The 5Rs approach to GCSE Maths resits, a two-arm cluster randomised trial

Statistical Analysis Plan

Evaluator: York Trials Unit, University of York
Principal investigators: Pam Hanley & Louise Elliott

September 2020

PROJECT TITLE	The 5Rs approach to GCSE Maths resits, a two-arm cluster randomised trial
DEVELOPER (INSTITUTION)	Julia Smith Ltd & Association of Colleges
EVALUATOR (INSTITUTION)	York Trials Unit, University of York
PRINCIPAL INVESTIGATORS	Pam Hanley & Louise Elliott
PROTOCOL AUTHORS	Pam Hanley, Louise Elliott, Caroline Fairhurst & Elizabeth Coleman
TRIAL STATISTICIANS & SAP AUTHORS	Elizabeth Coleman & Caroline Fairhurst
TRIAL DESIGN	Two-armed cluster randomised controlled trial with random allocation at the setting level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	KS5; age 16 -19
NUMBER OF SCHOOLS	80 (88 randomised, 78 included)
NUMBER OF PUPILS	Maximum 6400 (actual 4486)
PRIMARY OUTCOME MEASURE AND SOURCE	GCSE maths attainment (GSCE raw maths score converted to a z-score as scale varies by exam board – raw scores obtained from settings)
SECONDARY OUTCOME MEASURE AND SOURCE	- Achievement of a grade 4 or above at GSCE resit (i.e. pass; GCSE maths score, scaled 9-1, binary, yes/no, obtained from settings); - Attendance at exam sessions (all 3 or 0-2, obtained from settings); - Student attitude towards maths (Adapted Attitudes Towards Mathematics Inventory (ATMI) (Tapia & Marsh, 2000) obtained via the pupils in settings.

This SAP was initially prepared post-randomisation of the settings, but prior to any collection of data at the pupil level or analysis, and prior to the cancellation of GSCE exams as a result of the COVID-19 pandemic. Due to this, the planned analysis for this trial has changed substantially as the primary outcome (teacher-assessed GCSE maths attainment) was deemed unusable. As such, this SAP presents the analysis originally planned (which will not be undertaken), and the (descriptive analysis) which will be undertaken to allow for some evaluation of the data (see page 12).

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	09/03/2020	-
1.1	07/09/2020	Clarification of the actual analysis of data for the 5Rs trial that will be undertaken. The changes result from the implications to the trial of the COVID-19 pandemic.

Protocol and SAP Changes

Any changes to the protocol which will impact the SAP, and any changes to the SAP after initial publication will be detailed here.

Changes from the protocol:

In the protocol it was stated that type and size of setting would be used as minimisation factors. Size of setting was measured by the number of students who resat GCSE maths in 2018/2019 and type of setting was categorised as Further Education College, Sixth Form College, and School Sixth Form. However, it came to light prior to randomisation that some of the settings were involved in Basic Maths Premium (BMP, additional funding to improve the quality of teaching of maths in post-16 education), and it was thought this may have an impact on the results – after approaching all settings, ten (11.4%) stated that they were involved. Thus, this was also included as a minimisation factor to ensure balance across the two arms.

Table of contents

Introduction

This trial will investigate the efficacy of the 5Rs approach. The aim of 5Rs is to improve maths skills, and in this trial the impact of the intervention will be evaluated on those in post-16 education (KS5) who have failed their maths GCSE and are resitting to try to achieve a pass grade (Grade 4 or above).

The primary objective in this trial is to evaluate the effectiveness of 5Rs compared to teaching as normal in improving maths GCSE outcomes for resitting pupils aged 16-19. Secondary objectives are included to evaluate if 5Rs:

- is effective at improving the pass rate for pupils aged 16-19 resitting GCSE maths;
- has an impact on student attitudes towards maths;
- has an impact on student exam retention rates;
- has a greater benefit for students resitting in May/June 2020 rather than November 2019;
- is effective in those who have ever been eligible for free schools meals (FSM), and whether the effect is differential to the effect of those who have never been eligible for FSM.

Design overview

This is a two-armed cluster randomised controlled efficacy trial, where the clusters are higher education settings. Minimisation was used to allocate settings to either of the two arms; intervention or control – and type of setting, number of students resitting maths in the previous academic year, and participation in BMP were used as minimisation factors. The primary outcome is GCSE maths attainment, with pass (yes/no), exam attendance, and students' attitude towards maths used as secondary outcomes.

