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
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
1.2 [*latest*]		
1.1		
1.0 [*original*]	Oct 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

Table of Contents

SAP version history 1

Introduction 3

Design overview 3

Sample size calculations overview 3

Analyses 7

Research Questions for the Impact Evaluation of RME 7

- Primary question 7
- Secondary questions 7

Intention to Treat (ITT) and other pupil samples 7

ITT (headline) analysis of primary outcome (RQ1) 10

Follow-on exploratory analyses of the primary outcome (RQ2 & RQ3) 11

Follow-on exploratory analyses of secondary outcome (RQ4) 13

Fidelity analyses (RQ5) 13

Compliers Average Causal Effect (CACE) analyses (RQ6) 16

- Imbalance at baseline 17
- Missing data 18
- Intra-cluster correlations (ICCs) 18
- Additional Sensitivity Analyses 18

References 20

Appendix I Realistic Maths Power Analysis 21

Appendix II Realistic Maths Multilevel Analyses & Calculation of Hedges g effect size 22

Tables and Figures

Table 1: Sample size calculations 6

Table 2: Summary of analysis plan for ITT impact analyses for primary outcome (RQ1) 10

Table 3: Summary of analysis plan for exploratory follow-on impact analyses for primary outcome (RQ2 and RQ3) 11

Table 4: Summary of analysis plan for ITT impact analyses for secondary outcome (RQ4) 13

Table 5: Summary of analysis plan for fidelity analyses for primary outcome (RQ5) 14

Table 6: Pupil and Teacher-level fidelity dimensions and thresholds used for the CACE estimate 16

Table 7: School level and single class and pupil level covariates 25

Figure 1: Pupil Samples and Subsamples for the RME impact evaluation 8

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 & 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 PTM13, 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 PTM13 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 proportion of variance in

[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/

the GL PTM13 outcome that is clustered at the school and class levels are measured using Intra-cluster Correlation Coefficients (ICCs):

1. All 119 school practice mixed ability; School level ICC =0.15-0.20; **class level ICC=0.05**.
2. Some of the 119 school practice mixed ability others use setting. School level ICC =0.15-0.20; **class level ICC=0.25.**
3. All of the 119 school 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 to use 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 PTM13 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 amongst a Y8 pupil sample. This means that the proportion of variance in GL PTM13 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 PTiM (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

[3] For the Y8 cohort, Boylan et al. (2015) reported a class level ICC of 0.47 which is seen to increase to 0.70 for Y9.

[4] GL Progress in Mathematics (PiM) was replaced by GL Progress Test in Mathematics (PTM) in 2015. For the Y8 measures (PiM13 and PTM13) a correlation of 0.83 between the two is reported. See p12 here https://www.gl-assessment.co.uk/media/1346/ptm-technical-information.pdf

classes per school (3) whilst at the randomisation stage, the details were taken 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 PTM13 outcome is between 0.21 and 0.23. For the FSM subgroup analyses the MDES estimate is between 0.22 and 0.24.

Table 1: 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

Analyses

Research Questions for the Impact Evaluation of RME

The objectives for the impact analyses for the RME impact evaluation were to answer the following six research questions:

Primary question

RQ1. Does the RME intervention improve pupil attainment in mathematics over 2 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 throughout the trial period?

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 trial period?

RQ4. What is the effect on attainment of components of the GL PTM13 excluding mental mathematics, given this is not a focus of the intervention?

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?

