Randomised Controlled Trial: Evaluation of the White Rose Maths Reception Jigsaw Statistical Analysis Plan

Evaluator: NFER

Principal investigator: Ben Styles

PROJECT TITLE	Randomised Controlled Trial: Evaluation of the White Rose Maths Reception Jigsaw
DEVELOPER (INSTITUTION)	White Rose Maths (Trinity MAT)
EVALUATOR (INSTITUTION)	National Foundation for Educational Research (NFER)
PRINCIPAL INVESTIGATORS	Ben Styles
PROTOCOL AUTHORS	Helen Poet, Ben Styles, Guido Miani, Joana Andrade, Anusha Ghosh
SAP AUTHORS	Joana Andrade, Ben Styles, and Chris Morton
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Reception pupils (age 4-5)
NUMBER OF SCHOOLS	138
NUMBER OF PUPILS	2689 (approximately 19 pupils per school)
PRIMARY OUTCOME MEASURE AND SOURCE	Maths attainment (New Progress in Understanding Mathematics – New PUMA)
SECONDARY OUTCOME MEASURE AND SOURCE	(1) Maths attainment (G11 and G12 elements of EYFSP) (2) Practitioners' confidence in teaching maths (bespoke survey, adapted from Chen et al., 2014)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.1 [*latest*]	April 2022	The G11/G12 EYFSP goals can only take the values 'Emerging' or 'Expected' but not 'Exceeding' in 2021/22. Text updated accordingly. Due to concerns that some PUMA items are not aligned with the areas of maths the Reception Jigsaw focuses on, a secondary outcome has been added. This is made up of 20/30 items from the new PUMA.
1.0 [*original*]	November 2021	*N/A*

Table of contents

Introduction

Research suggests that high quality early numeracy education in the early years can have long lasting beneficial effects and the early numeracy strand of the EEF Early Years Teaching and Learning Toolkit highlights the importance of professional development in supporting early numeracy approaches.

Teachers are now required to complete the Early Years Foundation Stage Profile (EYFSP) at the end of the reception year by reviewing the evidence gathered on their pupils against each Early Learning Goal (ELG) and evaluate if each child is meeting the expected level of development or not yet reaching the level (emerging)[1].

The new ELGs became statutory in 2021, with schools able to opt in to use them from September 2020. However, a 2019 evaluation[2] of the pilot of the draft ELGs revealed that although most pilot schools viewed the draft ELGs as positive, there was no consensus about whether reception children would be better prepared for Key Stage 1 as a result. There was some confusion in schools on whether changes to the ELGs constituted a change to the reception curriculum and further guidance accompanying the ELGs was called for.

Part of the Trinity Multi Academy Trust, West Yorkshire based White Rose Maths (WRM) provides a range of consultancy services, schemes of work, and maths resources for primary schools, including the delivery of training packages.

Schools have fed back to WRM that reception teachers often miss out on training opportunities and also the need for a training package addressing the changes to the reception year brought up by the introduction of the ELGs. More broadly, it is recognised that early years' teachers frequently require support to develop their subject knowledge, knowledge of children's development, and pedagogical knowledge. As a result of this, WRM have recently developed the Reception Jigsaw which has been piloted in a small number of schools. There has been no formal evaluation of the Reception Jigsaw before, but it is founded in strong principles of early years' maths teaching and the modules are designed to be evidence based.

The independent evaluation of the Reception Jigsaw is necessary to assess its impact on pupils' maths attainment and practitioners' confidence in teaching maths to reception-age children. The impact evaluation comprises a randomised controlled trial, with school-level randomisation as the intervention is delivered to the whole of reception in each school. The trial will use a teacher-completed baseline assessment (checklist) based on their observation of each sampled child[3]. The follow-up test will use the New Progress in Understanding Mathematics Assessment (PUMA) for reception which will be administered by NFER test administrators. The New PUMA is produced by Rising Stars (part of Hodder).

As secondary outcomes relating to teachers' confidence and self-actualization, we will assess their confidence in teaching mathematics and confidence in their own maths ability by means of a survey administered to early years practitioners at baseline and at the end of the reception year. The instruments used to measure confidence in teaching and teachers' maths ability will

[1] For more information see the EYFSP 2022 handbook: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1024319/Early_years_foundation_stage_profile_handbook_2022.pdf.

[2] https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Early_Years_Foundation_Stage_Profile_(EYFSP)_Reforms.pdf

[3] The Reception Baseline Assessment was completed in schools in September/October 2020 but the data from it is not going to be available for researchers in the NPD (see RBA privacy notices: https://www.nfer.ac.uk/for-schools/participate-in-research/information-about-the-reception-baselineassessment/).

be adapted from the questionnaire and instruments developed by Chen et al. (2014) and used recently in EEF's Maths Champions trial (Robinson-Smith et al., 2018).

The primary research question of the evaluation is:

RQ1: What is the impact of the White Rose Maths Reception Jigsaw on reception children's maths attainment, as measured by New PUMA tests at the end of the reception year, compared to 'business as usual'?

The secondary research questions are:

RQ2: What is the impact of the White Rose Maths Reception Jigsaw on pupils' maths attainment as measured by New PUMA tests at the end of Year 1, compared to 'business as usual'?

RQ3: What is the impact of the White Rose Maths Reception Jigsaw on pupils' maths attainment as measure by the maths elements of the Early Years Foundation Stage Profile (EYFSP), compared to 'business as usual'?

RQ4: What is the impact of the White Rose Maths Reception Jigsaw on practitioners' confidence to teach maths to children in their reception year, compared to 'business as usual'?

RQ5: What is the impact of the White Rose Maths Reception Jigsaw on practitioners' confidence in their maths ability, compared to 'business as usual'?

RQ6: Are effects of the White Rose Maths Reception Jigsaw on maths attainment different for pupils eligible for free school meals (FSM) in the reception year (as per RQ1)?

