pupil-level FSM ever continues to show a minor level of imbalance. We will also collect further

school-level data from previous years to average school-level characteristics over time to check whether any imbalance (particularly with percentage of FSM ever) is consistent across a five year period. If this imbalance continues we will run further analyses with appropriate (school) group-level predictors within the varying intercept model.

Pupil-Level Background Characteristics Imbalance Analysis		
	Treatment	**Control**
Male	452 (51%)	455(50%)
Female	436 (49%)	452(50%)
FSM	264(30%)	232(26%)

Table 2: Initial imbalance analysis for pupil-level characteristics

School-Level Background Characteristics Imbalance Analysis		
	Treatment	**Control**
Percentage FSM Ever	M=31.6 SD=16.9	M=26.0, SD=15.8
Average KS2 Score	M=77.5, SD=13.0	M=76.4, SD=16.0
Percentage of EAL Pupils	M=18.7, SD=25.0	M=18.3, SD=25.6

Table 3: Initial imbalance analysis for school-level characteristics

Missing data

Our report will present the results from the complete case analysis (fully observed cases only). However, we will also provide a sensitivity analysis which will examine the robustness of the reported results against multiply imputed data examining the point estimates and credible interval coverage. We intend to use imputation to investigate the robustness of the reported results whatever the level of missingness. It should be noted that there are no agreed cut-offs or thresholds for acceptable percentages of missingness (Dong and Peng 2013) and as Tabachnick and Fidell (2012) argue, the pattern of any missingness is more critical than its extent.

In a Bayesian framework, there are two main options for the handling of missing data. Firstly, missing data can be treated as another random parameter and estimated by building in a missing data sub-model. Secondly, multiple datasets can be imputed by a separate statistical package, then each MCMC chain can be assigned an imputed dataset and the posterior simulations mixed together. While both Stan and MLwiN can handle this process relatively straightforwardly, Stan can only impute covariates with missingness that are continuous, and it is easier to incorporate auxiliary variables using a separate imputation procedure. The likelihood is that minor amounts of missing data will be confined, at least in the primary model to the pre and post-test, but for consistency and compatibility with the classically derived estimates the most appropriate approach is to use a separate imputation procedure.

We will use the software Stat-JR and its new n-level template which is based on the joint modelling which assumes a multivariate normal distribution (MVN), and is capable of fully

imputing our 3+level datasets. While it is impossible to determine the missingness mechanism we will use the imputation tools within Mice, an imputation package within R (van Buuren & Groothuis-Oudshoorn, 2011), to conduct descriptive analyses[2] and to construct a drop-out model. The final imputation model will make full use of the additional auxiliary data (at the individual level – KS1 scores and EYFS data, and at the school-level proportion of pupils that have ever been FSM, KS1 and KS2 scores) within the NPD to increase the plausibility of the Missing at Random assumption. As discussed in Gelman and Hill (2007: 531) it is impossible to be absolutely sure that data is Missing at Random and so it is important to increase the plausibility of this mechanism by including relevant predictors in an imputation analysis. Logistic regression analyses of missingness will not give us a definitive indication of the mechanism, but we will use this approach to help select appropriate auxiliary variables, along with appropriate correlation analyses. We will also check the plausibility of imputed values using the diagnostic techniques outlined in Abayomi, Gelman and Levy (2008).[3]

Non-compliance with intervention

Non-compliance with the intervention can come from pupil-level factors such as chronic absenteeism, as well as school related factors such as availability of TA time and the supply of wider resources. However, for the Catch Up Numeracy intervention, the issue of non-compliance is not straightforward because, by design, the intervention involves different levels of dosage depending on when pupils are rolled off the intervention. Hence, pupils receive different levels of dosage as a result of both compliance and non-compliance. We have collected data to help us estimate dosage at the pupil level by schools submitting time spent with each pupil per week (collected at half-termly intervals) for both intervention and control. At the school level, we have collected fidelity data from schools to create a three-fold category of high, medium and low fidelity. These will be entered into multilevel regression model with pupil dosage at the individual level and fidelity score entered as a school-level covariate.

