Effectiveness Trial of Mathematical Reasoning
Statistical Analysis Plan

Evaluator (institution): National Foundation for Educational Research
Principal investigator(s): Helen Poet

Education Endowment Foundation

PROJECT TITLE	Effectiveness trial of Mathematical Reasoning
DEVELOPER (INSTITUTION)	University of Oxford
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Helen Poet
SAP AUTHOR(S)	Chris Morton, Andrew Smith
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	Year 2
NUMBER OF SCHOOLS	240
NUMBER OF PUPILS	6,168
PRIMARY OUTCOME MEASURE AND SOURCE	Progress Test in Maths 7 (PTM7) test administered by teachers at baseline and by NFER test administrators at endpoint
SECONDARY OUTCOME MEASURE AND SOURCE	GL Assessment Progress Test in Maths (PTM7) 'process' categories (subscales): (i) fluency in facts and procedures, (ii) fluency in conceptual understanding, (iii) problem-solving, and (iv) mathematical reasoning.

Effectiveness Trial of Mathematical Reasoning
Statistical Analysis Plan

Evaluator (institution): National Foundation for Educational Research
Principal investigator(s): Helen Poet

Education Endowment Foundation

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]		*N/A*

Effectiveness Trial of Mathematical Reasoning Statistical Analysis Plan

Evaluator (institution): National Foundation for Educational Research
Principal investigator(s): Helen Poet

Education Endowment Foundation

Table of contents

Introduction

Mathematical Reasoning (MR) is a programme for pupils in Year 2 that aims to improve mathematical attainment by developing their understanding of the logical principles underlying mathematics. The programme focuses on quantitative reasoning and number sense and is based on the KS1 National Curriculum, with no new content introduced. One maths lesson per week for 12 weeks is replaced with a programme session. Each session has whole-class teacher-led time, followed by differentiated group time. In the differentiated group time, half the class receive tailored teacher support and half play programme-specific computer games to embed learning. The programme was developed by Professor Terezinha Nunes and Professor Peter Bryant at the University of Oxford.

The teacher and teaching assistant (TA) that will deliver the programme together in the classroom are trained as a pair, receiving the same training via e-learning and professional development support. The programme aims to improve teacher and TA pedagogical knowledge around mathematical and numerical reasoning and to increase understanding of the importance of teaching these concepts and skills from a young age. While the core purpose of the CPD is to train teachers and TAs to effectively deliver the MR programme, it is also intended to empower teachers and TAs to apply their learning from the programme to other areas of their work and to share it with their colleagues. The CPD element is delivered as an online training programme made up of five core modules and an additional four modules (only the last of which is compulsory) accompanied by tailored implementation support from trained professionals.

The Mathematical Reasoning programme has been the subject of two randomised controlled trials commissioned by the Education Endowment Foundation (EEF), both with a high-security rating. The efficacy trial (Worth *et al.*, 2015) was a three-arm trial conducted by the National Foundation for Educational Research, with 17 out of 55 schools allocated to the MR group. The Oxford University team trained teachers directly through a one-day in-person training session, with one follow-up visit to each participating school to observe delivery and provide personalised feedback. The efficacy trial found that the programme achieved a positive impact on pupils' numeracy abilities (effect size of 0.20), compared to pupils who had not received the programme. A subsequent effectiveness trial (Stokes *et al.*, 2018), carried out by the National Institute of Economic and Social Research involved 160 schools. A train-the-trainer model was employed, with the National Centre for Excellence in the Teaching of Mathematics helping to develop the training model, which was delivered through the national network of 'Maths Hubs'. The trial found a smaller impact than the efficacy trial (effect size of 0.08) and the results were not statistically significant.

Given the positive impact identified in the efficacy trial, the EEF is interested in determining how to effectively implement the Mathematical Reasoning programme on a larger scale. This second effectiveness trial seeks to understand whether a new training model may better retain the scale of impact seen at the efficacy stage by enabling direct contact with the Oxford University team's teaching material. In the updated training model, the developers have removed the train-the-trainer element and developed an online training course with support for schools from Teacher Leaders employed and trained by the University of Oxford team. The primary focus of the impact evaluation will be to estimate the impact of the programme, when teachers are trained via the online course, on short-term pupil mathematical attainment outcomes. This is in line with the Theory of Change (ToC) (see the trial protocol (Flemons *et al.*, 2024)). The design of the impact evaluation is broadly congruent with the previous

effectiveness trial to allow for some comparison of findings and as the potential basis for inference about the revised training model.

Design overview

Trial design, including number of arms		Two-arm cluster randomised controlled trial with random allocation at school level
Unit of randomisation		Schools
Stratification variables (if applicable)		N/A
Primary outcome	variable	Maths attainment
	measure (instrument, scale, source)	Progress Test in Maths 7 (PTM7) (test administered by NFER test administrators, 0-43 raw score, GL Assessment)
Secondary outcomes	variable(s)	Maths attainment (specific domains)
	measures (scales)	Progress Test in Maths 7 (PTM7) 'process' categories (subscales): (i) fluency in facts and procedures (0-6 raw score) (ii) fluency in conceptual understanding (0-13 raw score) (iii) problem-solving (0-4 raw score) (iv) mathematical reasoning (0-20 raw score).
Baseline for primary outcome	variable	Maths attainment
	measure (instrument, scale, source)	Progress Test in Maths 6 (PTM6) (test administered by teachers, 0-31 raw score, GL Assessment)
Baseline for secondary outcomes	variable	Maths attainment (specific domains)
	measure (instrument, scale, source)	Progress Test in Maths 6 (PTM6) 'process' categories (subscales): (i) fluency in facts and procedures (0-2 raw score) (ii) fluency in conceptual understanding (0-12 raw score) (iii) problem-solving (0-3 raw score) (iv) mathematical reasoning (0-14 raw score).

This impact evaluation is conducted as a Randomised Controlled Trial, in which schools are randomised to either the intervention or control arm. The trial pupils are Year 2 pupils in participating classes: schools specify at least one class where the teacher will receive the Mathematical Reasoning training (if the school is randomised to the intervention). All trial pupils in intervention schools are eligible to receive the Mathematical Reasoning intervention, with intervention sessions being delivered to whole classes. Control schools continue with 'business as usual'; in these schools Year 2 pupils do not access the Mathematical Reasoning programme but may receive other maths interventions and support

(as decided by their school), including interventions that aim to promote mathematical reasoning.

Research questions

The **primary impact research question** asks:

> **RQ1**: *What is the impact of the Mathematical Reasoning programme on Year 2 pupils' attainment in mathematics (measured using GL PTM7)?*

RQ1 is aligned with the programme's Theory of Change (Flemons *et al.*, 2024), which identifies 'improved mathematical performance' as a short-term outcome for pupils. The question is also very similar to the primary impact research question in the previous effectiveness trial (Stokes *et al.*, 2018), including the use of the same outcome measure. Answering this will, therefore, also allow for some comparability of findings.