Design overview

Trial design, including number of arms		Two-arm, cluster randomised
Unit of randomisation		School
Stratification variables (if applicable)		Geographic area (Yorkshire/Essex), Mastering Number participation
Primary outcome	variable	Maths attainment
	measure (instrument, scale, source)	New Progress in Understanding Mathematics Assessment (New PUMA), [0-30], Rising Stars (part of Hodder Education)
Secondary outcome(s)	variable(s)	1) Binary measure of reaching the expected level in both G11 (number) and G12 (numerical patterns) Early Learning Goals (ELG) elements[4] of EYFSP, [0/1], NPD 2) Practitioner confidence in teaching maths, [11-55], bespoke survey, adapted from Chen et al., 2014

[4] See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/790580/EYFSP_Handbook_2019.pdf.

		3) Practitioner confidence in their maths ability, [9-45], bespoke survey, adapted from Chen et al., 2014 4) New Progress in Understanding Mathematics Assessment (New PUMA), [0-20], (selected 20 of 30 items, bespoke adaption to instrument)
	measure(s) (instrument, scale, source)	1) Binary measure of reaching the expected level in both G11 (number) and G12 (numerical patterns) Early Learning Goals (ELG) elements[5] of EYFSP, [0/1], NPD 2) Practitioner confidence in teaching maths, [11-55], bespoke survey, adapted from Chen et al., 2014 3) Practitioner confidence in their maths ability, [9-45], bespoke survey, adapted from Chen et al., 2014 4) New Progress in Understanding Mathematics Assessment (New PUMA), [0-20], (selected 20 of 30 items, bespoke adaption to instrument)
Baseline for primary outcome and secondary outcomes (1) and (4)	**variable**	1) Emerging numeracy
	measure (instrument, scale, source)	1) Emerging numeracy checklist, [0-20], bespoke checklist adapted from Early Years Outcomes and Early Learning Goals
Baseline for secondary outcomes (2 and 3)	**variables**	2) Practitioners' confidence to teach maths to reception pupils 3) Practitioners' confidence in their maths ability
	measure (instrument, scale, source)	2) Practitioner confidence in teaching maths, [11-55], bespoke survey, adapted from Chen et al., 2014 3) Practitioner confidence in their maths ability, [9-45], bespoke survey, adapted from Chen et al., 2014

School-level pupil sampling

Following the rationale that considerable cost savings could be made by only testing a random sample of 20 pupils per school, the number of small-group tests that an NFER test administrator can accomplish in a day's visit, a random school-level sampling of pupils was introduced in the trial design. The implications of this choice of design in terms of power and MDEs is discussed in the Sample Size Calculations overview section below.

As evaluating the impact of educational interventions on children from deprived backgrounds is one of the main objectives of EEF trials it made sense, when designing the trial, to establish a sampling frame of high FSM schools to increase the power of FSM sub-group analyses. During the design stage of the trial the evaluation team assumed that 25 percent of pupils

[5] See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/790580/EYFSP_Handbook_2019.pdf.

participating in the trial would be eligible for FSM[6] which lead to a rough estimation of a cluster size of five FSM pupils per school.

However, in subsequent analysis based on the historical figures for proportion of FSM pupils in the schools that signed up for the trial the trial's statistician concluded that the average proportion of FSM eligible pupils was overestimated and that to ensure a cluster size of five FSM pupils per school it would be necessary to randomly select 28 rather than 20 pupils per school.

Based on the updated calculation, the evaluation team adopted the following sampling scheme:

1. We randomly sampled an initial sample of 28 pupils per school, the randomisation being stratified to reflect the reported FSM proportion of the individual schools (the proportion of reported FSM was calculated out of Y/N responses for each individual school – ignoring missing data).

2. In each school the pupils explicitly identified as eligible for free school meals in the initial sample were automatically included in the sample of 20 pupils selected for the trial. The remaining pupils in the initial sample were then allocated to the trial sample or assigned to a group of replacements of up to eight children per school.

3. The pupils in the replacement groups were treated as standbys to be included in the trial sample in case the pupils in the trial samples of their schools could not take part on the trial or were withdrawn from it.

The R code used to draw the pupil is included in this statistical analysis plan as an appendix. All the calculations were performed in R 4.0.3.

Randomisation

As originally designed by the evaluation team, the trial's randomisation consisted of a stratified randomisation with geographical area as the sole stratifier. However, during the design stage of the trial the team was made aware that the Maths Hubs network[7] was recruiting schools for the Mastering Number[8] programme (MN), an initiative launched by the Department for Education and aiming to provide training to teachers in order to promote "firm foundations in the development of good number sense for all children from reception through to Year 1 and Year 2".

Consulting with the participating schools, the evaluation team concluded that a number of schools that were recruited for the Maths Reception Jigsaw trial had also signed up for Mastering Number. Given the similarities between the two programmes in terms of objectives and methodology, it is advisable to ensure that the participation on Mastering Number is balanced across intervention and control schools.

On the one hand, it is necessary to control for possible instances of contamination (teachers in control schools participating in Mastering Number receiving training in similar moulds to teachers in intervention schools), and, on the other, it is also necessary to account for

[6] The figure of 25 percent of pupils eligible for FSM is still considerably higher than the national proportion that stands at 20.8 percent as of June 2021, having gone up from 17.3 percent in 2020 (https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics).

[7] See https://www.ncetm.org.uk/maths-hubs/.

[8] See https://www.ncetm.org.uk/maths-hubs-projects/mastering-number/.

confounding or even synergies between the two interventions (the average effects detected in trial may be the result of the Mastering Number interventions implemented in intervention schools rather than those of the Reception Jigsaw).

Bearing in mind the circumstances described above, the evaluation team decided to update the stratified randomisation design described in the protocol to a stratified randomisation that comprised four strata: Yorkshire schools that signed up for the Mastering Number programme, Yorkshire schools that did not sign up for the Mastering Number programme, Essex schools that signed up for the Mastering Number programme, and Essex schools that did not sign up for the Mastering Number programme.