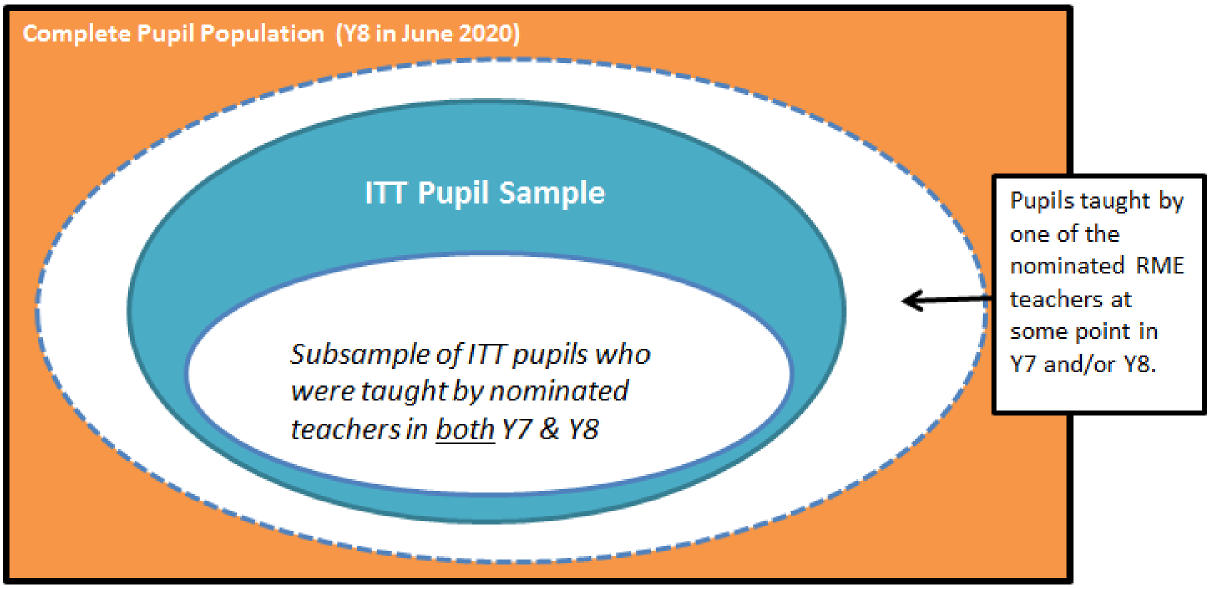
Intention to Treat (ITT) and other pupil samples

Figure 1 is a visual representation of the Y7/Y8 pupil population for the 119 schools in the RME impact evaluation. The orange box represents the whole Y7/Y8 pupil population. Within this are three 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 PTM13 will use this ITT sample (RQ1). This 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 on three occasions (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).

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


Complete Pupil Population (Y8 in June 2020)
ITT Pupil Sample
Subsample of ITT pupils who were taught by nominated teachers in both Y7 & Y8
Pupils taught by one of the nominated RME teachers at some point in Y7 and/or Y8.

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 evaluation period. This sample is a subset (or subsample) of the ITT sample and is illustrated in Figure 1 by the while ellipse within the ITT sample. Specifically, using pupil class lists collected at three points in the trial period (start, mid-trial and end); 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 on at least one of the three observation points (start, middle or end of trial). This sample will contain the ITT sample but is likely to be larger[5]. At the

[5] 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.

same time, this pupil sample will be smaller than the total Y8 pupil population. 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 PTM13 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 start/middle/end class list data collection) and the PTM13 outcome will be examined. A four point scale will be derived for the whole of the Y8 pupil population that measures the number of occasions a pupil was located in a classroom taught by an RME teacher: from 0 (representing the best evidence of a pupil receiving no exposure to RME - shown in Figure 1 shaded in orange outside the larger white ellipse) to 3 (representing the best evidence of a pupil receiving consistent exposure to RME - shown in Figure 1 as the small central white ellipse). If evidence is found that there is a positive correlation between pupil-level exposure to RME and the PTM13 outcome (including and excluding the mental maths items), follow-on exploratory analyses will look closer at pupil-level exposure by further defining the 4-point scale to take account of time. For example, whilst for the extremes of the 4-point scale a single pupil sample can be identified across the three time points, the values of 1 and 2 can be further defined (e.g. a value of 1 might indicate a pupil located in an RME classroom at the start middle or end of the trial and this distinction might matter).

At the teacher level, three dimensions of fidelity will be used (attendance of RME training days; RME material coverage & RME curriculum time).

