Statistical Analysis Plan

Realistic Maths Education (RME)

Evaluator (Sheffield Hallam University):

Principal investigator(s): Sean Demack & Mark Boylan

PROJECT TITLE	Realistic Maths Education
DEVELOPER (INSTITUTION)	Manchester Metropolitan University
EVALUATOR (INSTITUTION)	Sheffield Hallam University
PRINCIPAL INVESTIGATOR(S)	Sean Demack & Mark Boylan
TRIAL (CHIEF) STATISTICIAN	Martin Culliney & Sean Demack
SAP AUTHOR(S)	Sean Demack, Mark Boylan, Martin Culliney & Claire Wolstenholme
TRIAL REGISTRATION NUMBER	ISRCTN21086912
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/realistic-maths-education/

SAP version history

VERSION	DATE	REASON FOR REVISION
1.1 [*latest*]	December 2020	Covid19 and partial closure of schools in England in March 2020 coincided with the last five months of planned RME delivery. Additionally, school closures meant that the planned outcome testing in Summer 2020 was no longer feasible. In response, the following changes have been made: • Outcome testing is rescheduled from Summer 2020 to Spring 2021. The GL PTM13 test will still be used. • The impact of Covid19 resulted in an increased time-lag between end of delivery (Summer 2020) and outcome testing (Spring 2021) - from around one month to between six and nine months. • The RME delivery period remained as specified (Y7 in 2019; Y8 in 2020) but with a loss of teaching time in Summer 2020. An optional 'capstone' module/set of teaching materials for consolidation of RME approaches into Y9 is being made available, along with associated training for Y9 teachers, to try to address the increased time-lag to outcome testing. Three such 90-minute online RME training sessions will be run. • Therefore, there are now two distinct timescales for the evaluation: ○ The RME delivery period: two years; Y7 in 2019 & Y8 in 2020.

		○ The protracted RME evaluation period; 2.5 to 3 years (Y7, Y8 and Y9). • The original SAP stated that pupil/class lists would be collected at three points within the RME delivery period. This has been updated to be two data collection points within the RME delivery period and one data collection point within the protracted (Y9) RME evaluation period.
1.0 [*original*]	June 2019	*[leave blank for the original version]*

Statistical Analysis Plan

Realistic Maths Education (RME)

Evaluator (Sheffield Hallam University):

Principal investigator(s): Sean Demack & Mark Boylan

Education Endowment Foundation

Table of Contents

Tables and Figures

Introduction

Realistic Maths Education (RME) is a pedagogical theory developed in the Netherlands that uses realistic contexts and a notion of progressive formalisation to help the mathematical development of pupils. The RME approach was taken up by Manchester Metropolitan University (MMU) and adapted to the context of the English education system for use in KS3 and KS4 from 2004. MMU's RME approach has undergone independent evaluations using mixed methods research designs (Searle & Barmby, 2012; Hough et al., 2017; Boylan & Jay 2017). This current evaluation represents the first time that RME has been evaluated using a Randomised Controlled Trial (RCT) research design.

Design overview

Trial type and number of arms		Two-arm, multisite four-level clustered randomised controlled trial (MSCRT)[1]
Unit of randomisation		School
Stratification variables (if applicable)		Geographic area, school type and use of setting/streaming
Primary outcome	variable	Maths Attainment
	measure (instrument, scale)	GL Assessment Progress Test in Mathematics, GL PTM13[2]
Secondary outcome(s)	variable(s)	Components of Maths Attainment
	measure(s) (instrument, scale)	GL PTM13 but excluding the timed mental maths items

Sample size calculations overview

Please note: MDES estimates are dependent on how the variance of the primary outcome variable is structured. Within the 3-level CRT design, variance of the primary outcome (GL PTM, maths attainment) will lie between schools; within schools (between classrooms) and finally within classrooms (between individual pupils). The strength of clustering at the school and class levels will determine the smallest effect size that the MSCRT design could detect as statistically significant ($p<0.05$) with a statistical power of 0.80 (the Minimum Detectable Effect Size or MDES). At this stage of the trial we are unclear about the strength of clustering of the primary GL PTM outcome at the school and class levels. Demack (2019) highlighted how the strength of this class level clustering will depend on the use of setting/streaming in Y7 and Y8 maths across the 119 schools in the trial. Within the protocol, a range of estimates were provided that related to three setting/streaming scenarios. The

[1] This can also be viewed as a 3-level CRT blocked by geographical area but is called a 4-level MSCRT by Spybrook et al. (2016)

[2] See www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/

proportion of variance in the GL PTM outcome that is clustered at the school and class levels is measured using Intra-cluster Correlation Coefficients (ICCs):

1. All 119 schools practice mixed ability; School level ICC =0.15-0.20; **class level ICC=0.05**.
2. Some of the 119 schools practice mixed ability others use setting. School level ICC =0.15-0.20; **class level ICC=0.25.**
3. All of the 119 schools practice setting. School level ICC =0.15-0.20; **class level ICC=0.50**.

From data collected prior to randomisation, 106 of the 119 schools in the trial (89%) reported using setting or streaming in Y7 and/or Y8 maths. This leads us to adopt protocol scenario 3 above for the MDES estimates; to assume that half of the variation in the GL PTM outcome will be within schools/between classrooms and 15% to 20% will be between schools. The class ICC of 0.50 agrees with recommendations from Demack (2019) for trials evaluating secondary mathematics programmes and is comparable to what was found by Boylan et al. (2015)[3] with a GL PIM13[4] outcome in a Y8 pupil sample. This means that the proportion of variance in GL PTM that is assumed to be at the pupil level (within classes between pupils) will be between 30% and 35%.

We estimate the correlation between covariates and the outcome to be 0.70 at all three levels (R^2=0.49). This is a conservative estimate that draws on Boylan et al. (2015) who found much stronger correlation between KS2 and GL PIM (12, 13 and 14) at school, class and pupil levels. We have gone with a conservative estimate because Boylan et al. (ibid) is a single study and we feel it preferable to err on the side of caution. Given that the class level clustering will largely be a result of school policies that sort pupils (at least in part) along attainment lines, it is highly likely that the class level correlation will be higher than 0.70. The school level correlation is more of an unknown and given that the Boylan et al. (2015) study included Grammar and secondary modern schools along with comprehensive schools, their estimates for school level explanatory power (R^2=0.88 for Y8) may well be an artefact of their sample. The RME sample only includes comprehensive middle and secondary schools and so is not directly comparable.

Table 1 presents the MDES summary and sample sizes at protocol and randomisation stages. At protocol stage the estimates are based on a predicted number of pupils per class (25) and classes per school (3) whilst at the randomisation stage, the details were taken

[3] For the Y8 cohort, Boylan et al. (2015) reported a class level ICC of 0.47.
[4] GL Progress in Mathematics (PiM) was replaced by GL Progress Test in Mathematics (PTM) in 2015. For Y8 measures (PiM13 and PTM13) a correlation of 0.83 between the two is reported (p12 of GL Assessments PTM technical report).

directly from data collected from schools before randomisation. Please see Appendix I for the MDES formula used.

At both protocol and randomisation stages, the estimated MDES for the primary GL PTM outcome is between 0.21 and 0.23. For the FSM subgroup analyses the MDES estimate is between 0.22 and 0.24.