As per protocol, a total of 55 schools were randomised as intervention schools, with the remaining schools being assigned, by default, to the control group. The number of intervention schools to be randomly selected out of each stratum was determined so that the representativeness of each block in the samples of intervention and control schools mirrors that of the overall sample of schools participating in the trial:

Table 1: Stratified randomisation

Number of Schools (%)	Yorkshire + Signed up for MN	Yorkshire + Did not sign up for MN	Essex + Signed up for MN	Essex + Did not sign up for MN	Total
Participating Schools	39 (28.3%)	75 (54.3%)	8 (5.8%)	16 (11.6%)	138 (100%)
Intervention Schools	16 (29.1%)	30 (54.5%)	3 (5.5%)	6 (10.9%)	55 (100%)
Control Schools	23 (27.7%)	45 (54.2%)	5 (6.0%)	10 (12.0%)	83 (100%)

The R code used to perform the stratified randomisation is included in this statistical analysis plan as an appendix. All the calculations were performed in R 4.0.3.

Baseline and outcome measures

Baseline measures

The baseline measure for the Primary and Secondary I analyses will be derived from a bespoke 20 item-checklist to be completed by reception teachers. The checklist, in Appendix C, was created by the evaluation team and was based on the Early Years Outcomes and the Early Learning Goals. All the items in the checklist have as possible responses “cannot do with support”, “can do with support”, and “can do independently” that will be recoded as numeric values (1, 2, 3). The recoded numeric variables will be used to derive the emerging numeracy scores.

The baseline measure will be computed by adding the 20 checklist variables. To ensure that the emerging numeracy measure is adequately reliable we will calculate the Cronbach’s alpha index for the group of 20 variables. If this reliability index is lower than 0.7, we will identify items that reduce the reliability of the ensemble and re-compute the instrument excluding the problematic items.

The baseline measure for the Secondary II analysis is the same as the outcome measure (see section below).

Primary outcome measure

As a follow-up test we will use the reception version (summer) of the New Progress in Understanding Mathematics Assessment (New PUMA[9]), a standardized test developed by Rising Starts (part of Holder Education). The New PUMA is the new edition of the 2014 PUMA released in 2020. There is no change in the content between the New PUMA and the previous edition at reception level.

Secondary outcome I measure

As a follow up measure for the first secondary analysis we will be using an aggregation of the two mathematics Early Learning Goals in the early years foundation stage profile (EYFSP), using data from the national pupil database (NPD).

The latest version of the early years profile has only two maths goals, G11 (Number) and G12 (Numerical Patterns); a pupil is recorded as either 'Emerging' (below the expected level) or 'Expected' (meeting the expected level) in each of these. We will compute EYFSP as a binary measure, taking the value 1 if a pupil meets the expected level for both maths goals, and 0 otherwise.

Secondary outcome II and III measures

Practitioners' confidence to teach mathematics to reception pupils and confidence in their own maths ability will be measured both at baseline and follow up using instruments adapted from Chen et al. (2014). The survey will be based on the survey developed by Chen et al. (2014), with minor adaptations to reflect the setting (mainly related to terminology, i.e. to say reception/school rather than nursery).

Secondary outcome IV measure

As for the primary outcome, this secondary outcome will use pupils' scores from the New PUMA, but in this case only 20 of the 30 items will contribute to the total score. It was decided after consultation with White Rose Maths that 10 items did not align with the current reception maths curriculum, upon which the Reception Jigsaw training is based. These 10 items, which were related to totalling money or using teen numbers, were not included in the secondary measure. As this is a bespoke scale that has not been validated in previous studies, its inclusion in the analysis is conditional on it demonstrating a Cronbach's Alpha of 0.7 or above.

Sample size calculations overview

	Protocol		Randomisation	
	OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)	0.22	0.27	0.21	0.26

[9] https://www.risingstars-uk.com/subjects/assessment/rising-stars-puma

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Pre-test/ post-test correlations	level 1 (pupil)	0.40	0.40	0.40	0.40
Intracluster correlations (ICCs)	level 2 (school)	0.17	0.17	0.17	0.17
Alpha		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
Average cluster size		20	5	19.49	5*
Number of schools	intervention	55	55	55	55
	control	74	74	83	83
	total	129	129	138	138
Number of pupils	intervention	1100	275	1076	275*
	control	1480	370	1613	415*
	total	2580	645	2689	690*

Notes: * Estimated

As described in the *School-Level Pupil Sampling* section above, the evaluation team aimed to establish a sampling frame with an average of five children eligible for free school meals per school. Based on the parameters reported in EEF's Maths Champions evaluation (Robinson-Smith et al., 2018), which used CEM's ASPECTS assessment, an ICC of 0.17 is considered appropriate. Using the emerging numeracy checklist at baseline will, in general, result in lower pre-post correlations than a properly developed test so we have assumed a correlation of 0.4.

Under the assumption that 20 pupils would be tested in each participating school and after discussions with the developer concerning their capacity to deliver to more than 50 schools, a sample size of 55 intervention schools and 74 control schools was agreed. This design had 80 per cent power to detect an effect size of 0.22.

The trial recruited around 5 per cent extra schools to allow for dropout at baseline before randomisation. This lead to a total of 138 schools being randomised, 55 to intervention and the remainder to control and the trial being powered to detect an effect size as small as 0.21.

As discussed in the *Sampling* section above, we expect that the sampling design adopted for this trial will yield an average cluster size of five FSM children per school, which corresponds to a sub-group analysis that includes all the 138 randomised schools being powered to detect a MDES of 0.26.

Analysis

The primary and both secondary analyses will follow EEF's 2018 guidelines and assume intention to treat (ITT).

