Statistical Analysis Plan for Young Enterprise: Mathematics in Context

University of Nottingham

INTERVENTION	**Young Enterprise: Mathematics in Context**
DEVELOPER	Young Enterprise
EVALUATOR	University of Nottingham
TRIAL REGISTRATION NUMBER	ISRCTN58590757
TRIAL STATISTICIAN	Dr Michael Adkins
TRIAL CHIEF INVESTIGATOR	Professor Geoff Wake
SAP AUTHOR	Dr Michael Adkins
SAP VERSION	1.0
SAP VERSION DATE	25/05/2018

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.

- The evaluation protocol (version 1.0) states: "We will use a school-level randomisation approach using a block design stratified by geographical area and either everFSM or attainment. We will model this process prior to randomisation and, if necessary amend the protocol in May 2017. … During 2016-17, we will investigate whether further stratification by school factors (e.g., FSM, GCSE, examination board) is necessary to achieve sufficiently balanced intervention and groups." (p.14). As a result of this modelling, school randomisation was stratified by region (North/South England) and by school-level everFSM.
- The evaluation protocol (version 1.0) states that randomisation would take place in Summer 2017 (p.13). In the event, due to recruitment difficulties, randomisation was conducted in two batches: 122 schools on 12th July 2017, and 3 on 18th September 2017.

Table of contents

Introduction

"Young Enterprise: Maths in Context" is an intervention that seeks to improve children's financial capability, and specifically their financial knowledge and understanding, applied numeracy and problem-solving skills. This large England-wide efficacy trial follows an earlier project funded by the London Schools Excellence Fund (PFEG, 2015). The earlier project involved a small-scale evaluation of the impact on student attainment involving comparison of the intervention group (260 students) to a control group (101 students) who were taught by the same teachers.[1] The intervention group made greater gains on a levelled GCSE-based assessment.

This statistical analysis plan outlines the planned analysis of a two-arm efficacy trial targeting secondary school pupils (Year 10) from 125 schools recruited. It will discuss 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 discuss our primary and secondary outcome analyses, effect size calculation, missing data and non-compliance issues, and finally sub-group analyses.

Study design

Young Enterprise: Mathematics in Context is being evaluated using a two-arm randomised controlled trial with an intervention arm comprising of 63 secondary schools against a business-as-usual control arm of 62 secondary schools.

Students who are part of the Young Enterprise intervention group will receive a series of 10-12 lessons, each focused on a specific area of mathematics in the context of financial capability from their mathematics teacher. To prepare them for delivering the intervention, each intervention school has identified a lead teacher. Lead teachers are expected to model the teaching approach by implementing the lessons and pedagogies introduced in the training in their own lessons *and* more widely within their schools by providing 'cascade' training to at least three other Year 10 mathematics teachers.

Incentives have been offered to schools allocated to the control group as detailed in the evaluation protocol. These schools will be provided with a payment of £1000 on receipt of final GCSE data in Autumn 2019.

[1] https://www.london.gov.uk/sites/default/files/pfeg_london_lead_teachers_in_financial_mathematics_final_report.pdf

For quantitatively evaluating the impact Young Enterprise: Mathematics in Context, we will use KS2 results as the pre-test, and will request GCSE Mathematics score as part of the National Pupil Database extract, along with Uniform Marking Scale (UMS) scores from schools. Subject to a satisfactory response rate, in order to improve discrimination and strengthen the statistical modelling, we will use UMS scores to assess the primary outcome as discussed in the evaluation protocol. We propose to impute UMS scores if necessary and consider that it is reasonable to impute where missingness on the UMS scores is up to 10%. However, if the response rate is such that attrition would affect the security rating of findings (where attrition is > 20%), we will use GCSE numerical grades as the primary outcome.

In the summer of 2018, we will conduct a dummy run of the GCSE UMS and item-by-item data collection process with several secondary schools that are independent of the project in order to ensure that extensive data collection of finer grained GCSE data can be undertaken effectively with trial schools. We will also explore the possibility of imputing GCSE UMS data as we should be able to build a strong predictive model with the inclusion of GCSE numerical grade and prior attainment data – e.g. KS2 and KS1 scores, and include the results as a sensitivity analysis in the final report. Given that we need to scope the potential of these approaches, we plan to update the SAP and protocol where appropriate later in 2018 (See missing data section for a further discussion).

