

Effectiveness Trial of 1stClass@Number 1

Statistical Analysis Plan

Evaluator (institution): National Foundation for Educational Research

Principal investigator(s): Pippa Lord



Education
Endowment
Foundation

PROJECT TITLE	Effectiveness trial of 1 st Class@Number 1
DEVELOPER (INSTITUTION)	Edge Hill University
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Pippa Lord
PROTOCOL AUTHOR(S)	Chris Morton, Aarti Sahasranaman
TRIAL DESIGN	Two-arm randomised controlled trial with within school random allocation at pupil-level
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	6 – 7 years old, KS1
NUMBER OF SCHOOLS	235
NUMBER OF PUPILS	1,880
PRIMARY OUTCOME MEASURE AND SOURCE	Quantitative Reasoning Test; test administered by teachers at baseline and by NFER test administrators at endpoint and data shared with NFER
SECONDARY OUTCOME MEASURE AND SOURCE	NA

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [<i>latest</i>]		
1.1		
1.0 [<i>original</i>]		N/A

Table of contents

SAP version history	1
Table of contents	2
Introduction	3
Design overview	4
Sample size calculations overview	5
Randomisation	6
Primary outcome measure	7
Analysis	7
Primary outcome analysis	8
Secondary outcome analysis	8
Subgroup analyses	8
Sensitivity check: impact of online training	9
Sensitivity check: allowing the impact of the intervention to vary between schools	10
IPE regression modelling	10
Longitudinal follow-up analyses	11
Imbalance at baseline and sample representativeness	11
Missing data	12
Compliance	13
Intra-cluster correlations (ICCs)	14
Effect size calculation	14
References	16
Appendix A	17
Appendix B	18

Introduction

1stClass@Number 1 is one of a suite of mathematical interventions developed by Every Child Counts¹, run by Edge Hill University (EHU) and set up with support from the Department for Education (DfE) in 2018. It is designed for children who need further support at the level of the Year 1 curriculum and is delivered by trained Teaching Assistants (TAs) to small groups of four pupils.

The Education Endowment Foundation (EEF) has previously commissioned an evaluation of 1stClass@Number 1 through an effectiveness trial. This RCT found that pupils who received 1stClass@Number 1 made, on average, two additional months' progress in maths compared to pupils who did not receive the intervention. Although this finding had a high security rating, the primary impact estimate was not statistically significant. Though this trial was not powered for FSM pupils, there was also evidence to suggest that 1stClass@Number 1 was not as effective for FSM pupils as for pupils not eligible for FSM, motivating a focus on FSM pupils in the current evaluation.

The current evaluation is a randomised controlled trial, with randomisation occurring within each school. Eight pupils are selected for randomisation at each school: the full selection process is described in the study protocol but is summarised here. An initial shortlist of between 10 and 16 suitable pupils was nominated by teachers at each school, from which the final eight were suggested by NFER and confirmed by schools. The selection of the eight pupils was done by administering the Sandwell Early Numeracy Test – Revised (SENT-R). Pupils were prioritised for selection according to their SENT-R score and FSM status. Pupils were selected according to priority groups, which were (ordered by highest to lowest priority): (i) FSM pupils with SENT-R scores of 49 or below, (ii) non-FSM pupils with SENT-R scores of 40 or below, (iii) non-FSM pupils with SENT-R scores of 41-49 and (iv) any pupils with SENT-R scores above 49. The eight pupils for randomisation were selected from group (i), then if more pupils were needed to make up the eight, these were taken from group (ii), and so on. The suggested list of eight pupils that was prepared as above was then shared with schools for confirmation.

Intervention pupils receive 30 half hour numeracy sessions over a 15-week period (with 10 delivery weeks, three half hour sessions a week) in addition to usual maths teaching. The sessions are delivered by teaching assistants, who are being trained in the 1stClass@Number 1 programme over the same 15-week period. Control pupils continue to be taught as usual and will only be eligible to receive 1stClass@Number 1 after the trial ends.

Due to the recruitment needs of the trial, the intervention will be delivered in two cohorts, with cohort 1 schools delivering the intervention to pupils starting in the 2023/24 autumn term and cohort 2 starting in the spring term 2024. The baseline measure for cohort 1 is collected in September 2023 and for cohort 2 it is collected in December 2023. Randomisation was balanced within schools and all pupils at a school were randomised in the same cohort, so the split of intervention and controls pupils will be approximately the same in each cohort. Pupils in both cohorts will be analysed together for the primary analysis, but a subgroup analysis will measure the impact of the intervention in each cohort.