Secondary impact research questions focus on FSM-eligible pupils, subscales, dosage, and how impacts may vary by pupil prior attainment and computer game usage.

The first secondary research question is similarly concerned with mathematical attainment measured using GL Assessment's Progress Test in Maths (PTM7) but focuses specifically on FSM-eligible pupils, as the programme ToC hypothesises FSM status to be a moderator of treatment effects.

> **RQ2.1**: *What is the impact of the Mathematical Reasoning programme on Year 2 FSM-eligible pupils' attainment in mathematics (measured using GL PTM7)?*

The second secondary research question, RQ2.2 (below), uses the same outcome measure as the primary research question but will measure impact against the PTM7 subscales. This analysis will provide additional findings for consideration alongside those of the primary research question, particularly to understand how the programme's impact is associated with specific aspects of mathematical attainment. One of the subscales in particular (mathematical reasoning) is hypothesised by the ToC to be directly associated with receipt of the programme, with the expectation that pupils participating in the programme will have a better understanding of quantitative reasoning.[1]

> **RQ2.2***: What is the impact of the Mathematical Reasoning programme on each of the PTM7 'process' categories (subscales): (i) fluency in facts and procedures, (ii) fluency in conceptual understanding, (iii) problem-solving, and (iv) mathematical reasoning (measured using GL PTM7)?*
>
> ***a)*** *for all pupils*
> ***b)*** *for FSM-eligible pupils*

The third and fourth secondary research questions concern dosage in order to understand how impacts vary by the number of sessions attended by the pupil and by pupils' use of computer games (which the ToC identifies as a potential moderator of impacts).

> **RQ2.3:** *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by the number of sessions attended by the pupil?*

[1] The developers of the MR programme specified that their definition of mathematical reasoning is not the same as that used by the PTM7 test designers, as discussed further in the 'Primary and secondary outcomes' section.

a) *for all pupils*
b) *for FSM-eligible pupils*

RQ2.4: *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by i) the number of computer games played by the pupil and ii) the number of different computer games played by the pupil?*

a) *for all pupils*
b) *for FSM-eligible pupils*

We will also investigate whether and how the impact of the programme varies by pupil prior attainment. Although our primary research question (RQ1) includes prior attainment as a baseline to control for it and to increase precision, this question specifically looks at how programme participation interacts with pupil prior attainment. Answering it will help us to understand (along with FSM-eligible status) for whom the programme is most effective.

RQ2.5: *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by pupil prior attainment?*

a) *for all pupils*
b) *for FSM-eligible pupils*

We will investigate the impact of teacher completion of the MR training as a research question, separate to the compliance analysis.

RQ2.6: *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by whether their teacher completed the training?*

a) *for all pupils*
b) *for FSM-eligible pupils*

Sample size calculations overview

Table 1 below illustrates our estimation of the minimum detectable effect size (MDES) given a randomised sample of 240 schools, with a 1:1 allocation of schools to the two trial arms. The MDES values given in Table 1 were calculated using the PowerUpR package in the R statistical software, using the function 'mdes.cra2' (Bulus *et al.*, 2021).

Given that the previous MR effectiveness trial (Stokes *et al.*, 2018) found an effect size of 0.08 amongst all pupils, we considered it appropriate to power this trial for a relatively low MDES at the protocol stage. We therefore recommended that the Oxford University team recruit schools to the upper limit of their capacity to deliver the intervention (120 schools). We assumed one class per school (25.7 pupils per class, 7 of whom are FSM-eligible[2]), with 10% of schools having mixed-year classes (which would mean fewer pupils eligible for the evaluation in these schools).

The number of schools randomised was slightly fewer than planned (238 versus 240). However, the number of pupils per class was slightly higher than expected, as was the proportion of FSM pupils (based on FSM data collected directly from schools). The overall

[2] 26.9% of pupils indicated as eligible for FSM by the EVERFSM_6_P NPD variable. 2022/23 DfE figures indicate 24.6% of Year 2 pupils were FSM eligible, a figure that has risen year-on-year.

effect was that the MDES amongst both all pupils and FSM pupils remained virtually unchanged between the protocol and randomisation stages.

While GL Assessment reports that the correlation between PTM6 and PTM7 is 0.67, we assumed a slightly lower correlation (0.62). This is due to the use of these assessments in the context of a programme, where we would expect different correlations between the intervention and control groups, resulting in an overall lower correlation than what would be found outside of a programme context. The formula used by 'PowerUpR' does not use the pre-post correlation directly, instead requiring the proportion of variance explained at level 1 and level 2 ((Bloom, Richburg-Hayes and Black, 2007) provides an explanation of these parameters). We assume the variance explained at both level 1 and level 2 will be equal[3], in which case both will be equal to the pre-post correlation (0.62) squared. We have also assumed a school-level ICC of 0.11, which aligns with analysis from the previous effectiveness trial (Stokes *et al.*, 2018).

Table 1: sample size calculations at the protocol and randomisation stages (no attrition assumed)

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.108	0.140	0.108	0.138
Pre-test/ post-test correlations	level 1 (pupil)	0.62	0.62	0.62	0.62
	level 2 (school)	0.62	0.62	0.62	0.62
Intracluster correlation (ICC)		0.11	0.11	0.11	0.11
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		25.7	6.8	26.4	7.2
Number of schools	intervention	120	120	119	119
	control	120	120	119	119
	total	240	240	238	238
Number of pupils	intervention	3,084	816	3,282	857
	control	3,084	816	3,010	867
	total	6,168	1,632	6,292	1,724

The figures in Table 1 do not account for post-randomisation attrition. To see the anticipated impact of attrition, school and pupils numbers in Table 2 have been reduced to allow for 10% school-level attrition and 15% pupil-level attrition. This assumes 90% of schools remain in the trial, and within these schools, 85% of pupils will be included in the primary analysis. For the sake of simplicity, it is assumed that all school-level attrition occurs post-randomisation. This

[3] The parameters in Appendix 5 of the MR efficacy trial (Worth et al., 2015) support this assumption.

shows that due to attrition, an MDES of approximately 0.117 is anticipated at the analysis stage

Table 2: sample size calculations at the protocol and randomisation stages, with pupil and school numbers reduced to include anticipated attrition

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Anticipated attrition rate	level 1 (pupil)	15%	15%	15%	15%
	level 2 (school)	10%	10%	10%	10%
Minimum Detectable Effect Size (MDES)		0.117	0.154	0.117	0.153
Average cluster size		21.8	5.8	22.4	6.1
Number of schools	intervention	108	108	107	107
	control	108	108	107	107
	total	216	216	214	214
Number of pupils	intervention	2,359	624	2,510	655
	control	2,359	624	2,302	663
	total	4,718	1,248	4,812	1,318

Further parameters (e.g. ICC) are the same as in Table 1.