Calculation of sample size

We performed two sets of power calculations for the primary and secondary outcomes. We used Raudenbush et al.'s (2011) Optimal Design software to estimate statistical power on the basis of recruiting schools in 2-arm and 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 – students are clustered in classes by teachers which are then further clustered in schools.

As the research design involves cascade training to multiple teachers, and the likely number of students involved is very high, the greatest change in power is from adding additional schools and so we varied the number of schools that would be part of the intervention.

We fixed the following parameters: α=0.05 (which refers to the probability of rejecting the hypothesis tested when it is true – 5%), 25 students per class and 4 classes per school, intra-cluster correlation for level 2 (class teachers) =0.05 and for level 3 (schools) =0.165 (which

refers to the variance between participants with the same teacher and for those in the same school). Here we are assuming that as students are generally in sets within mathematics, this should reduce the variation observed at the class-level. Since the EEF (2015) guidance on ICC indicate an ICC for GCSE mathematics of 0.165, we consider our assumptions overall to be relatively conservative.[2] We also included an additional pre-test (KS2 Mathematics score) covariate as a school level aggregate with the assumption that the post and pre-test have a correlation of 0.7 setting the level 3 variance explained at 0.70^2=0.49. This has the effect of reducing the overall variance and boosting the expected statistical power of the study.

This produces an estimated minimum detectable effect size (MDES) of 0.167 (for 130 schools) and 0.175 (for 120 schools). For the FSM sub-group analysis, a conservative estimate of 8 FSM students per class (approximately 30%) produces an MDES of 0.18.

Young Enterprise: Mathematics in Context Power Analysis			
	120 Schools	**125 Schools**	**130 Schools**
All pupils with outcome only	0.222	0.217	0.213
FSM pupils only and KS2 Covariate	0.188	0.183	0.180
All pupils with KS2 covariate	0.174	0.171	0.167

Table 1: Minimum Detectable Effect Size (MDES) for outcome only and with PTM covariate. Estimates are subject to rounding.

Randomisation

All state schools were eligible as long as the school had not already taken part in Young Enterprise's previous Maths in Context trial, funded by London Schools Excellence Fund pilot (PFEG, 2015) and could provide a minimum of four classes of year 10s who are eligible for the intervention. Recruitment aimed to maximise the number of schools with an above average proportion of students qualifying as everFSM in order that the proportion in the sample as a whole was at least 29.3%. In addition, in order to be entered into the randomisation the schools had to provide:

- A signed Memorandum of Understanding

[2] Education Endowment Foundation. (2015). Intra-cluster correlation coefficients.