The impact of Covid 19 on trial sensitivity

The Covid19 pandemic, partial closure of schools, and resulting need to reschedule the outcome testing for the RME evaluation, brought a number of profound issues into the trial. The impact of these issues was in relation to the delivery of RME and the trial sample size.

RME was delivered as intended up to February 2020. However, as schools were closed from March 2020 until the end of the academic year[5], RME delivery (to May/June 2020) was disrupted and a final training day in May was cancelled. In response, during the school closures MMU delivered two sets of training sessions online aiming to support what RME teaching might be occurring either online or in classrooms where a small minority of students were still being taught. However, the classroom engagement with RME materials was close to nil during this time, and online RME teaching very patchy. Almost all schools had neither begun nor delivered the final (10th) module (on functions/algebra), and the preceding (9th) module (on data) had also not been completed by a number of schools.

A number of RME intervention and control schools were unable to continue their involvement with the RME evaluation after Summer 2020. In some instances (8 middle schools), this was because pupils were no longer at the schools they attended in Y7 and Y8. In other instances, this related to schools that felt unable to continue with the evaluation because of the impact of Covid19. To try and limit the number of schools in this second group, additional financial incentives were offered in acknowledgement of the additional time commitment from schools.

At the time of writing (October 2020), of the 119 schools involved at the start of the trial, 59 are confirmed as remaining in the trial during the protracted Y9 evaluation period (34 RME intervention and 25 control schools). Schools known to have dropped out of the trial include the eight middle schools (4 intervention and 4 control); 11 further intervention and two further control schools; a total of 21 (15 intervention and 6 control). It is unclear whether the remaining 39 schools (11 intervention and 28 control) will be included in the outcome testing. Following discussions between MMU, SHU and EEF, an upper estimate of 70 schools was

[5] Schools were closed to most pupils from March 2020 but for children of key workers and other children, schools remained open. For simplicity, when we refer to school closures we mean partial closure - where most pupils did not experience any schooling.

agreed which assumed that 11 additional schools (one intervention and 10 control) would agree to the outcome testing.

Simply assuming a smaller sample size and proceeding to calculate MDES estimates is not technically suitable in this context. This is because the sample size has been determined by something other than randomness (Covid19 and rescheduling testing). However, some indicative MDES estimates are provided for consistency purposes. Please note that these MDES estimates assume that only randomness determined RME intervention and control group membership. Given that this was not the case, readers are advised to exert critical caution in interpreting these estimates.

Table 1: Original Sample size calculations

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.21-0.23	0.22-0.24	0.21-0.23	0.22-0.24
Pre-test/ post-test correlations	level 1 (pupil)	0.70	0.70	0.70	0.70
	level 2 (class)	0.70	0.70	0.70	0.70
	level 3 (school)	0.70	0.70	0.70	0.70
Intracluster correlations (ICCs)	level 2 (class)	0.50	0.50	0.50	0.50
	level 3 (school)	0.15-0.20	0.15-0.20	0.15-0.20	0.15-0.20
Alpha		0.05	0.05	0.05	0.05
Power		0.80	0.80	0.80	0.80
One-sided or two-sided?		Two	Two	Two	Two
Average cluster sizes					
Average classes per school		3	3	2.8	2.8
Average pupils per class		25	5	24.8	5
Average pupils per school		75	15	68.4	15
Number of schools	intervention	60	60	60	60
	control	59	59	59	59
	total	119	119	119	119
Number of classes	intervention	180	180	159	159
	control	177	177	169	169
	total	357	357	328	328
Number of pupils	intervention	4,500	900	4,011	900
	control	4,425	885	4,131	885
	total	8,925	1,785	8,142	1,785

Indicative MDES estimates (Covid19)

To calculate the indicative MDES estimates, correlations, ICC values, statistical significance and power all remain as shown in Table 1. The average classes per school, pupils per class and pupils per school also remain as shown in Table 1.

In calculating these indicative MDES estimates, two sample size estimates have been used. First, the 59 schools confirmed at the time of writing. Second, an upper estimate of 70 schools that assumes one additional intervention and 10 additional control schools will be confirmed.

The 59 confirmed schools are comprised of 34 intervention and 25 control schools and so the proportion of schools receiving the intervention is 0.57 [P=34 / (34+25)]. This leads to an indicative MDES estimate of 0.31-0.33 sds for all pupils and 0.32-0.34 for FSM pupils.

For the upper estimate of 70 schools, the assumption is that this will be comprised of 35 intervention and 35 control schools (and so P=0.50). This leads to an indicative MDES estimate of 0.28-0.30 sds for all pupils and 0.29-0.30 for FSM pupils.

On first reading, these estimates might suggest that the loss in sensitivity is not too dramatic. However, it is important to note that these are post-hoc indicative MDES estimates following the drop in sample size because of the Covid19 pandemic and therefore cannot have the same statistical validity as the original MDES estimates. This is because something other than randomness has determined the sample size. Whilst sample attrition is not uncommon within educational trials, the attrition resulting from Covid19 is more profound and included some systematic sample losses, such as the eight middle schools unable to continue in the evaluation as participating pupils have progressed into Y9 and therefore left the school.

Primary Outcome

Under normal circumstances, the most suitable GL PTM test for pupils in the Spring term of Y9 would be PTM14. According to GL, the PTM14 test is designed for pupils in Spring/Summer terms in Y9 or Autumn term in Y10[6]. The original primary outcome for the RME evaluation was GL PTM13 which is designed for pupils in Spring/Summer terms in Y8 or in Autumn term in Y9. The closure of schools during the Spring/Summer terms of Y8 and ongoing need for pupils and teachers exposed to Covid19 to stay away from schools since the start of Y9, led to the decision to maintain GL PTM13 as the primary outcome rather than change to PTM14. There are three (related) reasons for this decision. First, the PTM14 test includes items that test mathematical knowledge and understanding for the curriculum following the Y7 & Y8 RME delivery period. Second, given that most pupils experienced

[6] See https://www.gl-assessment.co.uk/support/ptm-product-support/progress-test-in-maths-test-level-age-guide/

little/ no schooling between March and July 2020, the progression into the Y9 mathematics curriculum is likely to be delayed. Third, Y9 pupils would be at the very start of the recommended age range for PTM14 and likely to have had limited exposure to much of the curriculum it is designed to test, even without taking account of the impact of Covid19. We also see an additional benefit for keeping PTM13 that relates to the Covid19 context; identifying gaps in mathematical learning in Y8 and earlier is likely to be of greater value to intervention and control schools as opposed to PTM14 confirming gaps in the Y9-10 curriculum not yet covered.

Analyses

Research Questions for the Impact Evaluation of RME

The original objectives for the impact analyses for the RME impact evaluation were to answer six research questions. These questions have been adapted so that they now focus on the RME delivery period but also acknowledge the protracted RME evaluation period[7]. Additionally, one further question (RQ2a) has been developed to look closer at the protracted RME evaluation period.

Primary question

RQ1. Does the two-year RME intervention improve pupil attainment in mathematics over the protracted 2.5 to 3 years as measured by the GL Assessment Progress Test in Mathematics (PTM13) in general and specifically for disadvantaged pupils?

Secondary questions

RQ2. What is the impact of RME on mathematics attainment for pupils known to have been taught by one of the nominated RME teachers in the RME delivery period (Y7 and Y8)

RQ2a. What is the impact of RME on mathematics attainment for pupils known to have been taught by one of the nominated RME teachers through the protracted RME evaluation period (Y7, Y8 and Y9).