¹ [About ECC - Every Child Counts \(edgehill.ac.uk\)](https://www.edgehill.ac.uk/every-child-counts)
Restricted

This trial is registered at ISRCTN and can be found here:

<https://doi.org/10.1186/ISRCTN14875632>

Design overview

Trial design, including number of arms		Randomised controlled trial with two arms
Unit of randomisation		Within school randomisation at pupil-level
Stratification variables (if applicable)		FSM eligibility status
Primary outcome	variable	Maths attainment
	measure (instrument, scale, source)	Quantitative Reasoning Test, 0-20, (Terezinha Nunes, Peter Bryant, Deborah Evans & Rossana Barros (2015) Assessing Quantitative Reasoning in Young Children, Mathematical Thinking and Learning, 17:2-3, 178-196, DOI: 10.1080/10986065.2015.1016815)
Secondary outcome(s)	variable(s)	NA
	measure(s) (instrument, scale, source)	NA
Baseline for primary outcome	variable	Maths attainment
	measure (instrument, scale, source)	Quantitative Reasoning Test, 0-20, (Terezinha Nunes, Peter Bryant, Deborah Evans & Rossana Barros (2015) Assessing Quantitative Reasoning in Young Children, Mathematical Thinking and Learning, 17:2-3, 178-196, DOI: 10.1080/10986065.2015.1016815)
Baseline for secondary outcome	variable	NA
	measure (instrument, scale, source)	NA

This impact evaluation is conducted as a randomised controlled trial, in which pupils are randomised to the intervention or control within schools. Intervention pupils receive 30 half-hour maths sessions over 15 weeks from teaching assistants that have been trained to deliver 1stClass@Number 1, while control pupils receive normal maths lessons. The primary research question for the evaluation is:

- RQ 1. What is the difference in maths attainment measured by the Quantitative Reasoning Test (Nunes *et al.*, 2015) of pupils receiving the 1stClass@Number 1 intervention in comparison to control pupils who do not receive the intervention?

The secondary research questions are:

- RQ 2.1. What is the difference in maths attainment measured by the Quantitative Reasoning Test of FSM-eligible pupils receiving the 1stClass@Number 1 intervention in comparison to FSM-eligible control pupils who do not receive the intervention?
- RQ 2.2. Is the 1stClass@Number 1 intervention as effective for pupils with Special Education Needs (SEN) as for pupils without SEN as assessed by the Quantitative Reasoning Test?
- RQ 2.3. Is the 1stClass@Number 1 intervention as effective for pupils with lower prior attainment as for pupils with higher prior attainment as assessed by the Quantitative Reasoning Test?

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.14	0.2	0.13	0.16
Pre-test/ post-test correlations	level 1 (pupil)	0.3	0.3	0.3	0.3
	level 2 (class)	-	-	-	-
	level 3 (school)	-	-	-	-
Alpha ²		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Number of schools		211	211	226	226
Number of pupils (per school)	intervention	4	2	4	2.6
	control	4	2	4	2.5
	total	8	4 ³	8	5.1
Number of pupils (all schools)	intervention	844	422	904	577
	control	844	422	893	573
	total	1,688	844	1797	1150

A central question for this evaluation is what impact 1stClass@Number 1 has amongst FSM-eligible pupils. Because of this, the sample size was specified so that the evaluation would be powered for a minimum detectable effect size (MDES) of 0.2 in the FSM-eligible pupil subgroup. The MDES of 0.14 amongst all pupils followed from this sample size specification.

³ Assuming 50% of pupils randomised across all schools are eligible for FSM the average number per school will be 4. However, the sample size calculation does not require exactly 4 FSM-eligible pupils at each school: only the overall proportion matters.

The school and pupil numbers given in the 'protocol' columns of the table above were those predicted at the point of randomisation and were made before recruitment began. The recruitment target for this study was 235 schools. Based on previous NFER evaluations we predicted approximately 10% of recruited schools will not reach the analysis stage. As eight pupils will be randomised per school this is 1,688 pupils randomised in total, split evenly between the intervention and control. Assuming 50% of pupils in total across all schools are eligible for FSM, this would be 844 such pupils randomised. Previous NFER evaluations suggest approximately 15% of randomised pupils will not be included in the primary analysis, which means 1,432 pupils (716 eligible for FSM) will be analysed. Assuming a pre-test/post-test correlation of 0.3, this sample size is sufficient to detect an effect size of 0.2 amongst FSM-eligible pupils (or 0.14 amongst all pupils), with a power of 0.8.

The most recent power calculation, based on the number of schools and pupils actually randomised, is given in the 'randomisation' column of the table above. 226 schools were randomised, which was more than anticipated. The proportion of FSM pupils was 64% (this was the same in both cohorts), which was also higher than anticipated. As a result, the updated MDES at the randomisation stage is slightly lower for all pupils (0.13) and substantially lower for FSM pupils (0.16). We anticipate that the majority of school attrition occurs between recruitment and randomisation, so we did not incorporate the 10% school attrition into these revised MDES calculations (i.e. little or no further school-level attrition assumed).