Primary outcome analysis

A multilevel random intercepts model with two levels (school and pupil) will be used to account for cluster randomisation. The primary analysis will investigate whether reception teachers having received training and support under the Reception Jigsaw programme had an effect on their pupils' maths attainment. This will be determined by fitting a model with maths attainment at follow-up, as measured by Progress in Understanding Mathematics (PUMA) scores, as the dependent variable.

To control for prior ability, pupil-level emerging numeracy (EN) scores assessed at baseline will be included in the model as a covariate. The model will also contain a dummy variable for stratum (school region versus school participation in the Mastering Number programme) to reflect the stratified randomisation.

The two level random intercepts model is given by:

$$\mathrm{PUMA}_{ij} = \beta_0 + u_{0j} + \beta_1 \mathrm{intervention}_j + \beta_2 \mathrm{baseline\ EN}_{ij} + \beta_3 \mathrm{stratum}_j + \epsilon_{ij}$$

Where PUMA_{ij} is the PUMA score of pupil i in school j, u_{0j} is the random intercept in school j, $\mathrm{intervention}_j$ is the school-level intervention/control dummy variable, $\mathrm{baseline\ EN}_{ij}$ is the baseline EN score of pupil i in school j, and $\mathrm{stratum}_j$ is a dummy variable for the randomisation stratum of school j.

The model will be run in R (version 4.1.2) using the package 'nlme'[10].

Secondary outcome analysis I

The first secondary outcome analysis will evaluate if their teachers' participation in the Reception Jigsaw Programme had an effect on pupils meeting the two early learning goals relating to mathematics. For this purpose we will fit a multilevel logistic regression model whose dependent variable is the EYFSP measure described in the *Secondary outcome I measure* subsection above**.**

The two level random intercepts regression model/logistic regression is given by:

$$\mathrm{EYFSP}_{ij} = \beta_0 + u_{0j} + \beta_1 \mathrm{intervention}_j + \beta_2 \mathrm{baseline\ EN}_{ij} + \beta_3 \mathrm{stratum}_j$$

Where the dependent variable EYFSP_{ij} is the binary indicator (1 if meeting expected level in both goals, 0 otherwise) for pupil i in school j, β_{0j} is the random intercept in school j and $\mathrm{intervention}_j$ is the school-level intervention/control dummy variable. $\mathrm{baseline\ EN}_{ij}$ is the baseline EN score of pupil i in school j, and $\mathrm{stratum}_j$ is a dummy variable for the randomisation stratum of school j.

The analysis will be run in R (version 4.1.2), either using the package 'lme4' (Bates et al., 2015), for a multilevel logistic regression model.

Secondary outcome analysis II and III

The second secondary outcome analyses will assess if teachers' confidence to teach maths to reception pupils and confidence in their own maths ability are affected by their participation in the Reception Jigsaw Programme. For this effect models of practitioners' confidence in

[10] Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2021). _nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-153, <URL: https://CRAN.R-project.org/package=nlme>.

teaching maths and confidence in their maths ability at follow up will be fitted. Both aspects of practitioners' confidence will be measured at baseline and follow up via the adapted bespoke survey described in the secondary outcomes analysis section above. In accordance with the EEF's 2018 directives the analysis models will include a baseline measure of practitioners' confidence and, taking into account the clustered randomisation design, will be run as a two-level (teacher and school) random intercepts models.

The two level random intercepts model for the secondary outcome analysis II is given by:

$$\mathrm{PCTM_{ij}} = \beta_0 + \mathrm{u_{0j}} + \beta_1 \mathrm{intervention_j} + \ \beta_2 \mathrm{baseline\ PCTM}_{\mathrm{\ ij}} + \beta_3 \mathrm{stratum_j} + \epsilon_{\mathrm{ij}}$$

Where $\mathrm{PCTM_{ij}}$ and $\mathrm{baseline\ PCTM}_{\mathrm{\ ij}}$ are, respectively, the practitioners' confidence in teaching mathematics scores derived from the survey taken by teacher i in school j at baseline and follow-up.

And the two level random intercepts model for the secondary outcome analysis III by:

$$\mathrm{PCMA_{ij}} = \beta_0 + \mathrm{u_{0j}} + \beta_1 \mathrm{intervention_j} + \ \beta_2 \mathrm{baseline\ PCMA}_{\mathrm{\ ij}} + \beta_3 \mathrm{stratum_j} + \epsilon_{\mathrm{ij}}$$

Where $\mathrm{PCMA_{ij}}$ and $\mathrm{baseline\ PCMA}_{\mathrm{\ ij}}$ are, respectively, the scores derived from the practitioners' confidence in teaching mathematics survey taken by teacher i in school j at baseline and follow-up.

$\mathrm{u_{0j}}$ is the random intercept in school j, $\mathrm{intervention_j}$ is the school-level intervention/control dummy variable, and $\mathrm{stratum_j}$ is a dummy variable for the randomisation stratum of school j.

The model will be run in R (version 4.1.2) using the package 'nlme'.

Secondary outcome analysis IV

The model used for secondary outcome IV will be identical to the primary analysis model in all respects, except that only 20 selected PUMA items will be summed to calculate the dependent variable, rather than all of 30 of them.

Subgroup analyses

As specified in the protocol, a subgroup analysis will be run to investigate possible differential effects of the Jigsaw Intervention on the attainment of children eligible for free school meals. EVERFSM6 collected from the National Pupil Database (NPD) will be used as the identifier for the main subgroup analysis.

As discussed in the *Randomisation* section above Mastering Number, a programme with similar characteristics to the Reception Jigsaw is being implemented in parallel to the trial in 47 of the 138 participating schools. To better inform the interpretation of the trial's results and investigate the existence of interaction and confounding effects a subgroup analysis considering school participation in Mastering Number.

The analyses will be approached in two distinct ways: by running models with interaction terms (i.e. models that include both the subgroup indicator and the product of the subgroup indicator and randomisation group), and by running separate primary outcome models on:

1) Just the FSM eligible pupils
2) Schools that signed up for the Mastering Number and on schools that didn't sign up for the alternative programme.