RQ3. What is the impact of RME on mathematics attainment for pupils who experience partial intervention effects due to pupil and/or teacher movement during the RME delivery period (Y7 or Y8)?

[7] The original GL PTM testing was scheduled for Summer 2020. Following Covid19, testing was rescheduled to Spring 2021. Most of the RME evaluation took place prior to March 2020 when schools were closed. Some RME training took place online between March 2020 and the end of the academic year but engagement with RME classroom materials was severely limited because of the closures. This means that there is a protracted time lag between the end of RME delivery and outcome testing, from an original assumption of one month to between six and nine months.

RQ4. What is the effect on attainment of components of the GL PTM13 aligned with the RME intervention?[8]

RQ5. What is the relationship between mathematics attainment and fidelity of implementation?

RQ6. What is the impact of RME on mathematics attainment for pupils taught by teachers identified as implementing the RME evaluation with high fidelity?

Intention to Treat (ITT) and other pupil samples

Figure 1 is a visual representation of the pupil cohort for the 119 schools in the RME impact evaluation. The orange box represents the whole pupil cohort (Y7 in 2019 to Y9 in 2021). Within this are four key pupil samples. First, the 'Intention to Treat' (ITT) pupil sample is shown as a blue ellipse. This ITT pupil sample is comprised of pupils in both intervention and control schools who were listed as being taught by one of the teachers nominated prior to randomisation.

The main (headline) impact analyses for the GM PTM will use this ITT sample (RQ1). Assuming that this sample remains intact through the trial period, the ITT analysis is the most robust for drawing causal conclusions about the impact of the RME evaluation on maths attainment.

However, within the context of widespread use of setting and streaming in secondary mathematics in England (Demack, 2019; Dracup, 2012), the integrity of this ITT sample may be undermined. For example, if notable pupil and/or teacher movement between classes takes place following randomisation (if setting/streaming is introduced post-randomisation whole classes might be reconfigured, or if teachers and/or pupils move between classrooms for other reasons), this may mean that some pupils within the ITT sample experience little or no RME teaching. This may also mean that some pupils outside of the ITT sample experience a lot of RME teaching. The evaluation design for RME has attempted to build in an approach that can explore this (and this links to RQ2 and RQ3).

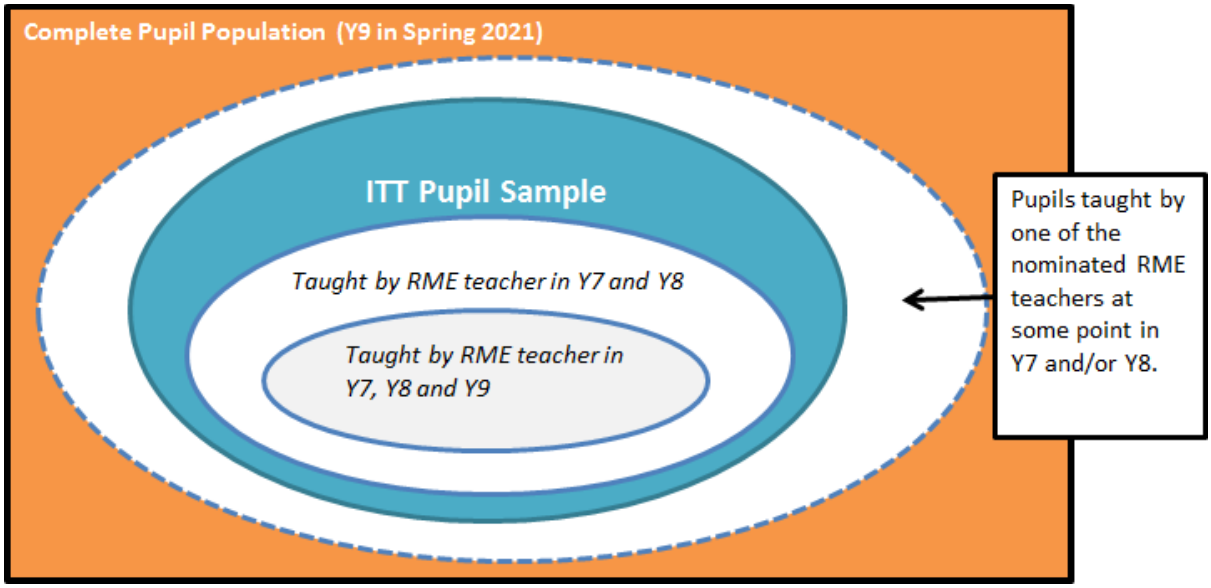
To enable these follow-on exploratory impact analyses, pupil class lists will be collected from both intervention and control groups. The pre-Covid protocol and SAP specified that these class list details would be collected on three occasions during the RME delivery period; before randomisation in September 2018 at the start of Y7; mid-trial at the start of Y8 in Sept/Oct 2019 and end of trial in June 2020. In reality, data was collected twice during the RME delivery period; before randomisation and in Y8 prior to Covid19. For RQ2, the ITT sample will be limited to pupils taught by one of the nominated maths teacher on both

[8] RQ4 links to two exploratory analyses; the first will explore attainment using PTM13 items excluding mental mathematics and the second will explore attainment using a selection of PTM13 items selected by MMU to be closely aligned to the RME programme.

occasions (Y7 and Y8). For RQ3, the ITT sample will be expanded to include all pupils taught by one of the nominated maths teacher on either occasion (Y7 or Y8).

To examine pupil movement between Y8 and Y9 during the protracted evaluation period, pupil class lists will be collected for a third time in Autumn 2020. For RQ2a, the ITT sample will be limited to pupils taught by one of the nominated maths teacher on all three occasions (Y7 to Y9).

Figure 1: Pupil Samples and Subsamples for the RME impact evaluation



For RQ2, the analyses will be limited to a subsample of pupils known to have been taught by one of the nominated RME teachers throughout the RME delivery period (Y7 and Y8). This sample is a subset (or subsample) of the ITT sample and is illustrated in Figure 1 by the white ellipse within the ITT sample. Specifically, using pupil class lists collected at two points in the trial period (start Y7, mid-Y8); the sample will be limited to pupils that were located in a classroom taught by an RME teacher on both occasions. For these exploratory follow-on impact analyses, both RME intervention and control samples will be restricted in this way.

For RQ2a, the analyses will be limited to a subsample of pupils known to have been taught by one of the nominated RME teachers throughout the protracted RME evaluation period (Y7, Y8 and Y9). This sample is a subset of the RQ2 sample and is illustrated in Figure 1 by the smallest grey ellipse within the RQ2 sample. Specifically, using pupil class lists collected at three points in the trial period (start, mid-Y8, mid-Y9), the sample will be limited to pupils that were located in a classroom taught by an RME teacher on all three occasions. For these exploratory follow-on impact analyses, both RME intervention and control samples will be restricted in this way.

For RQ3, the pupil class lists will be used to identify a sample of pupils who were in a class taught by an RME nominated teacher for at least one of the two data collection points[9] during the RME delivery period (Y7 or Y8). This sample will contain the ITT sample but is likely to be larger[10]. At the same time, this pupil sample will be smaller than the total pupil cohort. This sample is shown in Figure 1 as the larger white ellipse that contains the ITT pupil sample.