Randomisation

Schools were randomised to the intervention and control groups in a 1:1 ratio. The randomisation was performed using the R statistical software by an NFER Statistician, using the code included in Appendix A. The statistician was not blinded to group allocation. Code was stored for transparency and replicability of the randomisation process.

A simple randomisation was performed for this trial, without stratification. The potential benefits of stratification are: (i) to aid physical delivery of the programme, or (ii) to increase power when a stratifier explains substantial amounts of outcome variance. In this case, the first benefit is not relevant because the Teacher Training programme is delivered online. The second is also unlikely to be relevant: evidence from educational data suggests that further socioeconomic variables explain little additional variance when a baseline measure is included in modelling (see Stallasch *et al.*, 2024 and appendix tables from Singh *et al.*, 2023). In the absence of these benefits, we considered simple randomisation to be preferable to stratified randomisation, which may increase the chance of selection bias and errors during the randomisation process (Hewitt and Torgerson, 2006). Whilst not stratifying, we will include a sensitivity check that adds socioeconomic variables to the primary analysis model (see 'Additional analysis and robustness checks' section).

Primary and secondary outcome measures

The **primary outcome** for the trial is maths attainment, measured using overall raw scores from GL Assessment's PTM7 (Form A). For this trial, we expect raw scores to be less

vulnerable to floor and ceiling effects than age-standardised scores[4]. The choice of raw score rather than age-standardised score also makes the primary and secondary outcomes more comparable (age-standardised scores are not available for the PTM7 subscales). There will be no adjustment for pupil age, either through standardisation or by including age as a model covariate, given that the maximum age gap between any two trial pupils is roughly one year. Randomisation will help ensure that results are not biased by age imbalances between the trial arms, and we do not expect age to explain a large amount of model variance after accounting for the PTM6 baseline.

The choice of 'maths attainment' for the primary outcome is consistent with the ToC which includes 'improved mathematical performance' as a short-term pupil outcome. GL Assessment's PTM offers a suitable measure for the primary outcome of mathematical performance, particularly as it aligns with national curriculum objectives in English schools. Furthermore, the use of PTM7 is consistent with the previous effectiveness trial, thus allowing for a degree of comparability with previous evaluations of the programme.[5]

Multiple versions of the PTM assessment are available, representing different levels (i.e. year groups). This means it can be used to measure progress within a single year or across multiple years, thus making it suitable for both a baseline and outcome measure in this study. GL's tests are designed to measure progress using different test versions, rather than by repeated administration of the same questions. The PTM questions were developed by the Mathematics Assessment Resource Service (MARS) team, which is a collaboration between the University of California, Berkeley and the Shell Centre, Nottingham. Standardisation was conducted with 34,762 pupils in the UK for all test versions (4,071 pupils for PTM7). Internal consistency is reported at 0.91 (Cronbach's Alpha), and the assessment has been found to have no significant difference in standard age scores between male and female pupils. Given the robust development process of the test and its established psychometric properties, we will use the instrument as developed by GL Assessment and without modification for our main analysis (see also 'Additional analyses and robustness checks' section below).

All PTM assessments are based on categories of mathematical proficiency, which have been derived by GL Assessment from the Curriculum Aims in the KS1, KS2 and KS3 National Curriculum for England (2013). The assessment comprises of 34 questions aligned with these categories, as follows (further details in Appendix B of the study protocol):

- Fluency in facts and procedures
- Fluency in conceptual understanding
- Mathematical reasoning
- Problem-solving

The categories will be used as subscales for **secondary outcomes**,[6] thus providing additional findings for consideration alongside those of the primary question, particularly to understand how the programme's impact is associated with specific aspects of mathematical attainment. One of the subscales in particular (mathematical reasoning) is hypothesised to be directly

[4] This is based on inspection of the distribution of baseline PTM6 raw and age-standardised scores for this trial, as well as comparing figures 3 and G.1 in the previous effectiveness trial.
[5] The efficacy trial used Progress in Maths 7 (PiM7), the forerunner to PtM7. GL Assessment report the correlation between the PiM7 and PTM7 to be 0.8, based on a sample of 350 pupils.
[6] For RQ2.2. The other secondary RQs use the same outcome as primary RQ1.

associated with receipt of the programme and is identified by the ToC as a short-term outcome ('Pupils have a better understanding of quantitative reasoning'). However, the developers of MR emphasise that the programme promotes reasoning about relations between quantities and about relations between numbers. This is a narrower focus than the 'mathematical reasoning' PTM7 subscale; not all questions on the subscale relate to programme content.

The PTM7 subscales - as defined by GL Assessment - will be used in the secondary analysis, rather than developed using data-driven methods such as exploratory factor analysis. We will analyse the trial test data for reliability as a sensitivity check (see confirmatory factor analysis in the analysis section) and expect it to be sufficient given GL Assessment's prior development work and published data.

Assessments will be administered at the endpoint by NFER Test Administrators, who will be blinded to school allocation to programme or control.[7] Pupils will complete the assessment within the classroom and under test conditions. PTM7 is not time-limited, but GL Assessment has suggested that approximately 35 minutes would be needed for pupils to demonstrate their abilities. Test Administrators will use a secure courier to send the assessment papers to GL Assessment, who will complete the marking and scoring, blinded to group allocation. GL Assessment will then use the NFER secure portal to share with NFER a spreadsheet containing the pupil-level data, including overall raw scores and sub-total raw scores for each of the aforementioned categories, alongside item-level scoring. Standard Age Scores, Stanine Scores and National Percentile Ranks will also be included in this data.

Baseline measures

The baseline measure will be GL Assessment's PTM6 Form A (for RQs 1, 2.1, 2.3, 2.4, 2.5, 2.6), which has been designed for administration at the end of Year 1 or the start of Year 2. This assessment is based on the same categories as PTM7, with a Pearson correlation between PTM6 and PTM7 of 0.67 (Bishenden, 2023). As a point of comparison, the pre-post correlation in EEF trials where a commercial test has been used at baseline and endpoint varies between 0.28 and 0.75, with a median of 0.54 (Singh *et al.*, 2023)[8]. As the PTM6 tests content covered in the previous academic year, it provides a more accurate assessment of pupil ability at baseline, compared to using the PTM7. For RQ2.2, the baseline measure will be the PTM6 subscales.

The baseline assessment will be administered by class teachers, who will be asked to do so within the classroom and under exam conditions.

Analysis

An Intention-To-Treat (ITT) approach will be followed throughout (except the compliance analysis), with pupils analysed according to their intervention or control group assignment, regardless of their degree of participation in the intervention. Analysis will be conducted on complete cases only; pupils with missing values for any variables used in an analysis model will be excluded from modelling. The missing data analysis will investigate the sensitivity of

[7] While schools will be asked not to share this information with the Test Administrator, it is possible that the Test Administrator may be made aware of the school's allocation through interaction with pupils and/or staff.