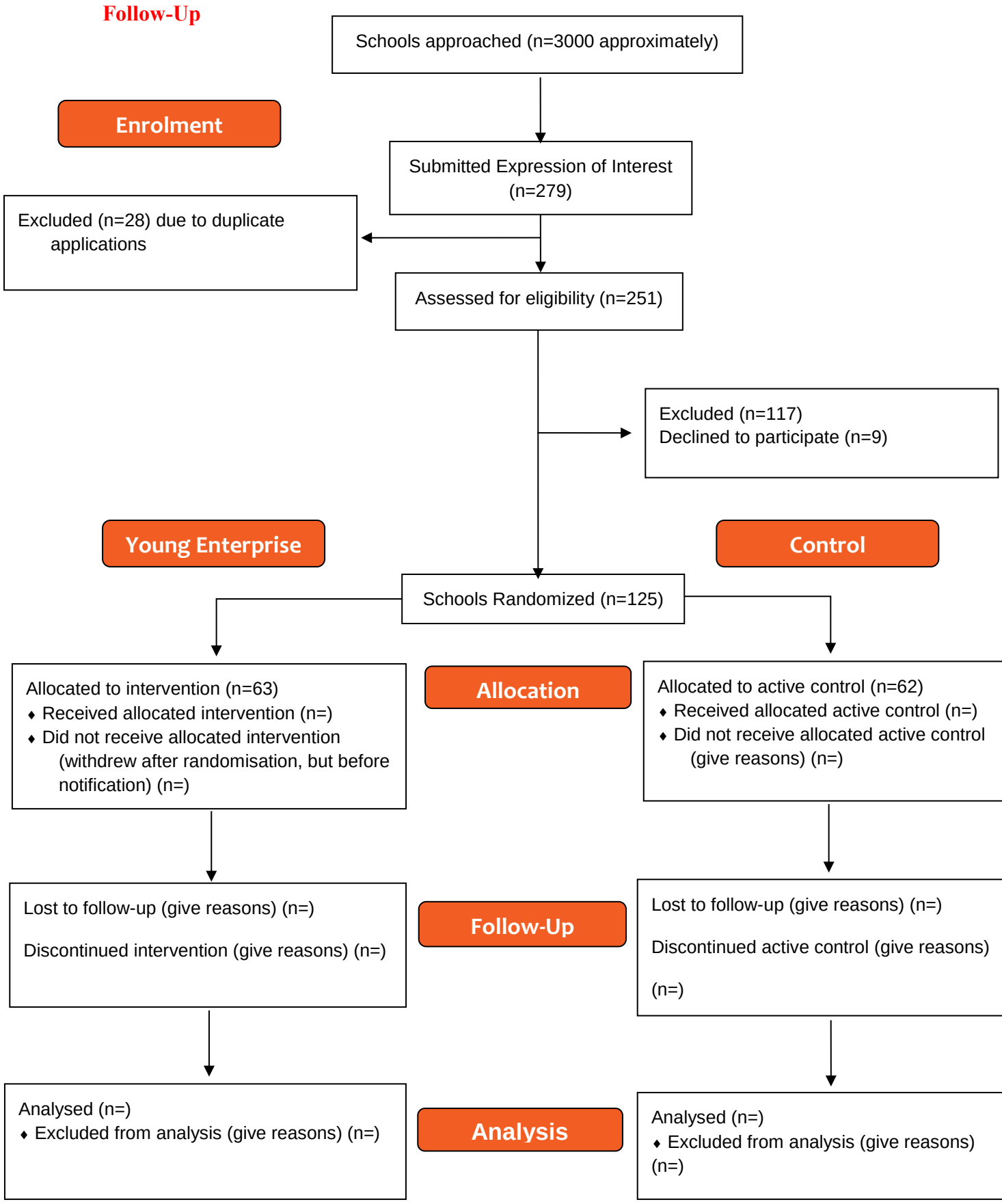
- Confirmation that consent forms have been sent out and any opt-outs
- Provision of pupil data for those identified as eligible: Class teacher ID, Unique Pupil Number (UPN), Forename, Surname, Date of Birth and Gender
- Pre-test data for all eligible pupils (financial capabilities assessment only)
- Names of lead teachers

Randomisation took place in July and September 2017, with baseline testing on the financial capabilities assessment taking place in June prior to schools being randomised. Due to issues of recruitment we randomised schools into the two conditions in two batches: 122 schools on 12th July 2017 using the block approach outlined below, and 3 schools on 18th September using a simple randomisation approach[3].

The evaluation protocol (version 1.0) proposed a school-level randomisation approach using a block design stratified by geographical area and either everFSM or attainment. The geographical location of schools was more diverse than planned, and, hence, following discussion with the developer, we decided to stratify schools into just two regions: North and South. Hence, an eight-block design was used. For the second batch, a simple randomisation procedure was adopted due to the small number of schools involved.

The main block randomisation procedure incorporated three core steps (see appendix 1), with an additional two error checking steps. First, we pre-processed the school-level data - checking school names, Unique Reference Numbers and postcodes against Edubase records. Second, we set up a split function in R which worked on the basis of simple randomisation to split schools equally into intervention and active control arms (for use within the everFSM and North-South blocks). Third, we then randomised the schools, using a random number generating (RNG) seed. This was the value of the FTSE 250 at midday on the day of randomisation (12th July 2017) (Seed=19267) which we used for both rounds of randomisation.

[3] Our aim here was to maximise the number of schools in the trial given our aim of recruiting 130 schools, but with 3 schools a second blocked randomisation approach was not possible.