The ITT pupil sample will be used for RQ4 to measure impact of RME on secondary outcome (GL PTM excluding the mental maths items).

For RQ5 the relationship between fidelity to RME and maths attainment will be examined. These analyses will only include pupils within RME intervention schools. Fidelity to RME will be measured on three levels: pupil, teacher and school.

At the pupil level, the relationship between exposure to RME (as measured by the class list data collection) and the GL PTM outcome will be examined. These analyses will focus first on the RME delivery period (Y7 and Y8) but also examine the protracted RME evaluation period (Y7, Y8 and Y9).

For the RME delivery period, a three-point scale will be used that will measure whether a pupil was in a class taught by an RME teacher in Y7 and Y8 (=2, representing best evidence that a pupil had consistent exposure to RME during the delivery period); whether a pupil was in a class taught by an RME teacher in either Y7 or Y8 (=1) or whether a pupil was not in a class taught by an RME teacher in either Y7 or Y8 (=0, representing best evidence of a pupil receiving no exposure during the delivery period). If evidence is found to show that exposure to RME is correlated with maths attainment, follow-on analyses will disaggregate the central value (1) to examine exposure in Y7 and in Y8 separately.

For the protracted RME evaluation period, the focus will be on the subsample of pupils taught by an RME teacher in Y7 and Y8 (=2 on previous scale). This subsample will be disaggregated to show pupils who continued to be taught by an RME teacher in Y9 (=1) and those who did not (=0).

At the teacher level, three dimensions of fidelity will be used; attendance of RME training days; RME material coverage & RME curriculum time- all during the RME delivery period.

[9] This is a change from the original design that specified three data collection points within the RME delivery period. A third data collection point will be undertaken in the protracted RME evaluation period (Y9).

[10] In addition to pupils within the ITT sample, pupil and/or teacher movement will mean other (non-ITT) pupils may experience some RME within their maths classes. The only instance where this (RQ3) and the ITT sample would be equivalent would be when there is zero pupil/teacher movement between classes in the trial period.

At the school level, the total number of times a school sent two or more teachers to one of the RME CPD days during the RME delivery period will be used.

For both teacher and school levels, the fidelity dimensions relate just to the RME delivery period (Y7 and Y8).

Whilst fidelity to RME will be examined at pupil, teacher and school levels, compliance will focus only on the pupil and teacher level dimensions of fidelity.

For RQ6, the pupil- and teacher-level dimensions of fidelity to RME will be used to identify pupils located in high fidelity RME classrooms. Specifically, these would be pupils located in classrooms 'consistently' taught by RME teachers during the delivery period[11]. These RME teachers will have attended a minimum of four of the eight RME training days, covered a minimum of six of the ten RME modules[12] using a minimum of 10 weeks maths curriculum time (at least one fifth of the total pre-Covid maths curriculum in Y7 and Y8)[13]. This variable will be used as the 'compliant'(=1) / 'non-compliant (=0) outcome in the first stage of the two stage least squares approach for estimating the Compliers Average Causal Effect (CACE)(see below).

ITT (headline) analysis of primary outcome (RQ1)

Table 2 below summarises the planned headline ITT impact analyses for the primary outcome of the RME evaluation. A multilevel approach will be taken, with pupils clustered into classes and classes clustered into schools (3-level random intercepts multilevel models). Multilevel linear regression models will be constructed for the GL PTM primary outcome using the STATA software version 15. KS2 maths attainment will be used as the baseline covariate. The first model will only include the school level group identifier (an outcome only model)[14]. The second model will also include KS2 maths attainment as a covariate at the pupil, class and school level[15]. The final model will also include the three

[11] In some cases this will be the same maths teacher in both Y7 and Y8; here the 'teacher level' dimension of fidelity will relate to a single teacher. When the original RME teacher is replaced by a new teacher (e.g. at the end of Y7, if a teacher leaves a schools), the 'teacher level' dimension of fidelity will relate to both teachers (i.e. their combined RME attendance, module use and curriculum time).
[12] This threshold has been reduced from seven to six modules to reflect the impact of Covid19 on classroom time and only includes the 10 original modules (not the 'capstone' module).
[13] This threshold has been reduced from 12 to 10 weeks to reflect the impact of Covid19 on curriculum time.
[14] A null (or empty) model will also be fitted in order to obtain the unconditional variance used to calculate the Hedges g effect size statistic, see Appendix II.
[15] These will be centred so that the school level will be centred on the (grand) mean for all 119 schools; the class level will be centred around the school mean; the pupil level will be centred around the class mean.

variables[16] used within the stratified randomisation (geographical hub area, setting/streaming dummy and middle/secondary dummy[17]). This final model will be used for the headline ITT impact analysis for the GL PTM primary outcome.

Follow-on ITT analyses will focus on the impact of the RME programme on maths attainment for pupils ever classed as FSM[18]. The same three model stages used for the headline ITT analyses will be used for these analyses.

For each model, the coefficient of the school-level dummy variable used to distinguish 'intervention group' pupils within the 60 schools who will receive the RME programme from 'control group' pupils within the 59 control schools. This coefficient will be converted into Hedges' *g* effect size statistics with 95% confidence intervals. Appendix II provides more technical detail on the multilevel model that will be used for the ITT (headline) analyses of the GL PTM primary outcome and how the Hedges' *g* effect size statistic will be calculated.

Table 2: Summary of analysis plan for ITT impact analyses for primary outcome (RQ1)

Primary Outcome: GL Progress Test in Maths (PTM13) taken in Spring 2021.

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (class) Variables*	*Level 3 (school) Variables*
ITT sample	-	-	• Group (1=RME; 0=control)
ITT sample	KS2 Maths (class centred)	Mean KS2 Maths (school centred)	• Group (1=RME; 0=control); • Mean KS2 Maths(Grand mean centred)
RQ1 Final (headline) Analysis ITT sample	KS2 Maths (class centred)	Mean KS2 Maths (school centred)	• Group (1=RME; 0=control); • Mean KS2 Maths (Grand mean centred) Stratification variables: • Geographical hub area • Setting/streaming dummy • Secondary/Middle dummy

[16] This will actually be seven binary variables; five to represent the six geographical hub areas, one for school phase (secondary / middle) and one for use of setting/streaming (yes/no)

[17] Please see Appendix III for a school level comparison of the RME intervention and control groups at baseline across these three controls. At the school level, the randomisation resulted in good balance across the three controls.

[18] The NPD EverFSM_6 variable will be used

Follow-on exploratory analyses of the primary outcome during the RME delivery period (RQ2 & RQ3)

For RQ2 and RQ3, exploratory analyses will be undertaken with the sample of pupils identified as being taught by an RME teacher in both Y7 and Y8. The same modelling stages shown in Table 2 for the ITT analyses will be repeated for this reduced sample.

Follow-on exploratory analyses of the primary outcome during the protracted RME evaluation period (RQ2a)

For RQ2a, added in response to the trial extension due to Covid19, exploratory analyses will be undertaken on the sample of pupils identified as being taught by an RME teacher in Y7, Y8 and Y9. The same modelling stages shown in Table 2 for the ITT analyses will be repeated for this reduced sample.