[8] EEF trials will implement an educational intervention for some pupils, which may affect the correlation between pre- and post-test, somewhat limiting the comparability of these results with the correlation between the PTM6 and PTM7.

results to this choice. The analyst will not be blinded to intervention assignment for any part of the analysis.

All analysis will be conducted in accordance with the EEF analysis guidance, using the R software (The R Foundation, 2023). Mixed effects models will be analysed using R package 'lme4' (Bates *et al.*, 2015).

Primary outcome analysis (RQ1)

The purpose of the primary analysis is to answer the research question:

> **RQ1**: *What is the impact of the Mathematical Reasoning programme on Year 2 pupils' attainment in mathematics (measured using GL PTM7)?*

A two-level (pupil and school) linear mixed effects model will be used for this analysis:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

In this model:

$\mathrm{PTM7_{ij}}$ = PTM7 score of pupil i in school j, measured at endpoint;

β_0 = intercept term;

$\mathrm{intervention_j}$ = indicator for whether school j was randomised to the intervention (1) or control (0);

$\mathrm{PTM6_{ij}}$ = PTM6 score of pupil i in school j, measured at baseline;

β_2 = average impact of each additional point of the baseline PTM6 score on the PTM7 score;

$\mathrm{b_j}$ = school-level error term (random intercept);

ϵ_{ij} = pupil-level residual error term.

The main estimate of interest is β_1, which represents the average impact of Mathematical Reasoning on PTM7 score, with a 95% confidence interval for this estimate calculated using the profile likelihood. The estimate and confidence interval will be converted to a standardised effect size, as described in 'Estimation of effect sizes' below.

Impact in the FSM subgroup (RQ2.1)

The effect of Mathematical Reasoning amongst FSM-eligible pupils[9] (RQ2.1) will be determined by repeating the primary analysis model, restricted to pupils from the FSM subgroup. Additionally, we will investigate the differential effect of the programme for FSM-eligible pupils relative to non-FSM pupils using the model:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6_{ij}} + \beta_3 \mathrm{FSM_{ij}} + \beta_4 \mathrm{FSM_{ij}} * \mathrm{intervention_j} + \mathrm{b_j} + \epsilon_{ij}$$

Where $\mathrm{FSM_{ij}}$ indicates whether a pupil is eligible for FSM (1) or not eligible (0). β_3 is the difference in PTM7 score for FSM pupils compared to non-FSM pupils, amongst pupils in the

[9] 'EVERFSM_6_P' from the 2024/25 spring census will be used to measure FSM eligibility throughout the impact analysis. The variable indicates whether a pupil has been eligible for FSM at any point in the last six years. For this evaluation of Year 2 pupils FSM records do not go back the full six years, since FSM status is first recorded in the reception year.

control group. The interaction term coefficient β_4 measures the differential impact of the intervention for FSM-eligible pupils relative to non-FSM pupils, which is the key aspect of interest from this model.

If β_4 is statistically significant, it suggests that the impact of the intervention differs between FSM-eligible and non-eligible pupils. A positive β_4 would then indicate a more favourable outcome for FSM-eligible pupils than for non-eligible pupils when receiving the intervention, while a negative β_4 would suggest the opposite. If the difference is not significant, there is no evidence that the intervention has a differential effect between FSM-eligible and non-eligible pupils.

Secondary outcome analysis (RQ2.2)

Four subscales of the PTM7 will be included as secondary outcomes (for RQ2.2 only – other secondary research questions will use the primary outcome of PTM7): (i) fluency in facts and procedures, (ii) fluency in conceptual understanding, (iii) problem-solving, and (iv) mathematical reasoning. These subscales will be the dependent variables in four linear two-level (pupil, school) regressions. For example, in the case of the fluency in facts and procedures subscale, the regression model will be:

$$\mathrm{PTM7\_FFP_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6\_FFP_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

where $\mathrm{PTM7\_FFP_{ij}}$ is the score of pupil *i* in school *j* for the fluency in facts and procedures subscale of the PTM7. Similarly, $\mathrm{PTM6\_FFP_{ij}}$ will be the pupil's score for the fluency in facts and procedures subscale of the PTM6 at baseline. This regression will then be repeated for each of the other three subscales, with $\mathrm{PTM7\_FFP_{ij}}$ and $\mathrm{PTM6\_FP_{ij}}$ replaced by the appropriate PTM7 and PTM6 subscales.

Due to the smaller range of possible scores for the subscales (the smallest range is 0-4 for problem solving), they may be more vulnerable to floor and ceiling effects, which was observed in the first evaluation of Mathematical Reasoning (Stokes *et al.*, 2018). Histograms will be produced for the PTM6 and PTM7 scores on each subscale and results will be caveated if potential floor or ceiling effects are observed. Regardless of whether there are floor or ceiling effects, there is a risk that the subscales with smaller ranges will not provide enough information to fully capture a pupil's ability in the relevant domain (e.g. problem solving). This could, in turn, prevent an impact of MR being observed for that subscale. The secondary analysis results will include this caveat, especially if the scales with smaller ranges show less or no impact.

As there are multiple outcomes in the secondary analysis, this will lead to an inflated 'family-wise error rate' (chance of one or more false positives) when conducting significance tests. We do not, however, intend to implement a multiple testing correction (e.g. Bonferroni) to address this, as we expect the resulting analysis would be severely underpowered to detect an intervention effect (inflated false negative rate). Rather, we will report any isolated low p-values amongst the secondary analysis results with caution and avoid a dichotomous 'significant' versus 'non-significant' interpretation of p-values.

Further moderator analyses (RQ2.3 – RQ2.6)

Impact of the number of sessions attended (RQ2.3)

The compliance analysis investigates the impact of the number of MR units attended; in this section, we separately look at the impact of the number of sessions attended (RQ2.3). There are 12 MR units in total, which may be delivered to pupils in a single session, or more than one session may be required to complete a unit. To investigate the impact of the number of sessions attended by a pupil, a two-level (pupil, school) linear regression will be run, restricted to intervention pupils only.

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{N\_sessions_{ij}} + \beta_2 \mathrm{PTM6_{ij}} + \beta_3 \mathrm{N\_absences_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

In this equation $\mathrm{N\_sessions_{ij}}$ is the number of Mathematical Reasoning sessions attended by of pupil i in school j and β_1 is the average change in PTM7 score per session attended[10]. $\mathrm{N\_absences_{ij}}$ is the number of days absent from school (not just Mathematical Reasoning sessions - in the autumn and spring terms of 2024/25) and β_3 is the average change in PTM7 score per additional absence. The rationale for including absences as a covariate in the model is that persistent absence is likely to impact both number of sessions attended and attainment (i.e. it is a confounder). The model above will be run for both all pupils and for the subgroup of FSM-eligible pupils.