Follow-Up
Schools approached (n=3000 approximately)
Enrolment
Submitted Expression of Interest (n=279)
Excluded (n=28) due to duplicate applications
Assessed for eligibility (n=251)
Excluded (n=117)
Declined to participate (n=9)
Young Enterprise
Control
Schools Randomized (n=125)
Allocated to intervention (n=63)
♦ Received allocated intervention (n=)
♦ Did not receive allocated intervention (withdrew after randomisation, but before notification) (n=)
Allocation
Allocated to active control (n=62)
♦ Received allocated active control (n=)
♦ Did not receive allocated active control (give reasons) (n=)
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

As mentioned above we will use GCSE mathematics as the primary outcome measure. In order to improve discrimination and strengthen the statistical modelling, where possible we will use GCSE Uniform Marking Scale (UMS) scores rather than grades and these will collected from directly from schools (subject to satisfactory returns).[4] We will use KS2 national test scores in mathematics (KS2_MATPOINTS**)** as a pre-test score for pupils, which will be matched to the UMS score through an extract of the National Pupil Database.

Secondary outcomes

We will use two secondary outcome measures:

I. An amalgamated scale for financial and problem-solving items from the GCSE mathematics papers. We will collect item-by-item data directly from schools.
II. A bespoke financial knowledge and understanding instrument based on the MAS Financial Capability Outcomes Framework (Bagwell et al, 2014). This secondary measure was administered by schools in May / June 2017 prior to randomisation, and will then be re-administered as a post-test in September 2018. As a secondary measure, we do not consider independent administration (or blinding) to be necessary.

Analysis

Our analysis will investigate the effect of Young Enterprise: Mathematics in Context against the business-as-usual control condition on the basis of intention-to-treat (ITT) using a linear multilevel model estimated by Bayesian Inference. 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

[4] GCSE grades are not designed to form a linear scale. The use of two tiers creates a censored variable for sub-populations of the year group. In order to simplify the modelling, we will use the UMS score, which can be modelled using linear regression techniques. The UMS score is a tool that all exam boards use to standardise marks awarded on papers across the different exam boards and paper tiers. The conversions are provided by the exam boards on their websites. For example: http://www.ocr.org.uk/i-want-to/convert-raw-marks-to-ums/

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 GCSE Mathematics score post-test (represented by β_0), which we allow to vary by membership of class 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 control condition and 1 which represents those pupils who received the Young Enterprise: Mathematics in Context intervention; a normally distributed and mean-centred pre-test covariate, two randomisation covariates – everFSM and North-South location (β_3 and β_4) and lastly an error term (ε_{ijk}).

$$y_{ijk} = \beta_0 + \beta_1 Treatment_i + \beta_2 Pre-test_i + \beta_3 South_k + \beta_4 everFSM_k + \underbrace{v_{0k} + u_{0jk}}_{Varying\ intercepts} + \epsilon_{ijk}$$

$$v_{0k} \sim N(0, \sigma^2_{School}) \; for \; k = 1 \ldots K$$

$$u_{0j} \sim N(0, \sigma^2_{Class}) \; for \; j = 1 \ldots J$$

$$\epsilon_i \sim N(0, \sigma^2_{\epsilon}) \; for \; i = 1 \ldots N$$

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 (WIP). The aim of using WIP is to "…to 'regularize' the posterior distribution, that is, to keep it roughly within reasonable bounds—but without attempting to fully capture one's scientific knowledge about the underlying parameter" (Gelman et al, 2014, 51). Our starting point will be the default priors – normal priors on the betas and half-Cauchy's on the variance parameters. However, to ensure consistency it will also be fitted using MLwiN's Gibbs sampler using diffuse priors (see Browne, 2015, 4-5), as well as classically using lme4's Maximum Likelihood and MLwiN's IGLS algorithm.

We will report Bayesian credible intervals for the main report, but we will also fit the models classically to allow comparison with other EEF trials and 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 analysis

Our initial analysis of imbalance shows good balance between treatment and control with regards to our two blocking variables school-level percentage of FSM and the regional dichotomy of North vs. South. We will extend this analysis further to check imbalance once the NPD sweep is complete.

School-Level Background Characteristics Imbalance Analysis		
	Treatment	**Control**
Percentage FSM Ever	M=30.47 SD=15.17	M=30.52 SD=15.57
Percentage of Schools in North and South	North=56% South=44%	North=56% South=44%

Table 3: Initial imbalance analysis for school-level characteristics

Missing data

We have designed our data collection procedures to minimise missingness by collecting additional data from schools (such as FSM), liaising further with schools to address missing data when returning to collect additional data, and providing a sufficient incentive to collect the final round of data at the end of the project. However as outlined above, the greatest threat relating to missing data is that we fail to collect sufficient UMS data from schools in Autumn 2019. If this is the case, we will use GCSE grades collected from the NPD for the primary analysis. In addition, we will impute UMS scores and report the results of this imputed dataset, comparing these to the primary analysis.