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

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[6]. These RME teachers will have attended a minimum of four of the eight RME training days, covered a minimum of seven of the ten

[6] 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).

RME modules using a minimum of 12 weeks maths curriculum time (at least one fifth of the total maths curriculum in Y7 and Y8). 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 PTM13 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)[7]. The second model will also include KS2 maths attainment as a covariate at the pupil, class and school level[8]. The final model will also include the three variables[9] used within the stratified randomisation (geographical hub area, setting/streaming dummy and middle/secondary dummy[10]). It will be this final model that will be used for the headline ITT impact analysis for the PTM13 primary outcome.

Follow-on ITT analyses will focus on the impact of the RME programme on maths attainment for pupils ever classed as FSM[11]. 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 GM PTM13 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 June 2020.

[7] 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.

[8] 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.

[9] 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)

[10] 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.

[11] The NPD EverFSM_6 variable will be used

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

Follow-on exploratory analyses of the primary outcome (RQ2 & RQ3)

Table 3 below summarises further planned follow-on impact analyses for the *RME* evaluation. These analyses draw on GL PTM13 test data for the whole of Y8 and data that monitor pupil/teacher movement between classes during the trial period to identify two subsamples in order to address RQs 2 and 3. A similar multilevel approach will be taken for analyses used to examine RQ2 (consistent exposure) and RQ3 (partial exposure) as used for the headline ITT analyses.

Table 3: Summary of analysis plan for exploratory follow-on impact analyses for primary outcome (RQ2 and RQ3)

Primary Outcome: GL Progress Test in Maths (PTM13) taken in June 2020.

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (class) Variables*	*Level 3 (school) Variables*
RQ2 **ITT Subsample** **Pupils taught by RME teacher in BOTH Y7 and Y8**	-	-	• Group (1=RME; 0=control)
	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
RQ3 **Y8 Subsample**	-	-	• Group (1=RME; 0=control)
	KS2 Maths (class centred)	Mean KS2 Maths (school centred)	• Group (1=RME; 0=control); • Mean KS2 Maths (Grand mean centred)

Pupils taught by RME teacher at some point in Y7 and Y8			Stratification variables: • Geographical hub area • Setting/streaming dummy • Secondary/Middle dummy

Follow-on exploratory analyses of secondary outcome (RQ4)

The GL PTM13 primary outcome includes a 'fixed time' mental maths section that runs counter to the RME ethos which encourages / promotes pupils to take time in order to develop their depth of mathematical understanding. Therefore, the secondary outcome for the RME evaluation will be the GL PTM13 outcome with all of the mental maths items removed. The analytical approach adopted will be similar to the ITT analyses of the primary outcome and is summarised in Table 4.

Table 4: Summary of analysis plan for ITT impact analyses for secondary outcome (RQ4)

Secondary outcome: GL Progress Test in Maths (PTM13) excluding the mental maths items

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)
Main 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

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 4-point scale will identify whether a pupil was located in a maths classroom taught by an RME teachers on up to three occasions (start, middle & end) during the trial[12].

[12] This may be a the same maths teacher throughout the trial period (Y7 and Y8) or when an RME maths teacher is replaced by another teacher during the trial period (Y7 or Y8).

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)

It will be the pupil and three teacher-level dimensions of fidelity to RME that 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. 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 5 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 (GL PTM13).

Table 5: Summary of analysis plan for fidelity analyses for primary outcome (RQ5)

Primary outcome: GL Progress Test in Maths (PTM13) taken in June 2020.