The pre-test/post-test correlation of 0.3 predicted for this study is lower than for many EEF trials and is based on the previous evaluation (Nunes *et al.*, 2018) of 1stClass@Number 1. The authors of that evaluation observed a pre-test/post-test correlation of 0.29. For the current evaluation, we took the additional step of screening pupils using a different instrument (the SENT-R) to that used for the baseline measurement (the QRT), which may improve this figure. Nevertheless, we made the conservative assumption that the pre-post correlation would be similar to that previously observed.

All sample size calculations were performed in the 'PowerUp' software (Dong and Maynard, 2013) using the sheet 'BIRA2_1c' for block randomised trials with a constant intervention effect across blocks (in this case schools). This calculation does not require estimation of the ICC or the proportion of level two variance explained, which is why these parameters are not estimated above⁴.

Randomisation

The eight pupils selected within each school were randomised in a 1:1 ratio as a block within each school, so that four were assigned to the intervention and four to the control. This was necessary to ensure both an equal randomisation ratio and that the intervention is delivered as intended, in tutoring groups of four. Randomisation occurred in three tranches: two for the first cohort of pupils and one for the second cohort. This did not introduce additional analytical considerations (e.g. the need to include cohort as a stratifier in analysis models), as randomisation is already blocked within each school and all pupils within a school were randomised in the same cohort.

⁴ A further parameter ω must also be estimated if the effect of the intervention is allowed to vary between schools, as described by Hedges and Rhoads (2009). This is not relevant here as the primary analysis does not include random slopes.
Restricted

The randomisation was also stratified by FSM eligibility within each school using FSM data collected from schools⁵. This stratification will maximise statistical power for the FSM subgroup: a 1:1 randomisation is superior to any other allocation ratio in terms of power. It will also help to correct any bias caused by chance imbalance in the FSM variable after randomisation.

The pupil selection process was designed so that exactly eight pupils were randomised at each school. However, in practice there were a small number of schools that were not able to provide eight pupils for randomisation (eight provided seven pupils, one provided five pupils). In these cases, four pupils were randomised to the intervention and the remainder to the control, ensuring that the intervention was delivered as intended to groups of four. This approach has a small chance of introducing a small amount of bias, as it introduces an association between randomisation allocation and any factors causing less than eight pupils to be randomised (e.g. school size). We believe our approach to be the best one available but acknowledge that others could have been taken. For example, allowing all four intervention pupils to participate in sessions but discarding their data at random until there was an equal number of control and intervention pupils analysed at the school. This approach would have ensured against bias but reduced sample size.

Randomisation was carried out by an NFER statistician using R Code, which was stored for reproducibility and transparency. The code is provided in Appendix A. The statistician was not blinded to group allocation.

Primary outcome measure

The primary (and only) outcome for the impact evaluation is QRT score, which is the same outcome as the previous evaluation of 1stClass@Number 1 (Nunes *et al.*, 2018). The QRT will be administered in all schools by NFER test administrators at the end of the intervention period. In the analysis, QRT scores will be treated as a continuous variable, with a range of possible scores between 0 and 20. The same instrument was used as a baseline measure, administered before randomisation. The baseline and follow-up QRT instruments will have 12 of the 20 questions in common.

Analysis

The analysis described in this plan will generally follow the EEF analysis guidance, although there is a notable divergence for the primary analysis model, which is discussed below. All analysis will be conducted using the R software (The R Foundation, 2023). Mixed effects models with a random intercept only will be analysed using R package 'lme4' (Bates *et al.*, 2015). Outputs from the model with a random intervention slope will be obtained using R package 'eefAnalytics' (Uwimpuhwe *et al.*, 2023). The two packages should agree in their output when calculating the same model, but 'eefAnalytics' provides additional functionality that makes certain processes (e.g. calculating effect sizes) more convenient.

All analysis will be conducted on 'complete cases' only (i.e. pupils with no missing analysis variables), although the missing data analysis will investigate the sensitivity of results to this choice. With the exception of the compliance analysis an 'intention-to-treat' approach will be

⁵ For all 'pre-analysis' purposes (including recruitment, pupil selection and stratified randomisation) the FSM variable was created by asking schools which of their pupils were currently eligible for FSM. Restricted

taken, with pupils being analysed according to their intervention/control assignment, regardless of their degree of participation in the intervention.

Primary outcome analysis

The purpose of the primary analysis is to investigate whether the 1stClass@Number 1 intervention improves maths attainment amongst eligible Year 2 pupils.

As this is an effectiveness trial that uses a multisite design, the EEF analysis guidance indicates that the appropriate analysis is a mixed effects model with a random effect of the intervention in addition to a random intercept. This would allow for the possibility that the impact of the intervention varies between schools. However, this would not align with the power calculation used to inform study recruitment, which assumed a constant effect of the intervention between schools. More importantly, there is a possibility that such a model will face technical issues such as lack of convergence, given that there are only 8 pupils randomised per school. This problem is likely to be exacerbated in the subgroup analyses below, which will have as few as one pupil per school. A random effect of the intervention is therefore not included in the primary analysis, although it will be added as a sensitivity check (see Additional analysis section).