Secondary outcome analyses

The intervention will also be assessed by the secondary outcome examining pupil anxieties and attitudes to Mathematics. Here we will model and report the scale in a similar form to the primary model (separately from the PTM post-test) to maintain consistency with previous EEF trials. However, we will also explore modelling this simultaneously with our main outcome using a multivariate multilevel model (also referred to as a multiple outcome model (Gelman et al. 2012)). This model adds an additional level of clustering to our previous analyses to account for multiple outcomes. Responses to the multiple outcomes are set at the first level providing the structure of the multivariate model, with level two being clusters of pupils, level three being clusters of TAs and level four being clusters of schools. 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

[2] This will include examining patterns of missingness, as well as generating influx, outflux and fico statistics which are useful for selecting predictors in an imputation model.

[3] These diagnostic techniques include overlaid density comparisons between observed and multiply imputed datasets; numerically comparing the empirical distributions of observed and imputed data using the Kolmogorov-Smirnov test; and bivariate scatter plots to check for internal consistency of missing and observed observations.

as the Bonferroni correction (Snijders and Bosker 2011, 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 MAAQ scale and the PTM post-test. The formula for our secondary outcome analysis is presented below.

We remain as consistent in notation as possible, again being broadly based on the standard notation of the Centre for Multilevel Modelling at the University of Bristol. We gain two additional elements - Z_{1ijkl} which is indicator where 1 is the Progress Test in Mathematics and 0 is the Mathematics Attitudes and Anxieties Questionnaire scale; and Z_{2ijkl} which is $1 - Z_{1ijkl}$. We estimate two intercepts - one for each outcome variable, denoted by β_{01} and β_{02}; two treatment effects (one for each outcome variable) - denoted by β_{11} and β_{21}; and two pre-test effects - again one for each outcome variable, denoted by β_{21} and β_{22}. As there is no level 1 variation specified because level 1 exists solely to define the multivariate structure, individual level error terms are denoted by the notation u, TA-level error terms are now denoted by the notation v and School-level error terms are denoted by the notation f. Error term levels are estimated for both outcome variables.

$$y_{ijkl} = \beta_{01}Z_{1ijkl} + \beta_{02}Z_{2ijkl} + \beta_{11}Z_{1ijkl}Treatment_j + \beta_{21}Z_{2ijkl}Treatment_j + \beta_{12}Z_{1ijkl}Pre-test_j + \beta_{22}Z_{2ijkl}Pre-test_j + \underbrace{u_{1j}Z_{1ijkl} + u_{2j}Z_{2ijkl}}_{\text{Individual level error terms}} + \underbrace{v_{1k}Z_{1ijkl} + v_{2k}Z_{2ijkl}}_{\text{TA level error terms}} + \underbrace{f_{1l}Z_{1ijkl} + f_{2l}Z_{2ijkl}}_{\text{School level error terms}}$$

In the group level models we assume bivariate normal distributions, with means of 0, and estimate three variance-covariance matrices. Diagonal elements are the variances for the two outcome variables at the individual, TA and school-level, and the off-diagonal elements are the correlations between the terms.

$$\begin{pmatrix} f_{1_l} \\ f_{2_l} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma^2_{f_1} & \\ \rho\sigma_{f_1}\sigma_{f_2} & \sigma^2_{f_2} \end{pmatrix}}_{\text{School-level variance-covariance matrix}}\right) \text{for } l = 1 \ldots L$$

$$\begin{pmatrix} v_{1_k} \\ v_{2_k} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma^2_{v_1} & \\ \rho\sigma_{v_1}\sigma_{v_2} & \sigma^2_{v_2} \end{pmatrix}}_{\text{TA-level variance-covariance matrix}}\right) \text{for } k = 1 \ldots K$$

$$\begin{pmatrix} u_{1_j} \\ u_{2_j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma^2_{u_1} & \\ \rho\sigma_{u_1}\sigma_{u_2} & \sigma^2_{u_2} \end{pmatrix}}_{\text{Outcome-level variance-covariance matrix}}\right) \text{for } j = 1 \ldots J$$

Additional analyses

As discussed in the imbalance and non-compliance sections we will fit further models incorporating dosage and group-level predictors such as TA Maths level, School KS2 average, School FSMever average, and fidelity information to investigate the sensitivity of the estimates.

Subgroup analyses

We will follow the primary and secondary outcome models with an investigation of the impact of the intervention on relevant subgroups adding two additional individual level

covariates - sex of participant and Free School Meal (FSM) entitlement. In addition we will include interaction effects for these with the binary treatment variable and we will add appropriate group-level predictors such as TA maths level at level 2, and school averages for KS1 and KS2 at level 3.