Both approaches conform to the EEF 2018 guidelines.

The multilevel level random intercepts model with interaction terms for the FSM subgroup analysis will be given by:

$$\mathrm{PUMA}_{ij} = \beta_0 + u_{0j} + \beta_1 \mathrm{intervention}_j + \ \beta_2 \mathrm{baseline\ EN}_{\ ij} + \beta_3 \mathrm{FSM}_{\ ij} +$$

$$+\beta_4 \mathrm{FSM}_{\ ij} * \mathrm{intervention}_j + \beta_5 \mathrm{stratum}_j + \epsilon_{ij}$$

With $\mathrm{FSM}_{\ ij}$ being a dummy variable for pupil i in school j's FSM eligibility status and the remaining variables as described in the Primary Outcome Analysis subsection above.

And the two multilevel level random intercepts models with interaction terms for the participation/non-participation in Mastering Number subgroup analysis will be given by:

$$\mathrm{PUMA}_{ij} = \beta_0 + u_{0j} + \beta_1 \mathrm{intervention}_j + \ \beta_2 \mathrm{baseline\ EN}_{\ ij} + \beta_3 \mathrm{MN}_{\ j} +$$

$$+\beta_4 \mathrm{MN}_{\ j} * \mathrm{intervention}_j + \beta_5 \mathrm{stratum}_j + \epsilon_{ij}$$

Where $\mathrm{MN}_{\ j}$ is the indicator of whether school j has signed up for the Mastering Number programme or not, and the remaining variables as described above in the *Primary Outcome Analysis* subsection.

Power analyses will also be performed to determine if subgroup analyses are underpowered. In accordance to the EEF 2018 guidelines, underpowered subgroup analyses will be reported as exploratory.

Longitudinal follow-up analyses

The PUMA will be re-administered to the cohort of pupils taking part in the trial at the end of Year 1 (in 2023). This timing will have the added benefit of including the additional influence of Year 1 teachers who were also trained alongside reception teachers. This follow-up will only proceed subject to the EEF guidance for longitudinal analysis[11].

These results would be analysed using a separate multilevel model similar to the one specified in the primary analysis except for the dependent variable, which in this case will be maths attainment as measured by the version of PUMA adopted for Year 1, rather than reception.

Imbalance at baseline

To assess imbalance between intervention and control groups at baseline we will produce cross-tabulations of background characteristics of the schools in the sample. We will examine the following background characteristics:

1. Proportion of FSM eligible pupils within the school
2. If the school is rural or urban
3. Type of school governance
4. Latest Ofsted rating.

To run this analysis, we will link the schools taking part in the trial to the relevant information contained on the most up to date edition of NFER's registry of schools.

[11] https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/longitudinal_guidance.pdf

We will also assess imbalance at baseline in terms of reception pupils' emergent numeracy by comparing the difference in means of the EN scores of the intervention and control groups and reporting them as Hedge's gs. For the purpose of computing the effect sizes we will be fitting two-level models, pupil and school, without covariates.

Missing data

Although a low level of pupil attrition is to be expected in the context of the Reception Jigsaw trial we will nevertheless tackle the problem of missing data if the proportion of participants with missing data exceeds five per cent.

After evaluating to what extent data are missing and counting the number of complete cases, we will proceed to identify patterns of missingness in terms of the primary outcome variable. By design only pupils who were assessed at baseline were included in the trial, and so we are not expecting to find missing cases in the data corresponding to any of the covariates of the primary analysis model (baseline emerging numeracy and randomisation group and block variables whose values are already known). As such, we will not investigate missingness in terms of any variables other than primary outcome.

We will start by investigating patterns of missing data by means of a two-level (pupil and school) logistic model where the outcome is missingness, with baseline EN, school region (Essex or Yorkshire), whether the school signed up to Mastering Number, and randomisation group indicators as covariates. Additional variables that may be associated with missingness, but which are not included in the primary analysis, will also be included as covariates.

After this stage the analyses will follow the road-map from EEF 2018 analysis guidance[12].

If necessary, sensitivity analysis built on a multilevel multiple imputation will be implemented. The missing primary outcome values will be imputed using predictive mean matching, with five plausible values derived for each case. The primary analysis model will then be re-run on the five sets of imputed plausible values and the estimates for 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.

The missing data analysis will be run in R (version 4.1.2) using the packages 'mice'[13] and 'smcfcs'[14] (pooling of the results of the plausible values models).

Compliance

As White Rose Maths will collect attendance registers at all training sessions, we will be able to link pupils to their reception teachers' attendance data. Compliance for the intervention will be defined at pupil level, in terms of the number of training sessions completed by a pupil's reception teacher.

[12] We are working under the expectation that there will be will no missing values among the models' covariate under MAR (missing at random), and that it will be possible to obtain valid estimates by including covariates predictive of non-response in the substantive models. The models' interpretation is conditional on these covariates being included.

[13] Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

[14] Jonathan Bartlett, Ruth Keogh and Edouard F. Bonneville (2021). smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification. R package version 1.6.0. https://CRAN.R-project.org/package=smcfcs

The evaluation team had initially conceptualized the compliance measures as school level measures based on the number of training sessions completed by each school. However, the team later revised this, as it was felt that school level measures would reflect poorly the experience of teachers and pupils, as well as the implementation of the intervention. While individual teacher attendance at the twilight sessions should be high, the nature of the coaching sessions is such that we would not expect all individuals to attend all coaching sessions. Furthermore, it is likely that the vast majority of the intervention schools will complete the full set of ten training sessions and school-level measures will to a high degree overlap treatment/control group allocation.