In the absence of a random effect of the intervention, the choice remains as to whether to represent schools as a random intercept, leading to a mixed effects model, or as a fixed covariate, leading to a fixed effects model. As this will be (nearly) a balanced design with a 1:1 allocation ratio within schools⁶, the two approaches should produce similar effect estimates and standard errors, although the standard errors for the mixed effects model will be slightly smaller (Feaster, Mikulich-Gilbertson and Brincks, 2011). It was decided that a two-level (pupil and school) linear mixed effects model will be used for the primary analysis. This model includes the baseline QRT measure and FSM stratification variable as covariates:

$$QRT_{ij} = \beta_0 + \beta_1 intervention_{ij} + \beta_2 QRT_BL_{ij} + \beta_3 FSM_{ij} + b_{0j} + \epsilon_{ij}$$

Here QRT_{ij} is the endpoint QRT score of pupil i in school j , and QRT_BL_{ij} is the baseline QRT score for that pupil. In this equation $intervention_{ij}$ is the random allocation indicator (intervention or control) for pupil i in school j and b_{0j} is the random intercept for school j . FSM_{ij} is a binary variable indicating whether the pupil is eligible for FSM⁷.

Secondary outcome analysis

There are no secondary outcomes for this evaluation.

Subgroup analyses

The effect of the intervention amongst pupils eligible for FSM (using the variable 'EVERFSM_6_P') will be determined by repeating the primary analysis model for this subgroup using the model:

⁶ Due to missing data there will not quite be the same number of pupils analysed in each school. The randomisation allocation was also not quite 1:1 at some schools, as described in the randomisation section.

⁷ The NPD variable 'EVERFSM_6_P' is used to measure FSM eligibility in all analysis models. This variable measures if the pupil has been eligible for FSM in the last six years. For this evaluation of Year 2 pupils FSM records do not go back six years, since FSM status is first recorded in the reception year.

$$QRT_{ij} = \beta_0 + \beta_1 intervention_{ij} + \beta_2 QRT_BL_{ij} + b_{0j} + \epsilon_{ij}$$

As the impact of 1stClass@Number 1 for FSM pupils is of particular importance, this evaluation is powered to detect an effect in the FSM subgroup. Subgroup analysis will also be performed to determine the effect of the intervention for pupils with special educational needs (SEN), defined by converting the 'SENprovisionMajor' variable gathered from the NPD into a binary indicator (SEN versus no SEN). This will follow the same specification as the primary analysis, but will only include pupils with SEN.

Another subgroup analysis will also be performed for pupils with lower SENT-R scores (scores from the eligibility test), using a score threshold of 40 and below to define the subgroup. For the primary analysis above, data for pupils in both analysis cohorts will be analysed together. As the two cohorts receive the intervention roughly a term apart, there is a possibility that the effect of the intervention differs between them (e.g. due to them being at different stages in their maths development). To investigate whether the impact of the intervention is the same across the two cohorts we will also perform subgroup analyses which are restricted to cohort 1 and cohort 2 respectively. Again, these analyses will use the primary analysis model, though restricted to the appropriate subgroup.

Additionally, for each potential moderator (variable defining a subgroup) a model with an interaction term will be used to estimate the differential impact of the intervention in the subgroup. In the case of FSM, the differential effect of the intervention for FSM-eligible pupils relative to non-FSM pupils will be investigated by including an interaction term in the primary analysis model:

$$QRT_{ij} = \beta_0 + \beta_1 intervention_{ij} + \beta_2 QRT_BL_{ij} + \beta_3 FSM_{ij} + \beta_4 FSM_{ij} * intervention_{ij} + b_{0j} + \epsilon_{ij}$$

In this equation β_4 quantifies the additional (or lesser) impact of the intervention for ever FSM-eligible pupils. We note that while this trial is powered to detect an intervention effect in the ever FSM subgroup, it may not be powered to test whether there is a different effect for FSM pupils relative to non-FSM pupils (i.e. whether $\beta_4 = 0$). A similar interaction model will also be calculated for each the other three subgroups (SEN pupils, pupils with SENT-R scores below 40, cohort 1 pupils) described above. For example, in the case of pupils with SEN the equation would be:

$$QRT_{ij} = \beta_0 + \beta_1 intervention_{ij} + \beta_2 QRT_BL_{ij} + \beta_3 FSM_{ij} + \beta_4 SEN_{ij} + \beta_5 SEN_{ij} * intervention_{ij} + b_{0j} + \epsilon_{ij}$$

where SEN_{ij} is an indicator for whether pupil i in school j has special educational needs.