Follow-on exploratory analyses of secondary outcomes (RQ4)

The GL PTM primary outcome includes a 'fixed time' mental maths section that runs counter to the RME ethos which encourages pupils to take time in order to develop their depth of mathematical understanding. Therefore, the first secondary outcome for the RME evaluation will be the GL PTM outcome with all 20 of the mental maths items removed. The same modelling stages shown in Table 2 for the primary analyses will be undertaken using the same ITT sample.

The second secondary outcome for the RME evaluation focuses in on the 23 PTM items excluding the 20 mental maths items. MMU selected a subset of items theorised to have a close alignment with the RME programme. This process resulted in selecting 11 of the 23 PTM13 items will be used to derive this secondary outcome (please see Appendix IV for more detail on this selection). The same modelling stages shown in Table 2 for the primary analyses will be undertaken using the same ITT sample.

Fidelity analyses (RQ5)

RQ5 focuses on the relationship between fidelity to RME and pupil outcomes. In these exploratory analyses, only the RME intervention schools, classes and pupils will be included. These analyses will not examine the causal impact of RME on maths attainment. Instead, the fidelity analyses will provide descriptive statistical detail on the relationship between fidelity to RME and maths attainment.

Fidelity to RME will be measured at three levels; pupil, class (teacher) and school.

At the pupil level, a 3-point scale will identify whether a pupil was located in a maths classroom taught by an RME teachers during the RME delivery period (Y7 and Y8)[19].

Drawing on IPE data, the following key teacher-level dimensions of fidelity to RME will be included as explanatory variables:

- Attendance of RME training days (number of days; maximum = 8)
- Coverage of RME teaching materials (number of modules; maximum = 10)
- RME curriculum time (percentage of maths curriculum)

The pupil and three teacher-level dimensions of fidelity to RME will be used to estimate the Compliers Average Causal Effect (CACE) of RME (see below for RQ6).

One school-level dimension of fidelity will be also examined. This variable is exploratory and will not be used for the CACE estimate.

- Dual attendance of RME CPD training during the RME delivery period (Y7 and Y8). This will measure whether schools send one (=0) or more than one (=1) teacher to an RME CPD day. The final measure will tally the number of times a school sent two or more teachers to an RME CPD day over the two years of the evaluation at the school level (maximum = 8).

Table 3 summarises the planned fidelity analyses for the RME evaluation. Multilevel analyses will be used to explore the relationship between fidelity to RME and the primary outcome.

Keeping the pupil, teacher and school level components of fidelity separate in these analyses reflects the exploratory context. This approach provides greater clarity in interpreting (and communicating) how different aspects of fidelity to RME are associated with the PTM outcome. We will also examine the strength of association between the five measures of fidelity as they may be highly correlated. For these exploratory analyses into the relationship between fidelity to RME and the PTM14 outcome, a single pupil-level measure of 'compliance' to RME will be derived within the Complier Average Causal Effect (CACE) analyses (see below).

[19] This may be the same maths teacher throughout the trial period (Y7 to Y8; Y7 to Y9) or when one RME maths teacher is replaced by another RME maths teacher during the trial period).

Table 3: Summary of analysis plan for fidelity analyses for primary outcome (RQ5)

Primary outcome: GL Progress Test in Maths (PTM13) taken in Spring 2021.

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (class) Variables*	*Level 3 (school) Variables*
RQ5 ***Fidelity Analyses [pupil level only]*** RME Intervention school ITT sample	KS2 Maths (class centred) Pupil Level Exposure during RME delivery period (Y7 & Y8) (0; 1; 2)	Mean KS2 Maths (school centred) Teacher-level dimensions of fidelity: • Attendance • Module use • Curriculum time	• Mean KS2 Maths (Grand mean centred) Stratification variables: • Geographical hub area • Setting/streaming dummy • Secondary/Middle dummy School-level dimension of fidelity: • Number of times a school sent 2+ teachers to RME training days

Compliers Average Causal Effect (CACE) analyses (RQ6)

The purpose of the CACE analyses is to estimate the impact of RME for pupils deemed to have 'complied' with the intervention during the delivery period (Y7 and Y8). In this instance 'compliance' relates to both pupils and their maths teachers and to dosage/exposure and fidelity to the intervention. Specifically, to be considered 'compliant', a pupil will need to have been located *consistently* in a maths class taught by an RME teacher who attended at least four of the eight RME training days, used RME in maths lessons for at least 10 weeks (20% or more of total available pre-Covid maths teaching time in Y7 and Y8) and covered at least six of the 10 RME modules. Whilst these thresholds are at two levels (pupil and teacher), the final compliance measure will be a binary variable at the pupil level (1=complied; 0=did not comply). This binary measure will be used to estimate the CACE for the RME intervention.

CACE will be estimated using two stage least squares (2SLS) regression (Gerber & Green, 2012). The first stage will model the pupil-level binary compliant variable using the same explanatory variables listed in Table 2 for the headline ITT analyses of PTM along with additional school level items available via the school census. This model will be used to generate predicted compliance (1 or 0). The second stage model will use the predicted compliance in place of the group variable in Table 2 in order to generate the CACE estimate for RME; the mean difference (in PTM) for pupils predicted (in stage 1) to have complied with the RME intervention compared with pupils predicted not to have complied. 'STATA IVRegress' will be used for this Instrumental Variable approach for estimating CACE.

Table 3: Pupil and Teacher-level fidelity dimensions and thresholds used for the CACE estimate

Component	Data	Maximum	Minimum Threshold	Notes
Pupil-Level (consistent exposure to RME in the classroom during RME delivery period)	Pupil / Teacher class lists collected prior to randomisation and mid Y8 (pre-Covid)	2 Instances	2 instances	Pupil located in class taught by an RME teacher on both class lists.
Teacher-Level 1. (Professional development attendance)	MMU attendance list	8 days	4 days	Half day or more counted as 1 **Unchanged from original SAP.**
Teacher-Level 2. (RME curriculum time)	Percentage of school lessons used based on number of maths lesson per week collected from PD day survey - Collected via annual teacher surveys.	20 weeks out of 60 - based on 2 weeks per 10 modules	20% of pre-Covid maths available teaching time in Y7 & Y8. (60 % of RME 17 weeks … 10 weeks or more RME)	Teacher self-report that a lesson as a RME module lesson counts as 1 lesson **Reduced from 12 weeks to acknowledge reduced classroom time following school partial closures.**
Teacher-Level 3. (RME module coverage)	Number of modules and what materials used. Collected via annual teacher surveys.	10 modules	6 modules	The amount of material/amount of teaching in each module to count as coverage to specified in the SAP **Reduced from 7 modules to acknowledge reduced curriculum time following school partial closures.**

Please note that these thresholds relate only to the RME delivery period (Y7 and Y8) and not the protracted RME evaluation period (Y9). Whilst a large majority of the RME programme had been delivered prior to March 2020, the thresholds have been adapted to account for the reduced mathematics curriculum time between March 2020 and the end of the academic year. Engagement and fidelity in Y9 will be examined as part of the IPE but will not be included in fidelity or CACE analyses.

Imbalance at baseline

The baseline balance between the RME intervention and control group samples will be examined at the school level with respect to the three variables used to stratify the randomisation (geographical hub, school phase, use of setting/streaming) and most recent OFSTED rating prior to randomisation. At the pupil level, FSM status and KS2 maths attainment for pupils in RME intervention and control schools will be statistically compared. For the scale KS2 maths covariate, the difference will be converted into a Hedges g effect size. For categorical variables, counts and percentages across categories will be reported.