Impact of computer game participation (RQ2.4)

To investigate the moderating effect of (i) the number of computer games played and (ii) the number of *different* computer games played (RQ 2.4), two further models will be analysed. These models will be restricted to intervention pupils only.

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{N\_games_{ij}} + \beta_2 \mathrm{PTM6_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{N\_different\_games_{ij}} + \beta_2 \mathrm{PTM6_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

Here $\mathrm{N_games_{ij}}$counts the number of games played by pupil i in school j and $\mathrm{N\_different\_games_{ij}}$ counts the number of different games played. As the benefit of additional games may depend on the number of different of games played and vice versa, a further model will be analysed:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{N\_games_{ij}} + \beta_2 \mathrm{N\_different\_games_{ij}} + \beta_3 \mathrm{N\_games_{ij}} \times \mathrm{N\_different\_games_{ij}} + \beta_4 \mathrm{PTM6_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

In this model β_3 is the coefficient for the interaction between $\mathrm{N\_games_{ij}}$ and $\mathrm{N\_different\_games_{ij}}$, representing how the impact of each additional game played changes with the number of different games played (and vice versa).

We note that the number of computer games played is not randomised, nor is it possible to use quasi-experimental methods such as instrumental variables analysis in this case. The analysis is therefore vulnerable to confounding from unobserved variables. For example, pupils who play more computer games may be more engaged to learn maths generally and so score higher in the PTM7 because of this (rather than benefitting from the games per se). This potential for confounding is reduced by the inclusion of the PTM6 covariates, measured shortly before the trial, but nevertheless, we consider this analysis to be exploratory.

[10] This relationship assumes the content of the individual sessions does not matter, which would not be the case if, for example, the first few sessions were pivotal to understanding the remainder of the sessions. However, we believe the linear relationship will be a reasonable approximation of the truth, in the absence of any strong prior theory to the contrary.

In addition to answering RQ2.4 we will undertake further exploratory analysis to investigate any potential relationships between when computer games are played (in school, or at home), FSM-eligibility and PTM7 score, as well as between computer game scores (e.g. 100% correct responses) and the pupil PTM7 score. The exact modelling approach will depend on the quality of the games data (at the time of writing, there have been difficulties accurately linking some pupils to their game data) and the patterns of game participation and scores observed in the final data.

Prior attainment moderator analysis (RQ2.5)

Further analysis will be performed investigating how the impact of the intervention varies with prior attainment (RQ2.5a). Prior attainment (PTM6) is already included as a continuous variable in the primary analysis model; this investigation requires the addition of an interaction term to that model:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6_{ij}} + \beta_3 \mathrm{PTM6_{ij}} * \mathrm{intervention_j} + \mathrm{b_j} + \epsilon_{ij}$$

Here $\boldsymbol{\beta_3}$ estimates the additional (or lesser) impact of the intervention per additional point of prior attainment on the PTM6.

This same analysis will also be repeated for only FSM-eligible pupils, investigating the differential impact of the intervention as prior attainment increases for this subgroup (RQ2.5b). This second analysis, restricted to FSM pupils, is exploratory as it is very likely to be underpowered, so results will be caveated in the final report. We will not attempt to dichotomise PTM6 scores to form a 'lower prior attainment' subgroup, instead restricting our investigation to the model above that treats PTM6 score as continuous.

Impact of teacher training attendance (RQ2.6)

Teacher training completion is not investigated in the compliance analysis, but is included as a research question here (RQ2.6). Training completion for the purpose of this trial is defined as the first 5 training modules and two out of three training webinars completed. We understand that completion of training modules 1-4 is a prerequisite for completing module 5, so it will not be possible to investigate the impact of completing fewer than 5 modules. A two-level (pupil, school) linear regression will be run, restricted to intervention pupils only.

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{N\_units_{ij}} + \beta_2 \mathrm{PTM6_{ij}} + \beta_3 \mathrm{Training\_completed_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

In this equation $\mathrm{Training\_attendance_{ij}}$ is a binary indicator for whether a pupil's teacher fully completed the MR training (as defined above). β_3 is therefore the average impact of receiving MR from a teacher with complete training, compared to one with incomplete training, on pupil PTM7 scores. This is adjusted for number of units a pupil attends, $\mathrm{N\_units_{ij}}$, so that β_3 better reflects the benefit of the training itself, rather than any correlation with the number of units delivered. The model above will be run for both all pupils and for the subgroup of FSM-eligible pupils.

Imbalance at baseline and sample representativeness

To assess imbalance in baseline school- and pupil-level characteristics a table will be produced describing the characteristics of the control and intervention groups after randomisation. The following characteristics will be described:

School level

- Proportion of pupils eligible for FSM in 2024/25
- Proportion of pupils with special educational needs (SEN) in 2024/25
- Whether the school is urban or rural
- School type (academy, maintained or independent) in 2024/25
- Most recent overall Ofsted rating in 2024/25[11]
- Proportion of pupils meeting the expected standard in their KS2 maths exam in 2024/25

Pupil level

- FSM eligibility
- Whether pupils have SEN
- Gender
- Whether pupils speak English as an additional language (EAL)
- Baseline PTM6 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 via publicly available data releases from the Department of Education, such as Get Information About Schools[12]. Pupil-level variables will be obtained from the 2024/25 spring census within the NPD (FSM, SEN, gender, EAL) or collected directly from schools (PTM6 score). The difference between the baseline PTM6 scores in the intervention and control groups will be estimated using a 2-level (pupil, school) linear model and expressed as a standardised effect size.

Wherever national data is available we will also describe the same school and pupil-level characteristics for the 'target' population (the wider population for which this trial seeks to draw conclusions). At a school level this is all primary schools in England and at a pupil level it is all Year 2 pupils in England. By comparing characteristics in the trial sample with those in the target population we can assess how representative the sample is, which may help infer the external validity of trial results.

Missing data

The number and proportion pupils with any missing primary analysis variables (PTM6 or PTM7 score) will be reported. PTM6 score could be missing for individual pupils included in the trial, but not for entire schools[13]. If this is less than five percent then the potential for bias in a complete case analysis will be considered minimal and (in line with EEF's statistical guidance; EEF, 2022) no further missing data analysis will take place. Otherwise, we will report the number and proportion of missing values, using a flow chart to specify both variables included in the primary analysis model and for both intervention and control groups.

[11] Despite the recent revisions to Ofsted inspections, an overall rating will still be available for all inspected schools until the end of the 2024/25 academic year.
[12] https://get-information-schools.service.gov.uk/Downloads
[13] PTM6 tests were sat shortly before randomisation. Randomisation occurred at a school level, with all eligible pupils within a randomised school considered to be part of the trial, even if they had a missing PTM6 score. However, a school that dropped out before PTM6 testing was known not to be participating in the trial and so was not randomised and would not be included in missing data analysis. This is why pupils have missing PTM6 scores but not entire schools.