Sensitivity check: impact of online training

There were N=16 schools where TAs attended only online rather than face-to-face training. Due to the inclusion of this online training group, we will perform a sensitivity analysis that excludes all pupils from schools that received online training, but otherwise follows the same model specification as the primary outcome:

$$QRT_{ij} = \beta_0 + \beta_1 intervention_{ij} + \beta_2 QRT_BL_{ij} + \beta_3 FSM_{ij} + b_{0j} + \epsilon_{ij}$$

Restricted

The sensitivity analysis will use the same modelling approach as the primary analysis. This approach could also be viewed as a subgroup analysis, but interest lies more in establishing whether the primary analysis impact estimate is robust to the inclusion of the online training group (rather than quantifying the moderating effect of online training *per se*).

Sensitivity check: allowing the impact of the intervention to vary between schools

For the primary analysis the choice was made to analyse this multisite trial using a mixed effects model with random school intercept only, as discussed in the primary analysis section. A limitation of this approach is that it assumes the impact of the intervention is the same for all schools (Feaster, Mikulich-Gilbertson and Brincks, 2011). If the assumption is not met, results may not be generalisable to schools outside the trial. To assess the sensitivity of the primary analysis to this assumption, the following model will be calculated:

$$QRT_{ij} = \beta_0 + \beta_1 intervention_{ij} + \beta_2 QRT_BL_{ij} + \beta_3 FSM_{ij} + b_{0j} + b_{1j} intervention_{ij} + \epsilon_{ij}$$

This adds a random slope b_{1j} to the primary analysis model, representing the effect of the intervention varying from school to school. All other terms are defined in the primary analysis section. This model will be calculated both for all pupils and for the subgroup of FSM eligible pupils, with the FSM indicator removed in the second case. The FSM-only model may face technical issues with estimation due to schools with few FSM pupils, so results will only be shared if meaningful estimates can be obtained. A plot estimating the effect of the intervention for each school⁸ will be included in the report.

IPE regression modelling

At the end of the intervention delivery period a survey will be sent to all TAs taking part in 1stClass@Number 1, which will include items on the TA's experience with teaching maths and their own confidence with the subject. Three measures derived from these items will be included in regression modelling, to investigate the extent to which they moderate the impact of 1stClass@Number 1 on maths attainment for the four pupils linked to each TA.

1. Whether the TA achieved at least a grade C or equivalent in their maths GCSE (binary, yes or no).
2. How many years of experience the TA has teaching maths in KS1 pupils (integer, modelled as a continuous variable).
3. How many years of experience the TA has teaching KS1 pupils in any subject (integer, modelled as a continuous variable).
4. The TA's self-reported confidence in their own maths abilities. 6 Likert scale items relating to aspects of the TA's self-reported confidence with maths (see Appendix B) will be summed to form a 6- 24 scale (modelled as a continuous variable).

These are measures included in the selection guidance for TA delivering 1CN1 and/or we believe may be important moderators for the impact of 1stClass@Number 1 based on the intervention's theory of change. Measure (4) is bespoke to this evaluation and while it will not be validated extensively (e.g. through confirmatory factor analysis), the Cronbach's Alpha of the scale will be reported. Each of these measures will be included as a predictor in a pupil-

⁸ That is, the overall fixed effect of the intervention plus the 'best linear unbiased predictor' of the random effect of the intervention for that school.
Restricted

level linear mixed effects model. These models will be similar to the primary analysis, except that only intervention pupils will be included, and the intervention indicator will be replaced by a variable representing measure (1), (2), (3) or (4) above. In the case of measure (1) the model will be:

$$QRT_{ij} = \beta_0 + \beta_1 TA_GCSE_math_j + \beta_2 QRT_BL_{ij} + \beta_3 FSM_{ij} + b_{0j} + \epsilon_{ij}$$

Here $TA_GCSE_math_j$ is an indicator variable, taking the value 1 if the TA at school j achieved a grade C or above in GCSE maths. The models for measures (2), (3) and (4) will follow similar lines, with $TA_GCSE_math_j$ replaced by a variable for the appropriate measure. For measures (2) and (3), the implied linear relationship between years of TA experience and pupil QRT score may not hold. We will plot this relationship for both measures and convert the measures into (e.g.) a categorical variable if a clear non-linear relationship is observed. Furthermore, as all four measures are likely to be correlated, they will be included simultaneously in a further model:

$$QRT_{ij} = \beta_0 + \beta_1 TA_GCSE_math_j + \beta_2 TA_years_math_teaching_j + \beta_3 TA_years_any_teaching_j + \beta_4 TA_math_confidence_j + \beta_5 QRT_BL_{ij} + \beta_6 FSM_{ij} + b_{0j} + \epsilon_{ij}$$

In this model $TA_years_math_teaching_j$, $TA_years_any_teaching_j$ and $TA_math_confidence_j$ represent measures 2, 3 and 4 respectively for the TA at school j . We will report the correlation between all four measures.