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, although we will bear in mind the thresholds in the EEF Security of Findings Guidance.

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 that 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 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, KS2 and GCSE pass rate 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).[5]

[5] 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

Non-compliance with intervention

The following definition of compliance and related evidence has been agreed with the developer:

- Adequate staffing: School identified one lead teacher, and (at least) three other teachers, and associated Y10 classes [Full compliance required, aside from schools with fewer than 4 teachers and classes; Evidence: School data / developer records / checks by Consultants]
- Attendance at training: Lead teachers attend 1 day training [Full compliance required; Evidence: developer attendance records]
- Cascade training: Lead teachers provide cascade training for three other teachers [Full compliance required; Evidence: Teacher survey / checks by Consultants]
- Consultant support: Consultants provide three days equivalent time of mentoring support delivered over up to eight visits [Full compliance required: Evidence: developer records]
- Lessons: All classes should be taught the Maths in Context lessons [Minimum / optimal compliance: 10 /12 lessons; Evidence: developer records / checks by Consultants / teacher survey]

We will investigate the effects of non-compliance by using an instrumental variables (IV) approach

We will also investigate the effects of "non-compliance" in the control group. The Mathematics in Context lessons are not publically available, so schools in the control group will not have access to the intervention materials. However, there may be some schools, or teachers, in the control group who teach significant amount of financial mathematics, and we will attempt to capture these "always compliers" using survey data. If sufficiently robust data are available, we will investigate the effect of this non-compliance in the control group using a per-protocol approach.

Kolmogorov-Smirnov test; and bivariate scatter plots to check for internal consistency of missing and observed observations.

Secondary outcome analyses

As discussed above, we have two secondary outcome measures – the amalgamated scale for financial and problem-solving items from the GCSE mathematics papers, and a bespoke financial knowledge and understanding instrument. We will model these two outcomes separately using a similar model specification to the primary outcome. For the amalgamated scale of financial and problem-solving items, we will use the KS2 mathematics result as the pre-test. For the bespoke financial knowledge and understanding instrument, we have designed and piloted a pre-test and will use this in place of the KS2 mathematics result.

Additional analyses

As discussed in the non-compliance section we will fit further models incorporating dosage and group-level predictors such as lead vs. cascade trained teacher, School KS2 average and school-level attainment, and fidelity information to investigate the sensitivity of the estimates.

We will also explore modelling the bespoke financial knowledge and understanding instrument simultaneously with our main outcome using a multivariate multilevel model (also referred to as a multiple outcome model (Gelman et al. 2012)). We have chosen to limit this as the first of the two secondary outcomes is derived from the main GCSE outcome variable. This model adds an additional level of clustering to our previous analyses to account for multiple outcomes – the UMS GCSE Mathematics score and the post-intervention financial knowledge and understanding survey. Responses to the multiple outcomes are set at the first level providing the structure of the multivariate model, with level two being pupils, level three being classes and level four being schools. This approach offers four significant advantages in understanding the relationship between GCSE mathematics performance and financial understanding. 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 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 financial knowledge and understanding instrument and the GCSE Mathematics result. 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}; two pre-test effects - again one for each outcome variable, denoted by β_{21} and β_{22}; and four randomisation stratifiers – two for each outcome variable denoted by β_{13}, $\beta_{14,}$ β_{23} and $\beta_{24.}$ 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*, class-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.

$$y_{ijkl} = \beta_{01}Z_{1ijkl} + \beta_{02}Z_{2ijkl} + \beta_{11}Z_{1ijkl}Treatment_j + \beta_{21}Z_{2ijkl}Treatment_j + \beta_{12}Z_{1ijkl}Pre-test_j + \beta_{22}Z_{2ijkl}Pre-test_j + \beta_{13}Z_{1ijkl}South_l + \beta_{23}Z_{2ijkl}South_l + \beta_{14}Z_{1ijkl}FSMever_l + \beta_{24}Z_{2ijkl}FSMever_l + \underbrace{u_{1j}Z_{1ijkl} + u_{2j}Z_{2ijkl}}_{\text{Individual level error terms}} + \underbrace{v_{1k}Z_{1ijkl} + v_{2k}Z_{2ijkl}}_{\text{Class level error terms}} + \underbrace{f_{1l}Z_{1ijkl} + f_{2l}Z_{2ijkl}}_{\text{School level error terms}}$$