To identify patterns of missingness, a mixed effects logistic regression[14] with two levels (pupil and school) will be run, where the outcome is the logit probability of the PTM7 outcome being missing. The PTM6 and intervention indicator variables from the primary analysis model will be included as covariates, together with these auxiliary variables that may be associated with missingness:

- FSM eligibility in the spring term of 2024/25
- SEN status in the spring term of 2024/25
- Days of authorised absence from school (authorised or unauthorised) in 2024/25
- Number of pupils at the school per full-time teacher in 2024/25
- Proportion of pupils eligible for FSM in 2024/25

A second logistic regression will be run in which the outcome is missingness of the PTM6 baseline. PTM7 score will replace PTM6 score as a predictor in this second model, but it will otherwise be the same as that described above.

Any of the auxiliary variables which demonstrate an association with missingness in the PTM7 scores (indicated by a p-value below 0.05) will be included as covariates to re-run the primary analysis model as a sensitivity check[15]. Additional variables associated with missingness in either PTM6 or PTM7 score will be included as predictors in the multiple imputation procedure described later in this section. If re-running the primary analysis including covariates associated with missingness in the PTM7 scores alters the substantive interpretation of the intervention effect, then the PTM7 data may be *'missing at random'* (MAR) conditional on the inclusion of those covariates (though it may still be *'missing not at random'* (MNAR)).

It will not be possible to determine whether the data are MAR or MNAR from the observed data[16]. We will, therefore, perform further sensitivity analyses, using multiple imputation to explore three simple missing data patterns for pupils with a missing PTM7 outcome:

a) Outcomes are MAR, conditional on available covariates.
b) Outcomes are MNAR and on average missing outcome scores are lower than non-missing scores by one standard deviation[17]. Missing scores are 'equally worse' between the intervention and control groups.
c) Outcomes are MNAR for intervention pupils only: intervention pupils with missing outcomes have their score reduced to the average amongst control pupils[18]. This models a situation where, amongst pupils with missing outcomes, intervention pupils do not receive any benefit from the intervention.

The plausibility of these missing data patterns depends on the reason why a pupil did not have a PTM7 score available. We will be able to distinguish between the reasons based on the endpoint test absence codes we collect from schools (we do not expect any missing data in the codes themselves). Each reason is given in the first column of Table 2, along with the missing data pattern we believe could apply in a reasonable 'worst case' scenario (worst in

[14] A school-level random intercept is included because this will lead to correct standard errors and p-values and for consistency with other analysis models.
[15] This will not replace the primary analysis result, it will be an additional result provided for context.
[16] Indeed, if the values of missing PTM7 scores are systematically different than those that are observed, we should be sceptical as to whether the limited list of available variables given above will account for these differences.
[17] One standard deviation of the PTM7 outcome amongst all pupils with a recorded outcome. This is a somewhat arbitrary choice: it is intended to be large but within the range of the observed data.
[18] For this scenario to be applicable a positive intervention effect of course needs to be observed in the primary analysis.

terms of biasing the primary analysis result). We consider it unlikely that schools would not complete the endpoint PTM7 tests due to delayed intervention delivery, but if it does then this scenario could also be considered as part of the missing data analysis.

We will use the full dataset to impute outcomes for pupils corresponding to pattern b and c according to the assumptions in Table 2; each imputation of PTM7 scores will be edited to reflect the relevant MNAR assumption (e.g. by having one standard deviation subtracted for pattern b). Missing data will be imputed using a two-level normal model with the '2l.lmer' function, using chained equations in the R package 'mice' (Buuren and Groothuis-Oudshoorn, 2011). The imputation model will include all variables from the primary analysis, together with any auxiliary variables associated with missingness of PTM6 or PTM7 scores. Ten datasets will be generated, each using twenty iterations. The mean and standard deviation of imputed variables will be plotted against imputation number, to check for convergence. If means or standard deviations are systematically increasing/decreasing with iteration number, or the lines from different imputations are not 'mixing' (crossing each other), this could indicate non-convergence. Using more iterations may then resolve the issue of non-convergence. The primary analysis model will then be re-run on each of the ten datasets generated and results from each model will be pooled into a single set of estimates and standard errors. This pooled estimate will then be compared to the primary analysis result.

Table 3: reason why a pupils' PTM7 outcome is not available and the corresponding missing data pattern in a reasonable worst-case scenario[19]

Reason for missing PTM7 outcome	Missing data pattern in reasonable worst-case scenario	Explanation
Pupil absent on the day of test	Pattern a) (MAR)	No reason to believe that absence on the day of the test is correlated with worse outcomes, except possibly via persistent absence (we potentially include absence rates in the imputation model, as described above).
Pupil withdrawal	Pattern a) (MAR)	Withdrawals are generally made by parents, presumably due to concerns around privacy or their child's welfare. There does not seem to be any reason to believe this would correlate with attainment.
Pupil did not want to take the test	Pattern b) (MNAR)	Reluctance to take test could indicate a lack of confidence and/or knowledge of test material.
Pupil present but excluded from the test	Pattern b) (MNAR)	Poor behaviour may be correlated with lower attainment. It may also be an expression of reluctance to take the test, as in the box above.
School withdrawal	Pattern c) (MNAR)	School withdrawal may indicate lack of resources (e.g. staff time) to implement the intervention properly or lack of engagement with the intervention.
Pupil left school	Pattern c) (MNAR)	Intervention pupils who leave the school during the trial period will miss some or all intervention sessions (effectively non-compliance).

[19] Patterns a) - c) are not an exhaustive list of missing data patterns. In particular, we do not consider a pattern where intervention pupils with missing outcomes perform worse than control pupils with missing outcomes (but the primary analysis estimate is positive). This pattern might be considered in, for example, a medical trial where side-effects of the treatment are linked to both attrition and poor outcomes, but we do not consider it likely for an educational intervention.

Compliance

Teachers at intervention schools will complete a delivery log of MR sessions. For each session, the unit completed in that session will be recorded (or 'unit not complete' if no unit was completed), together with which pupils attended the session. By combining this information, it is possible to count the number of units attended by each pupil and across how many sessions. Where a unit spans multiple sessions, a pupil's attendance at all those sessions will be required for the unit to be counted. A histogram showing the number of MR units attended by pupils will be included in the final report. There will be three compliance measures for this evaluation, all of which are based on a pupil's attendance of the 12 units delivered by their teacher.

1. Number of units attended by a pupil (continuous measure)
2. Whether a pupil attended 10 or more units (dichotomised variable derived from continuous measure)[20]
3. Whether a pupil attended one or more units (dichotomised variable derived from continuous measure)

For this evaluation, the main compliance measure (1) will be continuous. Compliance measure (2) investigates the impact of attending most units. Compliance measure (3) exists as a lower bound for measure (2), as described below.