Additionally, data from the pre-randomisation teacher survey will be used to examine the balance at the teacher level at baseline.

The school, teacher and pupil level analyses examining the baseline balance will be replicated with the final ITT (headline) sample for the GL PTM primary outcome.

Missing data

The baseline and ITT samples will be compared to help illustrate the impact of missing data. This will be done descriptively by examining missing cases across categories of variables included in the ITT analysis. Reasons for any missingness will be summarised and a multilevel logistic regression model (1=in ITT model; 0=not in ITT model) will examine whether missingness is associated with school and/or pupil level covariates including KS2 attainment, FSM status, geographical hub area, school phase and use of setting/streaming.

If over 5% of cases in the baseline sample are missing from the headline ITT analysis, we will adopt the following approach for imputing missing data:

- Stage 1: A series of binary variables will be generated for all variables in the final ITT analysis that measures whether a case is missing (=1) or not (=0). Logistic regression will be used to examine whether missing data can be statistically accounted for using the other variables in the ITT analysis with additional school-level variables included[20]. The purpose of this first stage is to assess whether it is reasonable to assume that the data is missing at random (MAR). When variables are found to account for a statistically significant proportion of variation in missing data, we will proceed to one of the next two stages.
- Stage 2: For instances where data is only missing in the PTM outcome, we will add any additional covariates that were found from stage 1 to the final ITT model and re-estimate the effect size.

[20] This will include school level data from the baseline 2018/19 academic year such as %FSM; %EAL; %SEN; school size.

- Stage 3: For instances where data is missing across covariates other than the PTM outcome and where stage 1 does identify variables that account for variation in this missing data, we will construct a Multiple Imputation model using all variables listed for stage 1. The Multiple Imputation model will be estimated using 'STATA MI' to create 20 imputed data sets. These imputed data sets will be used to re-estimate the effect of RME and the standard error (Rubin, 2004).

Intra-cluster correlations (ICCs)

The pre-test for RME will be KS2 maths attainment (Y6, age 10/11) and the post-test will be the GL PTM 13 (Y8, age 12/13). For both pre and post-test, ICCs at the school and class levels will be estimated using a null (empty) 3-level multilevel variance components model. Within the analyses, a table will present the variance decomposition for the three levels (school, class and pupil) along with the ICC estimates.

Additional Sensitivity Analyses

In addition to the missing data sensitivity analyses, the main (headline) ITT analyses will be repeated but ignoring clustering at the class level. The purpose of doing this is to examine whether the impact finding differs if class level clustering is ignored and to inform the design of future educational trials. For these analyses, the variance structure of the null models (3-level and 2-level) will be compared and ICC values calculated. Additionally, the covariate explanatory power (provided by the KS2 Maths pre-test) will be estimated for both the 3-level and 2-level models. Finally, if findings from the 2-level model contrast with what is found with the 3-level model, this will be discussed in the final report.

Additional Exploratory Analyses

The GL PTM outcome is measured across four maths dimensions: fluency in facts and procedures, fluency in conceptual understanding, mathematical reasoning and problem solving[21]. These four dimensions of PTM have not been psychometrically designed as validated 'standalone' subscales, and so are not suited for use as outcomes for the impact analyses in the RME evaluation. However, we will undertake exploratory analyses of the two PTM dimensions that do not draw on the mental maths aspects of PTM (mathematical reasoning and problem solving). As exploratory analyses, the analyses of the mathematical reasoning and problem solving dimensions of PTM will be included as an appendix in the final evaluation report for the RME programme.

Hedges g effect size

The impact of RME on pupil maths attainment will be measured using the Hedges g effect size statistic.

[21] See https://www.gl-assessment.co.uk/media/1385/ptm14-links-to-nc.pdf

The impact of Covid19, reduction of RME curriculum time and loss of schools will reduce the validity of drawing causal conclusions from this impact evaluation. Whilst attempts have been made to minimise the impact of Covid19, the delivery of RME and the randomisation have both been undermined. Therefore, caution should be exerted in interpreting findings from this evaluation.

Hedges g standardises the difference between the attainment of pupils in RME schools and pupils in control schools into units of standard deviations. As specified in the EEF analyses guidance, the unconditional variance will be used to obtain the standard deviation. Specifically, the variance in the GL PTM13 outcome that is clustered at school, class and pupil levels will be used, as set out in equation II.4:

Equation II.4 $$ES = \frac{(T-C)_{adjusted}}{\sqrt{\delta^2_{sch}+\delta^2_{class}+\delta^2_{pup}}}$$

Where:

- δ^2_{sch} is the school level variance, δ^2_{class} is the class level variance and δ^2_{pup} is the pupil level variance for the GL PTM13 outcome from the empty/null multilevel model.
- $(T-C)_{adjusted}$ is the mean difference between the attainment of pupils in RME schools and pupils in control schools in the original raw (GL PTM) units. This is obtained from the coefficient for the school-level 'group' variable from the final (headline) analyses.

References

Boylan, M. and Jay, T. (2017) Independent evaluation of 'Investigating the impact of Realistic Mathematics Education approach on achievement and attitudes in Post-16 GCSE mathematics resit classes' In Hough, S Solomon, Y Dickinson, P and Gough, S (2017) *Investigating the impact of a Realistic Mathematics Education approach on achievement and attitudes in Post-16 GCSE resit classes*. Nuffield.

Boylan, M., Demack, S., Willis, B., Stevens, A., Adams, G. & Verrier, D. (2015) Multiplicative reasoning professional development programme. https://www.gov.uk/government/publications/multiplicative-reasoning-professional-development-programme

Demack, S. (2019) Does the classroom level matter in the design of educational trials? EEF Research Paper No. 3 available at https://educationendowmentfoundation.org.uk/public/files/Publications/Does_the_classroom_level_matter.pdf

Dracup, T. (2014) The Politics of Setting. Available at https://giftedphoenix.wordpress.com/2014/11/12/the-politics-of-setting/

Gerber, A., & Green, D. (2012). Field Experiments: Design, analysis and Interpretation. W.W. Norton & Company.

Hedges and Hedberg (2013) Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster randomised experiments in education. Evaluation Review 37(6) pp445-489.

Hough, S Solomon, Y Dickinson, P and Gough, S (2017). *Investigating the impact of a Realistic Mathematics Education approach on achievement and attitudes in Post-16 GCSE resit classes*. MMU/Nuffield.

Jo, B., Asparouhov, T., Muthen, B.O., Ialango, N.S. & Brown, C.H. (2008) Cluster Randomized Trials with Treatment Noncompliance. Psychology Methods 13(1) pp1-18

Kelcey, B., Spybrook, J., Phelps, G., Jones, N. & Zhang, J. (2017) Designing large scale multisite and cluster randomized studies of professional development. The Journal of Experimental Education, 85(3) pp389-410

Searle, J. and Barmby, P (2012) *Evaluation Report on the Realistic Mathematics Education Pilot Project at Manchester Metropolitan University.* Durham: Durham University.