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, class and school-level, and the off-diagonal elements are the correlations between the terms.

$$\begin{pmatrix} f_{1l} \\ f_{2l} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \underbrace{\begin{pmatrix} \sigma^2_{f1} & \\ \rho\sigma_{f1}\sigma_{f2} & \sigma^2_{f1} \end{pmatrix}}_{School-level\ variance-covariance\ matrix} \right) for\ l = 1 \ldots L$$

$$\begin{pmatrix} v_{1l} \\ v_{2l} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \underbrace{\begin{pmatrix} \sigma^2_{v1} & \\ \rho\sigma_{v1}\sigma_{v2} & \sigma^2_{v1} \end{pmatrix}}_{Class-level\ variance-covariance\ matrix} \right) for\ k = 1 \ldots K$$

$$\begin{pmatrix} u_{1l} \\ u_{2l} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \underbrace{\begin{pmatrix} \sigma^2_{u1} & \\ \rho\sigma_{u1}\sigma_{u2} & \sigma^2_{u1} \end{pmatrix}}_{Outcome-level\ variance-covariance\ matrix} \right) for\ j = 1 \ldots J$$

Subgroup analyses

Additional models of greater complexity will be fitted which will include sex of participant, 'FSM ever' entitlement (defined as any pupil who has ever been classified as in receipt of free school meals), foundation or higher tier paper, 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 including whether the teacher was a school-lead or cascade trained and students' GCSE examination tier.

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.

Minimum detectable effect size at different stages

Stage	N [schools/ pupils] (n= intervention ; n=control)	Correlation between pre-test (+other covariates) & post-test	ICC	Blocking/ stratification or pair matching	Power	Alpha	Minimum detectable effect size (MDES)
Protocol							
Randomisation							
Analysis (i.e. available pre- and post-test)							

Baseline comparison

Variable	Intervention group		Control group	
School-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
	…	…	…	…
School-level (continuous)	n (missing)	[Mean or median]	n (missing)	[Mean or median]
	…	…	…	…
Pupil-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
	…	…	…	…
Pupil-level (continuous)	n (missing)	[Mean or median]	n (missing)	[Mean or median]

Primary analysis

	Raw means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (intervention; control)	Hedges g (95% CI)	p-value
	…	…	…	…	…	…	…

References

Abayomi, K., Gelman, A. & Levy, M. (2008) "Diagnostics for multivariate imputations" Appl. Statist. 57(3), pp.273-291

Bagwell, S., Hestbaek, C., Harries, E., Kail, A. (2014) Financial capability outcome frameworks, NPC [url] http://www.thinknpc.org/publications/financial-capability-outcome-frameworks/

Browne, W. (2015) "MCMC estimation in MLwiN Version 2.32" Bristol: Centre for Multilevel Modelling, University of Bristol. [url] http://www.bris.ac.uk/cmm/media/software/mlwin/downloads/manuals/2-32/mcmc-web.pdf.

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

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

Gelman, A., Hill, J., & Masanao, Y. (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.

Gelman, A., Carlin, J., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2014) *Bayesian Data Analysis*, 3rd edition, Boca Raton, FL: CRC Press.

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.