Software

As noted above, while the intention is to fit these models using Bayesian inference, 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).

Effect size calculation

We will use the standard practice of existing EEF trials in reporting effect sizes to calculate using total variance. The formula is presented below:

$$ES = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{(\sigma^2_{school} + \sigma^2_{ta} + \sigma^2_{y})}}$$

Effect size quantities will be computed directly in Stan within the "generated quantities" of the model. In MLwiN, effect sizes will be computed from the saved MCMC simulation values within R and from the classically derived estimates in lme4, these will be computed using the same methodology, but through the sim() function from the Applied Regression Modelling package (arm) in R. Across all three processes, credible/confidence intervals can be read off the summary report.

Report tables

We will report ICC statistics including credible intervals using the standardised EEF tables.

References

Dong, Y. and Peng, C-Y.J (2013). "Principled missing data methods for researchers". *SpringerPlus* 2 (222).

Dowker, Ann., Bennett, K., and Smith, L. (2012). "Attitudes to Mathematics in Primary School Children". *Child Development Research*. doi:10.1155/2012/124939.

Gelman, Andrew and Hill, Jennifer (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, Andrew, Hill, Jennifer, and Yajima, Masanao (2012). "Why We (Usually) Don't Have to Worry About Multiple Comparisons". *Journal of Research on Educational Effectiveness* 5.2, pp.189–211.doi:10.1080/19345747.2011.618213.

Holmes, Wayne and Dowker, Ann (2013). "Catch Up Numeracy: a targeted intervention for children who are low-attaining in mathematics". *Research in Mathematics Education* 15.3, pp. 249–265. doi: 10.1080/14794802.2013.803779.

McBee, M. (2010). Modeling Outcomes With Floor or Ceiling Effects: An Introduction to the Tobit Model. Gifted Child Quarterly, 54(4), 314-320. doi:10.1177/0016986210379095

Rutt, Simon, Easton, Claire, and Stacey, Oliver (2014). Catch Up Numeracy: Evaluation Report and Executive Summary. London: Education Endowment Foundation.

Snijders, T.A.B. and Bosker, R. (2011). *Multilevel Analysis: An Introduction To Basic And Advanced Multilevel Modeling*. London: Sage.

Tabachnick, B.G. and Fidell, L.S. (2012). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.

Thomas, Gemma and Dowker, Ann (September 2000). "Mathematics anxiety and related factors in young children". Proceedings of the British Psychological Society Developmental Section Conference, Bristol, UK.

Tymms, Peter (2004). "Effect sizes in multilevel models". Ed. by Ian Schagen and Karen Elliot. Chap. *But what does it mean? The use of effect sizes in educational research*, pp. 55–66. url: https://www. nfer.ac.uk/publications/SEF01/SEF01.pdf

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. 2011, 45(3), 67. doi:10.18637/jss.v045.i03

Appendix 1

```
# Function to randomly split schools dataset into intervention vs control.
splitdf <- function(dataframe, seed=NULL) {
        if (!is.null(seed)) set.seed(seed)
        index <- 1:nrow(dataframe)
        interventionindex <- sample(index, trunc(length(index)/2))
        Intervention <- dataframe[interventionindex, ]
        Control <- dataframe[-interventionindex, ]
        list(Intervention=Intervention,Control=Control)
}

# Apply the function
splits <- splitdf(schools, seed=NULL)

#It returns a list - two data frames
str(splits)

# There are x observations in each data frame
lapply(splits,nrow)

#View the first few columns in each data frame
lapply(splits,head)

#Save the Intervention and Control sets as data frames
Intervention <- as.data.frame(splits$Intervention)
Control <- as.data.frame(splits$Control)

# Add treatment indicator and merge dataframes
Intervention$Treatment <-"Treatment"
Control$Treatment <-"Control"
schools<-rbind(Intervention, Control)
rm(Intervention, Control, splits, splitdf)
```

Appendix 2

Using a Likert scale of “Very easy” to “Very hard”, pupils were asked to circle the description that best described their feelings towards Maths and Reading:

1. How easy do you find maths?
2. How much do you like maths?
3. How easy do you find reading?
4. How much do you like reading?