We will consider that a pupil has been taught by a reception teacher who completed the intervention if the teacher participates in nine or more sessions out of the ten available (five Twilight in-depth training sessions and five half-day coaching sessions); this binary measure will form an optimal compliance indicator in the CACE analysis. In addition we will run a separate CACE analysis using a pseudo-continuous dosage measure of compliance (0-10 sessions completed by the teacher).There is a small group of children in the trial's sample that were taught by more than one teacher at reception. We will not define compliance measures for these children or include their data on the CACE analyses.

To evaluate if there is an association between teachers having completed the intervention and maths attainment of their pupils we will adopt the instrumental variables methodology (IV) prescribed by the EEF 2018 guidelines: we will run an instrumental variable regression by two-stage least squares model with group allocation as the instrumental variable. The same approach will also be used to investigate the presence of an association between dosage (number of training sessions attended by a teacher) and pupil maths attainment. Both models will be fit using the function ivreg from the R package 'ivreg'[15] and the estimation of causal effects will be done resorting to the functions contained on the 'ivpack'[16] package.

The analyses will, as before, be run in R (version 4.1.2).

Intra-cluster correlations (ICCs)

ICCs will be estimated from the variance of the random intercept and residual variance of multi-level models by means of the formula:

$$ICC = \frac{\sigma_{intercepts}^2}{\sigma_{intercepts}^2 + \sigma_{residuals}^2}$$

Pre-test ICCs will be computed considering random intercepts two-level (school and pupil) models with no covariates, and post-test ICCs will be derived from the primary ITT and secondary ITT models described above**.**

Effect size calculation

As specified in the EEF 2018 guidelines, the results of the analyses of continuous outcomes by means of multi-level regression models will be reported as Hedges' g. The effect size will be calculated according to the formula

[15] John Fox, Christian Kleiber and Achim Zeileis (2021). ivreg: Instrumental-Variables Regression by '2SLS', '2SM', or '2SMM', with Diagnostics. R package version 0.6-1. https://CRAN.R-project.org/package=ivreg

[16] Yang Jiang and Dylan Small (2014). ivpack: Instrumental Variable Estimation.. R package version 1.2. https://CRAN.R-project.org/package=ivpack

$$\boldsymbol{g = \frac{\bar{o}i\ -\ \bar{o}c}{s^{*}}}$$

The numerator for the effect size calculations can be calculated as the coefficients of the intervention group from the regression models, and the denominator as the unconditional total variance from the corresponding models without covariates. The effect sizes thus computed are equivalent to Hedges' g.

If a logit model is used in the analysis of the first secondary outcome, the effect size will be reported as an odds ratios (OR). The odds ratios will be calculated according to the formula

$$\boldsymbol{OR = \exp\left(\beta_1\right)}$$

Where β_1 is the coefficient of the intervention/control dummy variable of the logistic regression model.

Confidence intervals for each effect size will be computed by multiplying the standard errors of the intervention group coefficient by the 2.5th percentile of a Student's t-distribution with the number of degrees of freedom associated with the sample size. The confidence intervals for the coefficient will be converted to effect size confidence intervals using the same formula as the effect sizes themselves.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), pp. 1-48. https://doi.org/10.18637/jss.v067.i01.

Chen, J.-Q., McCray, J., Adams, M., and Leow, C., 2014. A Survey Study of Early Childhood Teachers' Beliefs and Confidence about Teaching Early Math*. Early Childhood Education Journal*, 42, pp.367–377.

Kish, L., 1965. *Survey Sampling*. New York: Wiley.

Robinson-Smith, L., Fairhurst, C., Stone, G., Bell, K., Elliott, L., Gascoine, L., Hallett, S., Hewitt, C., Hugill, J., Torgerson, C., Torgerson, D., Menzies, V. and Ainsworth, H., 2018). *Maths Champions: Additional Appendices*. London: EEF [online]. Available at: https://educationendowmentfoundation.org.uk/public/files/Maths_champions-appends.pdf [Accessed 27 July 2021].