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

Example code from randomisation blocked design using everFSM and North vs. South:

```
census <- subset(census, select=c(URN, LA, ESTAB, PNUMFSMEVER))
attainment <-subset(attainment, select=c(URN, LEA, ESTAB, PTL2BASICS_LL_PTQ_EE))
Schools<- merge(Schools, edubase, by=c("URN"), all.x=TRUE)
Schools<- merge(Schools, census, by=c("URN"), all.x=TRUE)
Schools<- merge(Schools, attainment, by=c("URN"), all.x=TRUE)
Schools <-subset(Schools, select=c(URN, School.Name, LA.ESTAB, GOR..name., PNUMFSMEVER,
PTL2BASICS_LL_PTQ_EE))
colnames(Schools)[4] <- "GOR"
Schools$GOR <- factor(Schools$GOR)
Schools$PNUMFSMEVER <- as.numeric(sub("%","",Schools$PNUMFSMEVER))

round(Schools$PNUMFSMEVER)
ApplyQuantiles <- function(x) {
  cut(x, breaks=c(quantile(Schools$PNUMFSMEVER,  probs = seq(0, 1, by = 0.25))),
      labels=c("0-25","25-50","50-75","75-100"), include.lowest=TRUE)
}

Schools$Quantile <- sapply(Schools$PNUMFSMEVER, ApplyQuantiles)

FSM1_N <-subset(Schools, Schools$Quantile=="0-25" & Schools$NorthSouth=="North")
FSM2_N <-subset(Schools, Schools$Quantile=="25-50" & Schools$NorthSouth=="North")
FSM3_N <-subset(Schools, Schools$Quantile=="50-75" & Schools$NorthSouth=="North")
FSM4_N <-subset(Schools, Schools$Quantile=="75-100" & Schools$NorthSouth=="North")
FSM1_S <-subset(Schools, Schools$Quantile=="0-25" & Schools$NorthSouth=="North")
FSM2_S <-subset(Schools, Schools$Quantile=="25-50" & Schools$NorthSouth=="South")
FSM3_S <-subset(Schools, Schools$Quantile=="50-75" & Schools$NorthSouth=="South")
FSM4_S <-subset(Schools, Schools$Quantile=="75-100" & Schools$NorthSouth=="South")

rm(attainment, census, edubase)


splitdffsm1 <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  is.odd <-function(x) !x %% 2 == 0
  index <- 1:nrow(dataframe)
  odd <- is.odd(length(index))
  noise <-rnorm((nrow(dataframe)),0,1)
  if (odd==TRUE) {size <- round(trunc(length(index)/2) + noise[1])} else { size <- trunc(length(index)/2)}
  interventionindex <- sample(index, 16)
  Intervention <- dataframe[interventionindex, ]
  Control <- dataframe[-interventionindex, ]
  list(Intervention=Intervention,Control=Control)
}

splitdffsm2 <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  is.odd <-function(x) !x %% 2 == 0
  index <- 1:nrow(dataframe)
  odd <- is.odd(length(index))
  noise <-rnorm((nrow(dataframe)),0,1)
  if (odd==TRUE) {size <- round(trunc(length(index)/2) + noise[1])} else { size <- trunc(length(index)/2)}
  interventionindex <- sample(index, 15)
  Intervention <- dataframe[interventionindex, ]
  Control <- dataframe[-interventionindex, ]
  list(Intervention=Intervention,Control=Control)
}

splitdffsm3 <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
```

```
  is.odd <-function(x) !x %% 2 == 0
  index <- 1:nrow(dataframe)
  odd <- is.odd(length(index))
  noise <-rnorm((nrow(dataframe)),0,1)
  if (odd==TRUE) {size <- round(trunc(length(index)/2) + noise[1])} else { size <- trunc(length(index)/2)}
  interventionindex <- sample(index, 15)
  Intervention <- dataframe[interventionindex, ]
  Control <- dataframe[-interventionindex, ]
  list(Intervention=Intervention,Control=Control)
}

splitdffsm4 <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  is.odd <-function(x) !x %% 2 == 0
  index <- 1:nrow(dataframe)
  odd <- is.odd(length(index))
  noise <-rnorm((nrow(dataframe)),0,1)
  if (odd==TRUE) {size <- round(trunc(length(index)/2) + noise[1])} else { size <- trunc(length(index)/2)}
  interventionindex <- sample(index, 15)
  Intervention <- dataframe[interventionindex, ]
  Control <- dataframe[-interventionindex, ]
  list(Intervention=Intervention,Control=Control)
}

splits_1 <- splitdffsm1(FSM1_N)
splits_2 <- splitdffsm2(FSM2_N)
splits_3 <- splitdffsm3(FSM3_N)
splits_4 <- splitdffsm4(FSM4_N)
splits_5 <- splitdffsm5(FSM1_S)
splits_6 <- splitdffsm6(FSM2_S)
splits_7 <- splitdffsm7(FSM3_S)
splits_8 <- splitdffsm8(FSM4_S)
```