For each of these measures 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 PTM7 scores. For instrumental variable analysis to produce unbiased results the 'exclusion restriction' must hold: intervention assignment can only affect PTM7 scores via the chosen compliance measure. One implication of the exclusion restriction is that when dichotomising a continuous measure (here number of units), intervention pupils receive no benefit below the chosen compliance threshold. This is unlikely to hold for compliance measure (2): we expect some benefit for pupils attending fewer than 10 units, which is likely to upwardly bias this result. For this reason we included compliance measure (3) to act as a 'lower bound' for measure (2). Because measure (3) is relatively robust to the exclusion restriction, the unbiased estimate for measure (2) must be higher than the estimate for measure (3) (Gerber and Green, 2012).

None of the compliance measures (1) - (3) involve a teacher's completion of training sessions. We considered including this information, either as a composite measure with pupils, units attended or as a standalone measure. However, it would be difficult to obtain an accurate estimate of the impact of teacher training using instrumental variable methods, as a pupil may benefit from the intervention just by receiving Mathematical Reasoning units, whether or not their teacher completes all the training (i.e. another form of exclusion restriction violation). Instead, we investigate the impact using regression modelling outside of the instrumental variable framework (see RQ 2.6).

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 the first stage,

[20] By default this will be compliance measure (2). However, if fewer than 40% of intervention pupils meet this condition we will instead use the median number of units as the compliance cut-off for measure (2). This is to insure against a scenario where very few pupils meet the compliance condition, which worsens the potential bias from exclusion restriction violations and limits the applicability of results.

the compliance variable will be regressed on the intervention indicator and baseline PTM6 score. This first stage linear regression will be:

$$\text{compliance}_{ij} = \beta_0 + \beta_1 \text{intervention}_j + \beta_2 \text{PTM6}_{ij} + \alpha_j + \epsilon_{ij}$$

In each model, the continuous or dichotomised compliance variable, compliance_{ij}, will be defined based on one of the three measures above and will take the value zero for control pupils (the trial design ensures control pupils cannot receive Mathematical Reasoning). α_j is school-level fixed effect, so will be estimated using one fixed effect parameter per school (minus the reference level). For the second stage PTM7 scores are regressed on each pupil's predicted compliance value $\widehat{\text{compliance}}_{ij}$ obtained from the first stage, in the following linear regression:

$$\text{PTM7}_{ij} = \beta_0 + \beta_1 \widehat{\text{compliance}}_{ij} + \beta_2 \text{PTM6}_{ij} + \alpha_j + \varepsilon_{ij}$$

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 PTM7 scores. For the linear CACE measure, this can be interpreted as the average change in PTM7 score per additional session attended.

Results from both regression stages will be reported for all six models. All instrumental variable analyses 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). The correlation between the intervention indicator and the compliance variables, as well as the F-test for the intervention indicator from the first stage of the two-stage regressions, will be reported.

Additional analysis and robustness checks

Confirmatory factor analysis on the PTM7 subscales

The PTM subscales, which are the outcomes for RQ2.2, were identified by GL Assessment for general usage by mapping assessment questions to national curriculum areas, rather than being developed using a data-driven approach such as exploratory factor analysis and so their psychometric properties are unknown. To investigate the reliability and validity of the PTM7 subscales a confirmatory factor analysis (CFA) will be performed using a four-factor solution, with the items from each subscale loading onto exactly one factor. As the underlying items are binary (mark awarded (1) or not awarded (0)), diagonally weighted least squares will be used for estimation, using the R package 'lavaan' (Rosseel, 2012).

Certain items may be deemed to perform poorly, as identified by low factor loadings and the removal of the item improving overall model fit. These poorly performing items will be removed from their secondary outcome subscale, if this is theoretically justified[21]. Any secondary analysis models that are affected by this change will then be rerun, with the relevant items removed from their PTM7 subscale outcome.

[21] That is, if the evaluation team decides that the scale is still interpretable and measures the same underlying construct when the item is removed. The removal of items will therefore not be decided solely by statistical results from the confirmatory factor analysis.

Until the confirmatory factor analysis is performed, the secondary analysis PTM7 subscales are being proposed under the working assumption that they represent meaningful underlying constructs in the PTM7 data (they are a 'working hypothesis'). If the subscales perform generally poorly in the confirmatory factor analysis, then the secondary analysis will be presented with the appropriate caveats. However, the further modelling described here is intended to provide additional context to the secondary analysis results; it will not cause secondary analysis results to be removed or replaced.

Addition of further covariates to the primary analysis

As described in the sections above, the randomisation for this trial will not be stratified and the primary analysis model will not include any covariates other than baseline PTM6 score. To assess the sensitivity of results to any chance imbalances in pupil-level characteristics[22], we will rerun the primary analysis with additional covariates:

- FSM eligibility
- Whether pupils have SEN
- Whether pupils speaks EAL
- Gender

The model will therefore be:

$$\mathrm{PTM7_{ij}} = \beta_0 + \beta_1 \mathrm{intervention_j} + \beta_2 \mathrm{PTM6_{ij}} + \beta_3 \mathrm{FSM_{ij}} + \beta_4 \mathrm{SEN_{ij}} + \beta_5 \mathrm{EAL_{ij}} + \beta_6 \mathrm{male_{ij}} + \mathrm{b_j} + \epsilon_{ij}$$

Where $\mathrm{FSM_{ij}}$, $\mathrm{SEN_{ij}}$, $\mathrm{EAL_{ij}}$ and $\mathrm{male_{ij}}$ are indicators for whether a pupil is FSM-eligible, has SEN, speaks EAL and is male, respectively (yes (1) or no (0)). If these variables explain a substantial proportion of model variance, on top of that explained by PTM6 score, there may also be an improvement in the precision of the estimate, although previous research suggests this is unlikely (Stallasch *et al.*, 2024 and appendix tables from Singh *et al.*, 2023). All four variables will be added to the model, regardless of the degree of imbalance between the control and intervention groups observed at baseline.

Intra-cluster correlations (ICCs)

The ICC for the primary outcome model will be calculated as the proportion of PTM7 outcome score variance attributable to level 2 (between-school) variation:

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

Here $\boldsymbol{\sigma_B^2}$ and $\boldsymbol{\sigma_W^2}$ 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).

[22] School-level variables were also considered but we decided against their inclusion, as typically these have less potential to explain outcome variance, as reflected by the small intra-cluster correlation coefficients seen in many educational studies.

Effect size calculation

Impact estimates from the models described above will be presented as an effect size, as described by Hedges (2007):

$$\mathrm{ES} = \frac{\hat{\beta}}{\sqrt{\sigma_{\mathrm{B}}^2 + \sigma_{\mathrm{W}}^2}}$$

$\hat{\beta}$ is the coefficient for the binary predictor of interest, typically the intervention indicator, which will be extracted from a conditional model (including any covariates). σ_B^2 and σ_{W}^2 are the between-school and within-school variance from the corresponding empty model (the same outcome but no covariates). To obtain a 95% confidence interval for the effect size, a confidence interval for $\hat{\beta}$ will first be calculated. The end points of this confidence interval will then be divided by the denominator in the above effect size formula. The coefficients for continuous predictors will not be converted into an effect size, they will be presented on their original scale.