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 scale (0-3)	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
RQ5 ***Fidelity Analyses [pupil, teacher &***	KS2 Maths (class centred)	Mean KS2 Maths (school centred) Teacher-level dimensions of fidelity: • Attendance	• Mean KS2 Maths (Grand mean centred) Stratification variables: • Geographical hub area • Setting/streaming dummy

school levels] RME Intervention school sample restricted to pupils taught by a nominated RME teacher throughout the trial period		• Module use • Curriculum time	• Secondary/Middle dummy School-level dimension of fidelity: • Number of times a school sent 2+ teachers to RME training days

Keeping the pupil, teacher and school level components of fidelity separate / distinct 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 PTM13 outcome. We will also examine the strength of association between the five measures of fidelity (they may be highly correlated). Whilst for these exploratory analyses into the relationship between fidelity to RME and the PTM13 outcome, a single pupil-level measure of 'compliance' to RME will be derived within the Complier Average Causal Effect (CACE) analyses (see below).

Compliers Average Causal Effect (CACE) analyses (RQ6)

The purpose of the Complier Average Causal Effect (CACE) analyses is to estimate the impact of RME for pupils deemed to have 'complied' with the RME intervention. In an educational context, the economic/health language of CACE is rather unclear but in this instance 'compliance' relates to both pupils and their maths teachers and to dosage/exposure and fidelity to the RME intervention. Specifically, for RME, to be considered as 'compliant', a pupil will need to have been located *consistently* in a maths class that is taught by an RME teacher who attended at least at least four of the eight RME training days, used RME in maths lessons for at least 12 weeks (20% or more of total available maths teaching time in Y7 and Y8) and covered at least seven 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 PTM13 along with additional school level items that are 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 PTM13) 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 6: 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)	Pupil / Teacher class lists collected prior to randomisation (Sept 2018); mid-trial (Sept/Oct 2019) and end of trial (June 2020).	3 Instances	3 instances	Pupil located in classroom taught by an RME teacher on all three class lists.
Teacher-Level 2. (Professional development attendance)	MMU attendance list	8 days	4 days	Half day or more counted as 1
Teacher-Level 2. (RME curriculum time)	Percentage of school lessons used based on number of maths lesson per week collected from PD day survey - based on 60 possible weeks RME could be taught Collected via annual teacher surveys.	20 weeks out of 60 - based on 2 weeks per 10 modules	20% of total maths available teaching time in Y7 & Y8. (60 % of RME 20 weeks … 12 weeks or more RME)	Teacher self-report that a lesson as a RME module lesson counts as 1 lesson
Teacher-Level 3. (RME module coverage)	Number of modules and what materials used. Collected via annual teacher surveys.	10 modules	7 modules	The amount of material/amount of teaching in each module to count as coverage to specified in the SAP

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 PTM13 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[13]. 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.
- For instances where only data is missing in the PTM13 outcome, we will add any additional covariates that were found from stage 1 to the final ITT model and re-estimate the effect size.
- For instances where data is missing across covariates other than the PTM13 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

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

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 PTM13 outcome is measured across four maths dimensions; fluency in facts and procedures; fluency in conceptual understanding; mathematical reasoning and problem solving. These four dimensions of PTM13 have not been psychometrically designed as validated 'stand alone' 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 PTM13 dimensions that do not draw on the mental maths aspects of PTM13 (mathematical reasoning and problem solving). As exploratory analyses, the analyses of the mathematical reasoning and problem solving dimensions of PTM13 will be included as an appendix in the final evaluation report for the RME programme.

Hedges g effect size

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.

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) Muliplicative 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. Psychol 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[14].

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 trial, this assumption is appropriate (Spybrook, 2016). If this efficacy trial finds evidence of positive

[14] 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.

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 June 2020 for pupil i in classroom j in school k.

The level 1 (pupil-level) model is:
Equation II.1 $Y_{ijk} = \pi_{0jk} + \pi_{1jk}(KS2\_pc_{ijk}) + e_{ijk}$ $e_{ijk} \sim N(0, \sigma^2)$

Where:

- $i = 1, \ldots, n$ pupils per class; $j = 1, \ldots, J$ classes per school; $k = 1, \ldots, K$ 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 $\pi_{0jk} = \beta_{00k} + \beta_{01k}(KS2\_clc_{0jk}) + r_{0jk}$ $r_{ijk} \sim N(0, \tau_\pi)$

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 7: 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
	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).

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