Spybrook, J., Shi, R., Kelcey, B. (2016) Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. International Journal of Research and Method in Education. 39 (3) pp255-267

Appendix I Realistic Maths Power Analysis

From Kelcey et al (2017), the Minimum Detectable Effect Size (MDES) for a 3-level CRT is

$$MDES_{3LCRT} \sim M_{K-L-2}\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{ICC_{sch}(1-R^2_{sch})}{K}+\frac{ICC_{class}(1-R^2_{class})}{JK}+\frac{(1-ICC_{sch}-ICC_{class})(1-R^2_{pup})}{nJK}}$$

From Spybrook et al., (2016), the MDES equation for a 4-level MSCRT assuming zero effect size variability across clusters and including covariate explanatory power at class and pupil levels is:

$$MDES_{4LMSCRT} \sim M_{(M(K-L-2)}\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{ICC_{sch}(1-R^2_{sch})}{MK}+\frac{ICC_{class}(1-R^2_{class})}{MKJ}+\frac{(1-ICC_{sch}-ICC_{class})(1-R^2_{pup})}{MKJn}}$$

It can be useful to re-organise this equation following Hedges & Rhoads (2010)…

$$MDES_{4LMSCRT} \sim M_{(M(K-L-2)}\sqrt{\frac{1}{P(1-P)MKJn}}\sqrt{1+(Jn-1)ICC_{sch}+(n-1)ICC_{class}-\left[R^2_{pup}+\left(JnR^2_{sch}-R^2_{pup}\right)ICC_{Sch}+\left(nR^2_{Class}-R^2_{pup}\right)ICC_{Class}\right]}$$

Where…

- P is the proportion of schools who receive the intervention (=0.50)
- R^2_{Sch} is the school-level covariate explanatory power (=0.49)
- R^2_{class} is the class-level covariate explanatory power (=0.49)
- R^2_{pup} is the pupil-level covariate explanatory power (=0.49)
- ICC_{Sch} is the school level Intra Cluster Correlation coefficient (=0.15 to 0.20)
- ICC_{class} is the class level Intra Cluster Correlation coefficient (=0.50)
- M is the number of geographical sites (=6)
- K is the number of schools per site (=20)
- J is the number of classes per school (=3)
- n is the number of pupils per class (=25)
- L is the number of school level covariates (=9)
- $M_{(M(K-L-2)}$ is the t-distribution multiplier with M(K-L-2) (54) degrees of freedom. Assuming a two-tailed test with a statistical significance of 0.05 (α/2=0.025) and statistical power of (1-β=0.80). M_{54} = 2.8532.

In the protocol for RME MDES estimates for relatively weak and strong clustering of the GL PTM13 outcome at the class level were provided. Drawing on data provided by the 119 schools, 106 (89%) reported to use setting or streaming in Y7 and/or Y8 maths. We therefore assume that class level clustering will be relatively strong (Class ICC = 0.50) and for the SAP have dropped the MDES estimate that assumes weak class-level clustering. Clustering at the school level has been allowed to vary between (school level ICC=) 0.15 and 0.20

- Weakest clustering ($ICC_{Sch} = 0.15$; $ICC_{class} = 0.50$); $MDES = 0.211 \sim 0.21$
- Strongest clustering ($ICC_{Sch} = 0.20$; $ICC_{class} = 0.50$); $MDES = 0.226 \sim 0.23$

Appendix II Realistic Maths Multilevel Analyses & Calculation of Hedges g effect size

Overview

This Appendix provides additional details for the planned ITT analyses of the primary (headline) outcome (GL PTM13) for the RME efficacy trial. Specifically, this Appendix includes:

- Specification of the multilevel regression model
- Example STATA code that will be used to fit the multilevel model
- Detail on how the Hedges g effect size statistic will be calculated,

Specifying the multilevel analyses

As shown in Appendix I, the RME efficacy trial had a 3 level CRT research design that was blocked across (six) geographical hub areas. In addition to geographical hub area, two further stratification variables were used in the randomisation; whether a school was a secondary (1) or a middle (0) school and whether a school reported to have a policy of setting or streaming in Y7 or Y8 maths (1) or not (0). Finally, the design assumed that a pre-test covariate (KS2 Maths) would be included at all three levels (pupil, class and school).

To avoid multicollinearity between the three KS2 maths covariates, they will all be centred as outlined by Hedges and Hedberg (2013). Specifically, this means that:

- Pupil level KS2 maths attainment will be centred around the class-level mean KS2 maths attainment.
- Class level KS2 maths attainment will be centred around the school-level mean KS2 maths attainment.
- School level KS2 maths attainment will be centred around the overall (unweighted) school-level grand mean[22].

This approach ensures that zero variance in the outcome will be shared across the three variables (i.e. the correlation between them will be zero).

To reflect the research design, a 3 level multilevel regression model will be fitted to the data that will aim to account for variation in the GL PTM13 primary outcome. This model will include covariates at all three levels. Most covariates will be included at the school level (RME/Control identifier; three stratification variables & school-level KS2 maths attainment) but KS2 maths will also be included at both class and pupil levels.

This will be a random intercepts model. This means that the analyses will assume that the impact of RME will be consistent across RME schools and classes. As this is an efficacy

[22] The unweighted school level grand mean is the mean obtained using all school means. This means that each school mean will count once in calculating the unweighted school level grand mean. An overall pupil-level mean would be weighted at the school level by the number of pupils in each school.

trial, this assumption is appropriate (Spybrook, 2016). If this efficacy trial finds evidence of positive impact for RME on pupil maths attainment, a future larger scale effectiveness trial may be funded that could reliably examine variation in impact across schools and classrooms using multilevel models with both random intercepts and slopes.

To formally specify the ITT model, let Y_{ijk}represent the score in the GL PTM13 outcome in Spring 2021 for pupil i in classroom j in school k.

The level 1 (pupil-level) model is:

Equation II.1 $$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(KS2\_pc_{ijk}) + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2)$$

Where:

- $i = 1, \dots, n$ pupils per class; $j = 1, \dots, J$ classes per school; $k = 1, \dots, K$ schools
- π_{0jk} is the mean score for class j in school k
- $KS2\_pc_{ijk}$ is the pupil-level (class-centred) KS2 maths pre-test covariate for pupil i in class j in school k. π_{1jk} is the coefficient for the pupil-level KS2 maths covariate for class j in school k
- e_{ijk} is the pupil-level error/residual
- σ^2 is the within-class variance

The level 2 (class-level) model is:

Equation II.2 $$\pi_{0jk} = \beta_{00k} + \beta_{01k}(KS2\_clc_{0jk}) + r_{0jk} \qquad r_{ijk} \sim N(0, \tau_\pi)$$

Where:

- β_{00k} is the mean score for school k
- $KS2\_clc_{0jk}$ is the mean class-level KS2 maths covariate (school centred) for class j in school k. β_{01k} is the coefficient for the class level KS2 covariate for school k
- r_{0jk} is the random effect associated with each class
- τ_π is the residual/error variance between classes within schools

The level 3 (School-level) model is:

Equation II.3 $\beta_{00k} = \gamma_{000} + \gamma_{001}RME_k + \gamma_{002}KS2\_Sch_k + \gamma_{003}Strat_k + u_{00k};\ u_{00k} \sim N(0, \tau_{\beta_{00}})$

Where:

- γ_{000} is the estimated adjusted grand mean
- RME_k is '1'for treatment and '0' for control schools, γ_{001} is the effect of RME participation
- $KS2\_Sch_k$ is the school level mean KS2 maths covariate (centred around the school level mean), γ_{002} is the coefficient for the school-level KS2 maths covariate.
- $Strat_k$ represents a vector for the three stratification variables (geographical hub area, middle/secondary phase; use of setting/streaming), γ_{003} is a coefficient vector for the school-level stratification covariates. In total, for the three stratification variables, eight binary dummy variables will be included (five to account for the six geographical hub areas, one to identify middle/secondary school phases and one to identify whether a school had a policy of setting/streaming in secondary maths prior to randomisation)
- u_{00k} is the random effect associated with each school mean
- $\tau_{\beta_{00}}$ is the residual/error variance between schools

Example of STATA SYNTAX that will be used to fit the multilevel model

The multilevel regression model will be fitted to the data using the STATA mixed command, an example of the code that will be used is shown below:

Empty / Null Model:

- mixed GLPTM || School_ID: || Class_ID:

Outcome Only:

- mixed GLPTM Group|| School_ID: || Class_ID:

KS2 to GL PTM13 Progress:

- mixed GLPTM Group KS2Maths_SchC KS2Maths_ClassC KS2Maths_PupC || School_ID: || Class_ID:

Final (headline) analyses:

- mixed GLPTM Group KS2Maths_SchC KS2Maths_ClassC KS2Maths_PupC b1.Hub Phase SettStream || School_ID: || Class_ID:

The empty/null model will be used to obtain the standard deviation for calculating the Hedges g effect size statistic.

Table 7 provides a brief description of the five school level and single class and pupil level covariates that will be included in the final (headline) multilevel analyses.

Obtaining the Hedges g effect size statistic

The causal impact of RME on pupil maths attainment (GL PTM13) will be measured using the Hedges g effect size statistic. Hedges g standardises the difference between the attainment of pupils in RME schools and pupils in control schools into units of standard deviations. As specified in the EEF analyses guidance, the unconditional variance will be used to obtain the standard deviation. Specifically, the variance in the GL PTM13 outcome that is clustered at school, class and pupil levels will be used, as set out in equation II.4:

Equation II.4 $$ES = \frac{(T-C)_{adjusted}}{\sqrt{\delta^2_{sch}+\delta^2_{class}+\delta^2_{pup}}}$$

Where:

- δ^2_{sch} is the school level variance, δ^2_{class} is the class level variance and δ^2_{pup} is the pupil level variance for the GL PTM13 outcome from the empty/null multilevel model.
- $(T-C)_{adjusted}$ is the mean difference between the attainment of pupils in RME schools and pupils in control schools in the original raw (GL PTM) units. This is obtained from the coefficient for the school-level 'group' variable from the final (headline) analyses.

Table 4: Covariates used in the RME impact analyses at School-, Class- and Pupil-level

Description	*Name*	*Notes*
Primary Outcome	**GL PTM**	Primary Outcome (GL Progress Test in Maths)
ITT Impact Variable	**Group**	School-level binary variable identifying RME (1) and Control (0) schools
Pre-test Covariate at school, class & pupil levels. **(KS2 Maths)**	**KS2Maths_SchC**	School-level mean KS2 maths (centred around school level grand mean)
	KS2Maths_ClassC	Class-level mean KS2 maths (centred around school mean)
	KS2Maths_PupC	Pupil-level mean KS2 maths (centred around class mean)
School-level stratification variables. **Geographical Hub, School Phase, Use of setting/streaming**	**b1.Hub**	Generates five school-level binary (dummy) variables to describe the six geographical hub areas. The first category (London) will be the reference category for this variable. • Hub dummy 1 identifies schools in the South (1) or not (0); • Hub dummy 2 identifies schools in the East Midlands (1) or not (0); • Hub dummy 3 identifies schools in the West Midlands (1) or not (0); • Hub dummy 4 identifies schools in the North West (1) or not (0); • Hub dummy 5 identifies schools in the Yorkshire or the North East (1) or not (0);
	Phase	~~School-level binary variable identifying Secondary (1) and middle (0) schools~~ (see below)
	SetStream	School-level binary variable identifying schools that reported to have a policy of setting or streaming in Y7 or Y8 secondary maths prior to randomisation (1) or not (0).

Change following Covid19

Eight middle schools were unable to continue in the RME trial because pupils left at the end of Y8 to begin Y9 in secondary school from September 2020. This means that the Phase covariate in the above table is now redundant in the impact analyses and will not be included. This results in reducing the number of school-level covariates included in the analyses from nine to eight.

Appendix III School Level Baseline Comparison of intervention & control groups.

Overview

This Appendix provides a Table to compare the RME intervention and control groups at randomisation.

Table 8: School level summary comparing the RME intervention and control school samples at randomisation.

Description	*Intervention*	*Control*	*Total*
Total	60	59	119
Geographical Hub:			
NE & Yorkshire	6	6	12
NW	17	16	33
West Midlands	10	11	21
East Midlands	6	5	11
London	11	12	23
South	10	9	19
Whether use setting / streaming in Y7/Y8 maths			
No	7	6	13
Yes	53	53	106
Middle or Secondary School			
Middle	4	4	8
Secondary	56	55	111

Appendix IV Deriving the secondary outcomes from the PTM13 test..

The GL PTM test comprises of two parts; a mental maths test with 20 items and a test that assesses mathematical application and understanding with 23 items[23]. The primary outcome will be the raw test score for both parts of the GL PTM13 test and will represent a measure of general maths attainment.

For the secondary outcomes the 20 mental maths test items will be dropped. First, all 23 'mathematical application and understanding' items will be included. The second secondary outcome will be a subset of 11 of the 23 'mathematical application and understanding' items that were selected by MMU as being closely aligned to the RME programme.

Both of these secondary outcomes will be used in exploratory analyses examining the impact of RME on maths attainment. Please note that observed differences on these secondary outcomes will not provide the same standard of causal proof of impact for the RME programme when compared with the primary outcome. This is because the PTM outcome is not psychometrically designed for these two subscales. Table 9 places the 23 PTM13 test items under one of three process categories; fluency in conceptual understanding; mathematical reasoning and problem solving. The subset of 11 of the 23 items is indicated using a bold ✓**.**

Table 9: GL PTM13 mathematical application and understanding test items.

Description	*Fluency in Concepts*	*Mathematical Reasoning*	*Problem Solving*
Q1 (Quarters)	✓		
Q2 (ordering)	✓		
Q3 (Cubes)		✓	
Q4 (Brackets)	✓		
Q5 (Decimal Cards)		✓	
Q6 (Sale Prices)		✓	
Q7 (Day Care)	✓		
Q8 (Equivalencies)	✓		
Q9 (Advertising)		✓	
Q10 (Salaries)		✓	
Q11 (Pentagon)		✓	
Q12 (Spinning)		✓	
Q13 (Bronze)		✓	
Q14 (Equations)		✓	
Q15 (Graphs)			✓
Q16 (Photographs)		✓	
Q17 (Cycling Holiday)		✓	
Q18 (Boxes)			✓
Q19 (Driving Schools)		✓	
Q20 (Cuboid)		✓	
Q21 (Mistakes)		✓	
Q22 (Mowing a Lawn)			✓
Q23 (Bigger or Smaller)			✓

[23] See https://www.gl-assessment.co.uk/media/1384/ptm13-links-to-nc.pdf