Appendix A

School-level pupil sampling syntax

```
library(tidyr)

setwd("…/01.Sampling")

## master  corresponds to the document used by NFER’s research and operations group to record
the collected information on reception pupils attending the participating schools

master=dir()[grep("Masterfile",dir())]
master=read.csv(master,skip=1,stringsAsFactors = FALSE)

## checking if the reported FSM eligibility (Y,N, U) was reported for all the pupils and setting missing
cases to unknown (U)

table(master$FSM..Y.N.,useNA = "ifany")
aux=nchar(master$FSM..Y.N.)==0
master$FSM..Y.N.[aux]="U"
table(master$FSM..Y.N.,useNA = "ifany")
master$FSM..Y.N.=factor(master$FSM..Y.N.,levels=c("Y","N","U"))

### checking that there are no duplicated unique pupil ids (Pupil.ID)

sum(duplicated(master$Pupil.ID))==0



### counting the number of reported FSM eligible and non-eligible pupils in each school  and also the
total number of pupils per school
### (NFERNO is an internal unique school identifier)

dist=as.data.frame(table(master[,c("NFERNO","FSM..Y.N.")]))
dist=pivot_wider(dist, values_from="Freq",names_from="FSM..Y.N.")
aux=as.data.frame(table(master$NFERNO))
colnames(aux)=c("NFERNO","total")
dist=merge(dist,aux)

### calculating the number of reported FSM eligible pupils that should be sampled in each school
### in schools where no pupil in  the cohort is FSM eligible or  non-eligible we will just randomly draw
a sample of participants and replacements

dist$FSM_cases=dist$Y+dist$N
dist$aux=dist$Y*dist$N
dist$FSM_cases=ifelse(dist$aux==0,0,dist$Y/dist$FSM_cases)

### The number of FSM-eligible pupils to be sampled needs to be adjusted for schools with less than
### 28 reception pupils

aux=dist$total
aux[aux>28]=28
dist$FSM_cases=aux*dist$FSM_cases
dist$aux=dist$FSM_cases
dist$FSM_cases=round(dist$FSM_cases,0)

### All the schools with at least 1-FSM eligible pupil amongst non-eligible will include at least one
FSM-pupil in the sample
dist$FSM_cases[dist$FSM_cases==0 & dist$aux>0]=1
```

```
## We will try to include at least 5 FSM-eligible pupils in each school sample

dist$FSM_cases[dist$FSM_cases>0 & dist$FSM_cases<5]=5
table(dist$FSM_cases,useNA ="ifany")

### In schools with less than 20 pupils the full cohort is to be included in the sample as selected to
take part in the trial

dist$aux=dist$total<21
table(dist$aux,dist$total)

### dist will be divided into 2 data frames, one containing the schools with less than 20 pupils (dist)
### and the other containing the schools with 21 or more pupils (dist0)

dist0=dist[!dist$aux,]
dist=dist[dist$aux,]
dist$FSM_sample=dist$Y
dist$sample=dist$total

### the select data frame will contain the info of the children selected to participate in the trial and the
replacements

select=master[master$NFERNO %in%dist$NFERNO,]

###confirming all the cases were accounted for

nrow(select)==sum(dist$sample)

select=select[,c(1:10,17)]
select$Selected_for_testing..Yes.No.Reserve.="Y"

### Separate schools where a we will just draw a random sample of selected + replacement pupils
### (dist0) from schools where we will be drawing a stratified random sample (dist00)

dist0$aux=dist0$FSM_cases
dist0$aux=dist0$FSM_cases==0
table(dist0$aux,dist0$FSM_cases)
dist0$FSM_sample=NA
dist0$sample=NA

dist00=dist0[!dist0$aux,]
dist0=dist0[dist0$aux,]

### setting random seeds to ensure replicability

set.seed(20210916)
aux=nrow(dist0)+2*nrow(dist00)
seeds=sample.int(n=1000000,size=aux)

# generating the selection labels (20x “Y, R1 to R8)

selection=1:8
selection=paste("R",selection, sep="")
selection=c(rep("Y",times=20),selection)

## drawing the random participant +replacement sample for each school in dist0
### (the first 20 elements of the random sample are assigned to participate in the trial and the
remaining ones as replacements)
```

```
for (i in 1:nrow(dist0)){

set.seed(seeds[1])
seeds=seeds[-1]
aux=master[master$NFERNO==dist0$NFERNO[i],]
n=min(28,nrow(aux))
aux1=sample(1:nrow(aux),n)
aux=aux[aux1,]
aux$Selected_for_testing..Yes.No.Reserve.=selection[1:nrow(aux)]
aux=aux[colnames(select)]
select=rbind(select,aux)
dist0[i,"FSM_sample"]=sum(aux$FSM..Y.N.=="Y")
dist0[i,"sample"]=nrow(aux)

}

dist=rbind(dist,dist0)
remove(dist0)

## drawing the stratified participant +replacement sample for each school in dist00
### (all the FSM eligible pupils are included in the assigned to participate in the trial group  and the remaining ones complete the assigned to participate group and replacement group)



for (i in 1:nrow(dist00)){

set.seed(seeds[1])
seeds=seeds[-1]
aux=master[master$NFERNO==dist00$NFERNO[i],]
aux0=aux[aux$FSM..Y.N.=="Y",]
aux=aux[aux$FSM..Y.N.!="Y",]
n=min(dist00$FSM_cases[i],nrow(aux0))
aux1=sample(1:nrow(aux0),n)
aux0=aux0[aux1,]
set.seed(seeds[1])
seeds=seeds[-1]
n=28-n
n=min(n,nrow(aux))
aux1=sample(1:nrow(aux),n)
aux=aux[aux1,]
aux=rbind(aux0,aux)
aux$Selected_for_testing..Yes.No.Reserve.=selection[1:nrow(aux)]
aux=aux[colnames(select)]
select=rbind(select,aux)
dist00[i,"FSM_sample"]=sum(aux$FSM..Y.N.=="Y")
dist00[i,"sample"]=nrow(aux)

}

dist=rbind(dist,dist00)
remove(dist00)

dist=dist[order(dist$NFERNO),]

### Exclude the withdrawn pupils form the selected to participate and replacement groups
### (record the withdrawn pupils Pupil.ID)

aux=master$Withdrawn.Pupil..Y.N.=="Y"
aux=master$Pupil.ID[aux]
```

```
### Merge the sampling information to master

master=master[,colnames(select)]
master=master[,-11]
select=select[,c(3,11)]
table(select$Selected_for_testing..Yes.No.Reserve.)
master=merge(master,select,all.x=TRUE)
table(master$Selected_for_testing..Yes.No.Reserve.,useNA="ifany")
master$Selected_for_testing..Yes.No.Reserve.[is.na(master$Selected_for_testing..Yes.No.Reserve.)]="N"
table(master$Selected_for_testing..Yes.No.Reserve.,useNA = "ifany")

### Check if all the withdrawn-pupil are not in the selected to participate+ replacement sample.
## (In this specific case there was only one pupil that was not selected as a participant or replacement)

master[master$Pupil.ID==aux,"Selected_for_testing..Yes.No.Reserve."]=="N"
```