These IPE regression methods have changed since they were described in the study protocol: measure (3) for years of TA experience teaching any subject has been added. This addition was made in response to concerns from Edge Hill University that many TAs have experience teaching KS1 pupils, but not maths specifically. This means measure (2) may have a high number of zeros and lack a sufficient spread of values to investigate a linear relationship. However, as measure (2) is more closely aligned with the intervention it was agreed that both measures (2) and (3) were worth investigating. A further change is that the model which includes all four measures as predictors was added since the protocol. This change was made after the SAP peer review process, when a reviewer noted that these measures are likely to be correlated.

Longitudinal follow-up analyses

There is currently no longitudinal follow-up planned for this evaluation.

Imbalance at baseline and sample representativeness

As pupils are randomised within schools there will not be a comparison between intervention and control schools. Instead, the following characteristics will be compared between the sample of schools randomised and all primary schools in England:

School level

- Proportion of pupils eligible for FSM in 2023/24
- Whether the school is urban or rural
- School type (academy, maintained or independent) in 2023/24
- Most recent overall Ofsted rating in 2023/24

Restricted

- Proportion of pupils meeting the expected standard in their KS2 maths exam in 2023/24

This will help assess how representative the sampled schools are of English primary schools in general.

To assess imbalance in baseline characteristics amongst pupils a table will be produced describing the characteristics of the control and intervention groups after randomisation. Pupil-level characteristics will be:

- FSM eligibility
- Whether pupils have SEN (using 'SENprovisionMajor' from the NPD)
- Baseline QRT score
- SENT-R score

Categorical variables will be described in terms of counts and proportions, while means and standard deviations will be given for continuous variables. School-level variables will be obtained from NFER's own database of school characteristics. Pupil-level variables will be obtained from the NPD (FSM) or collected directly from schools (baseline QRT and SENT-R scores). The difference between the baseline QRT scores and SENT-R scores in the intervention and control groups will be expressed as a Hedges' g effect size.

Missing data

The missing data analysis will be performed for the primary analysis model (i.e. all pupils) and separately for the subgroup of FSM-eligible pupils. Missing data analysis is described below for the primary analysis model, but the same procedure will be conducted for the FSM subgroup.

The number and proportion of missing cases in the primary analysis will be reported. If this is less than five percent then the potential for bias in a complete case analysis will be considered minimal and no further missing data analysis will take place, following EEF's statistical guidance. Otherwise, the number and proportion of missing values for each variable included in the primary analysis models will be reported. There is potential for missing data in the baseline and follow-up QRT scores, as well as the FSM indicator.

Assuming both baseline and follow-up QRT scores have missing data, a mixed effects logistic regression with two levels (pupil and school) will be run for each; the outcomes will be the logit probability of the respective QRT scores being missing. All other variables from the maths primary analysis model will be included as covariates, together with these additional variables that may be associated with missingness:

- Pupil SEN in 2023/24
- Pupil SENT-R score
- Pupil's school type (academy, maintained or independent) in 2022/23
- Proportion of pupils eligible for FSM in 2023/24 at the pupil's school

The additional variables which demonstrate an association with missingness in the QRT follow-up score, as indicated by a p-value below 0.05, will be included as covariates in the primary analysis model. If re-running the primary analysis with these extra covariates included alters the substantive interpretation of the intervention effect, then the QRT outcome may be

'missing at random' conditional on the inclusion of those covariates. This would then need to be discussed in the report.

If missingness of the baseline QRT score is associated with any of the additional variables from the above list, a sensitivity analysis will be performed using multiple imputation. Missing data will be imputed using a two-level normal model with the '2l.imer' function in the R package 'mice' (Buuren and Groothuis-Oudshoorn, 2011). The imputation model will include all variables from the primary analysis, together with any additional variables associated with missingness of baseline or follow-up maths scores. Ten datasets will be generated, each using ten iterations, and the primary analysis model will then be re-run on each. Estimates from each model will be pooled into a single set of estimates and standard errors that will be compared to the results of the original analysis.

Where follow-up QRT scores are unavailable due to school withdrawals we will seek to establish, as far as possible, the reason(s) for these withdrawals, and summarise this information in the report.

Compliance

Three compliance measures will be used for this evaluation, all of which are binary: a pupil either meets the compliance condition(s) or they do not. All three measures are at least partially based on the pupil's attendance in the small group sessions:

- (1) Pupil attended at least 30 sessions in total.
- (2) Pupil attended at least 5/6 sessions in each of the five topic areas (implying at least 25 sessions attended in total).
- (3) Pupil attended at least 5/6 sessions in each of the five topic areas, as for (2) above. Additionally the pupil's TA attended all 5 of the initial training sessions in some form (catch-up sessions, online or otherwise, are included⁹).