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, pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Bishenden, O. (2023) 'RE: Progress Test in Maths in primary schools'.

Bloom, H.S., Richburg-Hayes, L. and Black, A.R. (2007) 'Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions', *Educational Eval;uation and Policy Analysis*, 29(1). Available at: https://journals.sagepub.com/doi/abs/10.3102/0162373707299550.

Bulus, M., Dong, N., Kelcey, B. and Spybrook, J. (2021) 'PowerUpR: power analysis tools for multilevel randomized experiments'. Available at: https://cran.r-project.org/web/packages/PowerUpR/index.html1.

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.

EEF (2022) *Statistical guidance for EEF evaluations*. London. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1699621596 (Accessed: 27 January 2025).

Flemons, L., Smith, A., Morton, C. and Poet, H. (2024) *Effectiveness trial of Mathematical Reasoning - Evaluation Protocol*. London. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/mathematical_reasoning_-_evaluation_protocol.pdf?v=1737980741 (Accessed: 27 January 2025).

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: 27 January 2025).

Gerber, A.S. and Green, D.P. (2012) *Field experiments: design, analysis and interpretation*. London: W. W. Norton & Company.

Hedges, L.V. (2007) 'Effect Sizes in Cluster-Randomized Designs', *Journal of Educational and Behavioral Statistics*, 32(4), pp. 341–370. Available at: https://doi.org/10.3102/1076998606298043.

Hewitt, C.E. and Torgerson, D.J. (2006) 'Is restricted randomisation necessary?', *BMJ*, 332(7556), pp. 1506–1508. Available at: https://doi.org/10.1136/bmj.332.7556.1506.

Rosseel, Y. (2012) 'lavaan: an R package for structural equation modeling', *Journal of Statistical Software*, 48(2), pp. 1–36. Available at: https://doi.org/10.18637/jss.v048.i02.

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Matur, T., Higgins, S., Einbeck, J., Culliney, M. and Demack, S. (2023) *Improving power calculations in educational trials*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/Work_Package_2023-WP6_18_09_2023_FINAL.pdf?v=1696410358 (Accessed: 21 June 2024).

Stallasch, S.E., Lüdtke, O., Artelt, C., Hedges, L.V. and Brunner, M. (2024) 'Single- and multilevel perspectives on covariate selection in randomized intervention studies on student achievement', *Educational Psychology Review*, 36(4), p. 112. Available at: https://doi.org/10.1007/s10648-024-09898-7.

Stokes, L., Hudson-Sharp, N., Dorsett, R., Rolfe, H., Anders, J., George, A., Buzzeo, J. and Munro-Lott, N. (2018) *Mathematical reasoning: evaluation report and executive summary*. Available at:

https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Mathematical_Reasoning.pdf?v=1696414461 (Accessed: 21 June 2024).

The R Foundation (2023) *R: the R project for statistical computing*. Available at: https://www.r-project.org/ (Accessed: 27 January 2025).

Worth, J., Sizmur, J., Ager, R. and Styles, B. (2015) *Improving numeracy and literacy: evaluation report and executive summary*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Oxford_Numeracy_and_Literacy.pdf?v=1696414466 (Accessed: 21 June 2024).

Zeileis, A. (2006) 'Object-oriented computation of sandwich estimators', *Journal of Statistical Software*, 16(9), 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(1), pp. 1–36. Available at: https://doi.org/10.18637/jss.v095.i01.

Appendix A: R randomisation code

```
library(openxlsx)

# 1. Set work directory
setwd("…")

# 2.identify project
project<-"EEMR"

# 3.identify classification: c, r or p
classification<-"C"

# 4. Number of the randomisation: 1st, 2nd, 3rd ...
randomisation<-1
randomisation<-as.character(as.roman(randomisation))

# 5. Load data
Experiment<- read.xlsx("…")

# 6.  Stratified sample?
# 6a. List the stratification variables if Yes
stratify <- "No" # Yes or No
if (stratify =="Yes"){
  stratification<-list() # list the stratification variables here
} else {
  Experiment$stratify_dummy <- 1
  stratification<-list("stratify_dummy")
}
n_strats<-length(stratification)

# 7. identify the cluster variable
cluster<-"NFER_No"

# 8. What time is now? (hh.mm)
time_now<-14.24

aux<-100*trunc(time_now)+100*(time_now-trunc(time_now))
set.seed(aux)
seeds<-sample(1:9999,size=(n_strats+2))

# Keep the original order of the columns
if (stratify=="Yes"){
  originalColOrder<-colnames(Experiment)
} else{
  originalColOrder<-colnames(Experiment[,c(1:ncol(Experiment)-1)])
}

# Adding a variable that will allow for the recovery
# of the original order of the data frame rows later on
Experiment$originalRowOrd<-1:nrow(Experiment)

# Ordering Experiment by cluster
Experiment<-Experiment[order(Experiment[[cluster]]),]

# Assigning a random order to the stratification
rands<-paste("rand",as.character(1:n_strats),sep="_")
```

```
for (i in 1:n_strats){

  aux<-as.data.frame(sort(unique(Experiment[,stratification[[i]]])))
  set.seed(seeds[1])
  seeds<-seeds[-1]

  aux[rands[i]]<-sample(1:nrow(aux))

  Experiment<-merge(Experiment,aux,by.x=stratification[[i]],by.y=colnames(aux)[1])
}

# Randomise by cluster
set.seed(seeds[1])
seeds<-seeds[-1]
Experiment["rand_cluster"]<-sample(nrow(Experiment))

# Reorder the rows of Experiment by rands and rancluster
rands<-c(rands,"rand_cluster")
aux<-do.call(order,Experiment[rands])
Experiment<-Experiment[aux,]

# Assigning Control or Intervention Group
aux<-rep(1:2,times=round(nrow(Experiment)/2))
Experiment$grp<-aux[1:nrow(Experiment)]

rands<-c(rands,"grp")

aux<-data.frame(group=c("control","intervention"))
set.seed(seeds[1])
aux$randgroup<-sample(1:2)

Experiment<-merge(Experiment,aux,by.x="grp",by.y="randgroup")

# Returning the data frame to its original order
Experiment<-Experiment[order(Experiment$originalRowOrd),]

# Removing the variables that are no longer necessary
if (stratify == "Yes"){
  rands<-c("originalRowOrd",rands)} else{
  rands<-c("originalRowOrd",rands,"stratify_dummy")
  }

rands<-which(colnames(Experiment)%in%rands)
Experiment<-Experiment[,-rands]
originalColOrder<-c(originalColOrder,"group")
Experiment<-Experiment[,originalColOrder]
```