Appendix B

School-level randomisation

```
setwd("…/03.Randomisation")

## master  corresponds to the document used by NFER’s research and operations group to record
the collected information on reception pupils attending the participating schools

###load data

data=dir()[grep("Masterfile",dir())]

data=read.csv(data,skip=1,stringsAsFactors =FALSE)

data=data[!is.na(data$Sample),]

### We will be drawing a sample of 55 intervention schools, the remaining ones are by default
### control. We are considering strata by region Essex/Yorkshire and Sign up to MN (yes/no): we want
the representativeness of each stratum in the control/intervention group to be similar to that of the
sample

#### Determine how many intervention schools in each stratum

data$Yorkshire.Essex=trimws(data$Yorkshire.Essex)

table(data$Yorkshire.Essex)

data$Signed.up.to.MN.programme.=trimws(data$Signed.up.to.MN.programme.)

table(data$Signed.up.to.MN.programme.)

data$Signed.up.to.MN.programme.[data$Signed.up.to.MN.programme.=="Unknown"]="No"

table(data$Signed.up.to.MN.programme.)

data$block=paste(data$Yorkshire.Essex,data$Signed.up.to.MN.programme.,sep="_")

table(data$block,useNA="ifany")

table(data$Yorkshire.Essex,data$Signed.up.to.MN.programme.,useNA="ifany")


##Find out how many intervention cases to draw from each stratum (start from the smaller ones)

aux=as.data.frame(table(data$block))

aux=aux[order(aux$Freq),]

aux$Freq=round(55*aux$Freq/nrow(data),digits=0)

sum(aux$Freq)
```

```
###In this case it's irrelevant, as the total adds up to 55, but just to be extra sure we will still have 55 schools in case we need to review the randomisation at a later stage (re-using the code already written)

aux$Freq[nrow(aux)]=55-sum(aux$Freq[-nrow(aux)])

###to make the code more legible, I'll change the names of the columns in aux

colnames(aux)=c("block","ncases")

data=data[,c(1:7,15,51)]

data=data[,c(1:5,7,6,8:9)]

#This will allow us to bring back the rows to their original order and also draw the samples

data$rand="control"

data=data[order(data$NFER_No),]

data$aux=1:nrow(data)

### Setting the seed

set.seed(20211025)

intervention=NULL



for (i in 1:nrow(aux)){

aux1=data$aux[data$block==aux[i,"block"]]

aux1=sort(aux1)

intervention=c(intervention,sample(aux1,size=aux[i,"ncases"]))

}

remove(aux1)


data$rand[intervention]="intervention"
```

Appendix C

Observational Check list for the Primary and Secondary I analyses

Please indicate for [insert pupil's name] whether, based on what you have observed since they started Reception, they can do each of the following tasks with support or independently.	cannot do with support	can do with support	can do independently
1. Recognises numerals 1 to 5			
2. Counts up to 3 or 4 objects by saying one number for each item in order			
3. Subitises (recognises quantities without counting) up to 5			
4. Counts objects up to 10			
5. Count objects beyond 10			
6. Counts out a smaller number of objects (up to 6) from a larger group - e.g. "give me four of the…"			
7. Selects the correct numeral to represent 1 to 5 objects			
8. Selects the correct numeral to represent 6 to 10 objects			
9. Estimates how many objects they can see and checks by counting them (up to 10)			
10. Uses the language of 'more than' and 'less than/fewer' to compare two sets of objects			
11. Finds the total number of items in two groups by counting all of them (up to 10)			
12. Finds one more or one less from a group of up to 5 objects			
13. Finds one more or one less from a group of up to 10 objects			
14. Records mathematical ideas using marks or objects they can interpret and explain			
15. Selects a particular named shape			
16. Can describe the relative position of objects or shapes such as 'behind' or 'next to'			
17. Uses familiar objects and common shapes to create and build models			

18. Continues, copies and creates repeating patterns			
19. Uses everyday language related to time			
20. Orders and sequences familiar events			

Appendix D

Practitioner Confidence in Teaching Maths Survey (adapted from Chen et al., 2014 and Robinson-Smith et al., 2018)

Question 1

Your confidence in helping reception aged children learn maths

1. Some early years practitioners have reported that they don't feel comfortable helping reception aged children learn maths. Others feel confident; still others say they are confident in some areas of maths but not in others.

In this table, please indicate how you feel about helping reception aged children learn maths.

For each of the following statements, rate your agreement by ticking the appropriate box.

I am confident in my **knowledge** of___________		[1]Strongly disagree	[2]Disagree	[3]Neither agree nor disagree	[4]Agree	[5]Strongly agree
1.1	what the children in my classroom know about maths when they enter reception					
1.2	reasonable maths goals for reception aged children					
1.3	the best practices and strategies for helping reception aged children learn maths					
1.4	national maths standards for reception aged children (for example, EYFS and EYFS profile results)					
1.5	the best ways to assess children's maths knowledge and understanding throughout the year					
I am confident in my **ability** to________		[1]Strongly disagree	[2]Disagree	[3]Neither agree nor disagree	[4]Agree	[5]Strongly agree
1.6	observe what reception aged children know about maths					
1.7	incorporate maths learning into common reception situations (such as art or dramatic play)					
1.8	plan activities to help reception aged children learn maths					
1.9	further children's maths knowledge when they make spontaneous maths comments/ discoveries *Example: When child says "I have four blocks" asking child how many blocks he would have if you gave him one more.*					
1.10	make sense of children's' confusions when they learn maths *Example: Why child thinks* ▲ *and* ◣ *aren't the same shape.*					
1.11	translate assessments into curriculum plans (i.e. turning assessments of children into next steps for learning)					

Question 2

Your confidence in your maths abilities

2. Some early years practitioners have reported that they just aren’t good at maths. Others say they love maths. Still others say how they feel depends on the specific area of maths. In this table, please indicate how you feel about maths and your maths abilities.

For each of the following statements, rate your agreement by ticking the appropriate box.

		[1]Strongly disagree	[2]Disagree	[3]Neither agree nor disagree	[4]Agree	[5]Strongly agree
2.1	Maths was one of my best subjects in school.					
2.2	Just the word “maths” can make me feel nervous.					
2.3	I’m not a “maths person”.					
2.4	I can easily rotate objects in my mind *Example: Figuring out how something would look from another angle.*					
2.5	I like coming up with creative ways to solve maths problems.					
2.6	I can easily convert fractions into percentages and decimal numbers.					
2.7	I have a bad sense of direction.					
2.8	I’m good at looking at numeric data and finding patterns.					
2.9	I’m good at estimating how tall something is or the distance between two locations.					