The impact of each compliance measure described above will be modelled for all pupils and for the FSM-eligible subgroup separately, resulting in a total of six models. For each of these models an instrumental variable analysis will be performed, using two-stage least squares methods (Angrist and Imbens, 1995) to estimate the effect of compliance with the intervention on QRT follow-up scores. For the first stage the compliance indicator will be regressed on the intervention indicator, together with covariates from the primary analysis model (baseline QRT score, FSM status). This first stage linear regression model will be:

$$\text{compliance}_{ij} = \beta_0 + \beta_1 \text{intervention}_{ij} + \beta_2 \text{QRT_BL}_{ij} + \beta_3 \text{FSM}_{ij} + \epsilon_{ij}$$

In each model the compliance indicator, compliance_{ij} , will be defined based on one of measures (1)-(3) above and will generally¹⁰ take the value zero for control pupils. For the second stage QRT follow-up scores are regressed on each pupil's predicted compliance value $\hat{\text{compliance}}_{ij}$ obtained from the first stage, in the following linear regression model:

⁹ Aspects of fidelity such as whether sessions were attended in person will still be recorded and presented descriptively as part of the IPE analysis but will not be included in these compliance measures.

¹⁰ As this trial is randomised within schools there is a small possibility of control pupils taking part in the intervention, despite efforts to prevent this. In such a case the control pupil would have a value of 1 for their compliance indicator if they took part in 30 sessions (for the first compliance measure).
Restricted

$$QRT_{ij} = \beta_0 + \beta_1 \widehat{compliance}_{ij} + \beta_2 QRT_BL_{ij} + \beta_3 FSM_{ij} + \epsilon_{ij}$$

The coefficient for predicted compliance β_1 in this second stage is the CACE (complier average causal effect) estimate for the effect of compliance on QRT follow-up scores.

Results from both regression stages will be reported. All instrumental variable analysis will be performed using the R package 'ivreg' (Fox *et al.*, 2021). These models do not include school-level random effects, so instead cluster-robust standard errors will be calculated using the R package 'sandwich' (Zeileis, 2006; Zeileis, Köll and Graham, 2020) (Zeileis, 2006; Zeileis, Köll, and Graham, 2020).

Intra-cluster correlations (ICCs)

The ICC for the primary outcome model will be calculated as the proportion of follow-up QRT score variance attributable to level 2 (between-school) variation.:

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

Here σ_B^2 and σ_W^2 are the between-school and within-school variation, which can be extracted directly from a mixed effects regression fitted by the 'lme4' package. The ICC will be calculated twice, once for the primary analysis model and once for an empty model (one with no covariates). These will be reported even though the ICC was not used in the sample size calculation for this evaluation.

Effect size calculation

For the model with a random effect of the intervention specified in the 'Additional analysis' section, the effect size and its variance will be calculated as described by (Singh *et al.*, 2022).

Impact estimates from all models described above will be presented as an effect size:

$$ES = \frac{\hat{\beta}_1}{\sqrt{\sigma_B^2 + \sigma_W^2}}$$

$\hat{\beta}_1$ is the estimated intervention or compliance coefficient from the appropriate model (conditional on covariates), while σ_B^2 and σ_W^2 are the between-school and within-school variance from an empty model. To obtain a 95% confidence interval for this effect size, a confidence interval for $\hat{\beta}_1$ will first be calculated by adding or subtracting 1.96 times its standard error. The end points of this confidence interval will then be divided by the denominator of the above effect size formula. One exception to this approach is when calculating a confidence interval for the effect size estimated from the random slope mixed effects model specified in the additional analysis section. In this case the variance of the effect size is calculated as:

$$Var(ES) = \frac{1}{M\sigma_T^2} \left(\sigma_E^2 + \frac{2\sigma_W^2}{n} \right) + \frac{ES^2}{4M(n-1)}$$

σ_T^2 is the total outcome variation and σ_E^2 is variation due to the differing effect of the intervention across schools. M and n are the number of schools and number of pupils per school respectively. A confidence interval is then obtained by adding or subtracting $1.96 \times \sqrt{Var(ES)}$ from the effect size. A derivation for the above variance, as well as arguments for using it to calculate confidence intervals (instead of the simpler approach above), is given by Singh *et al.*, 2022.¹¹

¹¹ It may be necessary to use the slightly more complicated formula for the effect size variance in the case of unbalanced data, which is also given by Singh *et al.* (2022)
Restricted

References

- Angrist, J.D. and Imbens, G.W. (1995) 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association*, 90(430), pp. 431–442. Available at: <https://doi.org/10.1080/01621459.1995.10476535>.
- Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. Available at: <https://doi.org/10.18637/jss.v067.i01>.
- Buuren, S. van and Groothuis-Oudshoorn, K. (2011) 'mice: multivariate imputation by chained equations in R', *Journal of Statistical Software*, 45, pp. 1–67. Available at: <https://doi.org/10.18637/jss.v045.i03>.
- Dong, N. and Maynard, R. (2013) 'PowerUp!: a tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies', *Journal of Research on Educational Effectiveness*, 6(1), pp. 24–67. Available at: <https://doi.org/10.1080/19345747.2012.673143>.
- Feaster, D.J., Mikulich-Gilbertson, S. and Brincks, A.M. (2011) 'Modeling site effects in the design and analysis of multi-site trials', *The American Journal of Drug and Alcohol Abuse*, 37(5), pp. 383–391. Available at: <https://doi.org/10.3109/00952990.2011.600386>.
- Fox, J., Kleiber, C., Zeileis, A. and Kuschnig, N. (2021) 'ivreg: Instrumental-Variables Regression by "2SLS", "2SM", or "2SMM", with Diagnostics'. Available at: <https://cran.r-project.org/web/packages/ivreg/index.html> (Accessed: 14 November 2023).
- Hedges, L.V. and Rhoads, C. (2009) *Power analysis in education research*. Available at: <https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf> (Accessed: 14 November 2023).
- Nunes, T., Barros, R., Evangelou, M., Strand, S., Mathers, S. and Sanders-Ellis, D. (2018) *1stClass@Number: Evaluation report and executive summary*. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/projects/1stClass@Number_evaluation_report.pdf?v=1630925665.
- Nunes, T., Bryant, P., Evans, D. and Barros, R. (2015) 'Assessing quantitative reasoning in young children', *Mathematical Thinking and Learning*, 17(2–3), pp. 178–196. Available at: <https://doi.org/10.1080/10986065.2015.1016815>.
- Singh, A., Uwimpuhwe, G., Li, M., Einbeck, J., Higgins, S. and Kasim, A. (2022) 'Multisite educational trials: estimating the effect size and its confidence intervals', *International Journal of Research & Method in Education*, 45(1), pp. 18–38. Available at: <https://doi.org/10.1080/1743727X.2021.1882416>.
- The R Foundation (2023) *R: the R project for statistical computing*. Available at: <https://www.r-project.org/> (Accessed: 14 November 2023).
- Uwimpuhwe, G., Singh, A., Vallis, D., Higgins, S., Xiao, Z. and Kasim, E.D.T. and A. (2023) 'eefAnalytics: robust analytical methods for evaluating educational interventions using randomised controlled trials designs'. Available at: <https://cran.r-project.org/web/packages/eefAnalytics/index.html> (Accessed: 15 November 2023).
- Zeileis, A. (2006) 'Object-oriented computation of sandwich estimators', *Journal of Statistical Software*, 16, pp. 1–16. Available at: <https://doi.org/10.18637/jss.v016.i09>.
- Zeileis, A., Köll, S. and Graham, N. (2020) 'Various Versatile Variances: an object-oriented implementation of clustered covariances in R', *Journal of Statistical Software*, 95, pp. 1–36. Available at: <https://doi.org/10.18637/jss.v095.i01>.

Appendix A: randomisation code

```
#### Stratified randomisation code for cohort 1, tranche 1
#### Load data
df=read.xlsx(...)

#### set random seeds using current time
time_now=17.56 # What time is it now? (hh.mm)(e.g: 11.35)

#### function that takes an NFER No (school ID) and returns all data for that school
#### with random allocation column completed
func=function(x, seed=time_now){
  aux=df[df$School.NFER.No==x,]
  aux$original_order=1:nrow(aux)

  seed=100*trunc(seed)+100*(seed-trunc(seed))+x^2
  set.seed(seed)

  aux=aux[sample(1:nrow(aux), nrow(aux)),]
  aux=aux[order(aux$Pupil.FSM.eligibility),]
  if(sample(1:2, 1)==1){aux=aux[order(aux$Pupil.FSM.eligibility, decreasing=T),]}

  groups=rep(c("Intervention", "Comparison"), 4)
  if(nrow(aux)==7){groups=groups[1:7]}
  aux$Randomisation.allocation=groups

  aux=aux[order(aux$original_order),]
  aux=aux[,1:12]
  return(aux)
}
out=lapply(unique(df$School.NFER.No), func)
out=do.call(rbind, out)
```

Appendix B: survey items measuring TA's confidence in their own maths abilities

D		YOUR CONFIDENCE IN YOUR MATHS ABILITIES				
<i>Please select one response for each statement.</i>		1	2	3	4	5
		Strongly disagree	Disagree	Agree	Strongly agree	I don't want to say
D.1	My mental maths abilities are good					
D.2	I am confident solving maths problems					
D.3	I am confident using maths in my daily life (for example estimating a bill; calculating a 20% off deal)					
D.4	I believe I'll be able to understand the content delivered in maths CPD sessions					
D.5	I believe I am good at maths					
D.6	Overall, I feel confident in my own maths abilities					