Trial design, including number of arms		Two-armed cluster randomised controlled efficacy trial
Unit of randomisation		Further Education Settings
Stratification variables (if applicable)		- Type of setting (3 levels; Further Education College, School Sixth Form, Sixth Form College); - Number of students who resat maths in the academic year 2018/2019 (2 levels; <173, >=173); - Participating in Basic Maths Premium (BMP) (2 levels; yes/no)
Primary outcome	variable	GCSE maths attainment
	measure (instrument, scale, source)	GCSE maths raw score converted to z-score for analysis (as the scale varies by exam board, provided by the settings)
Secondary outcomes	variables	GCSE maths pass; Exam attendance; Students attitudes towards maths.
	measures (instrument, scale, source)	- Achievement of a grade 4 or above at GSCE resit (GCSE maths score, scaled 9-1, binary, yes/no, obtained from settings); - Attendance at exam sessions (two-levels; all 3, 0-2, obtained from settings); - Student attitude towards maths (Adapted Attitudes Towards Mathematics (ATMI)

		(Tapia & Marsh, 2000) obtained via the pupils in settings).
Baseline for primary outcome	**variable**	Previous maths attainment score
	measure (instrument, scale, source)	KS2 maths score (NPD; KS2_MATSCORE)
Baseline for secondary outcome	**variable**	Previous maths attainment score
	measure (instrument, scale, source)	KS2 maths score (NPD; KS2_MATSCORE)

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	**FSM**	**OVERALL**	**FSM**
Minimum Detectable Effect Size (MDES)		0.22	0.25	0.23	0.26
Pre-test/ post-test correlations	level 1 (pupil)	0.6	0.6	0.6	0.6
Intracluster correlations (ICCs)	level 3 (setting)	0.17	0.17	0.17	0.17
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two
Average cluster size		60	13	51	9
Number of settings	intervention	40	40	44	44
	control	40	40	44	44
	total	80	80	88	88
Number of pupils	intervention	2400	504	2300	483
	control	2400	504	2186	459
	total	4800	1008	4486	942

Notes:

1. Calculations for the number of pupils in the FSM group in the table above assume that 21% of all higher education pupils are, or have ever been, in receipt of FSM.
2. We randomised 88 settings (44 to each arm); however, not all randomised settings provided pupil details (those that did not provide pupil details before 10th January 2020 were withdrawn). As such, the 'as randomised' number of pupils and MDES are based on the actual numbers for the 78 (41 intervention, and 37 control) settings which remain in the trial, and approximations for the remaining ten settings. The approximated number of resitting pupils for the ten settings was taken to be the

number which they stated had resat in 2018-2019 (provided for randomisation); if this number was greater than 80, it was set to be 80 as we accepted a maximum of 80 per setting. The mean cluster size, and range of cluster size used in the calculations of MDES include these estimates.

FROM PROTOCOL

Overall

For this efficacy trial, the programme developers specified that they would have capacity to deliver the intervention within a maximum of 40 settings. Therefore, the aim was to recruit 80 settings into the trial using 1:1 allocation. A maximum of 80, and a minimum of 15, learners per setting were to be enrolled into the trial; we assumed there will be an average of 60 students per setting at randomisation. A large variation in cluster sizes has the potential to increase the minimum detectable effect size (MDES) that the trial is able to detect. It is possible to account for variation in cluster size in the calculation of the MDES by considering the coefficient of variation of cluster size in the design effect (DE) as per Eldridge *et al* (2006). The DE, also called the variance inflation ratio, is the factor by which the sample size (at analysis) for a comparable individually randomised trial should be multiplied to estimate the required sample size (at analysis) for a cluster RCT. A simple yet conservative estimate of the DE accounting for variable cluster sizes is:

$$DE = 1 + \{(cv^2 + 1)\tilde{m} - 1\}\rho$$

Where the coefficient of variation (cv) is the ratio of the standard deviation of the cluster sizes, s_m, to the mean cluster size, $\tilde{m}$. The average cluster size at randomisation is assumed to be 60; however, if we account for 15% pupil-level attrition at post-test (i.e. pupils withdrawing from sitting the exam) then we expect an average of 51 pupils per setting at analysis ($\tilde{m} = 51$). Since we do not know s_m in advance, we estimated it by dividing the likely range of cluster sizes (at analysis) by 4 (so (68-13)/4=13.75) (Eldridge *et al,* 2006). Therefore, we assumed a cv of 13.75/51=0.27. The symbol ρ represents the intracluster correlation coefficient (ICC). Allen *et al* indicate that the ICC for maths increases with each Key Stage. At Key Stage 4 they calculated an ICC of 0.15 for Maths based on over 500,000 pupils from 3058 schools. Since this trial involves a KS5 population, we conservatively assumed a slightly increased ICC of 0.17. Within the analysis for this trial we shall adjust for pupil-level KS2 maths score (KS2 has been chosen as a more discriminating baseline measure because most of the previous GCSE results will be Grade 3). This will correlate with the outcome to increase the power of the trial (i.e. decrease the MDES the trial is able to detect). There are limited data on which to estimate the likely correlation between maths KS2 and GSCE resit score for pupils who fail their initial GSCE attempt. There are data to suggest that, nationally, the correlation between KS2 maths and GSCE maths is high (0.76[1]); we conservatively assumed a lower correlation of 0.6 in this scenario.

Therefore, we anticipated to recruit 4800 students (80 settings with average of 60 students per setting), which will reduce to 4080 at analysis after 15% attrition. Inflating this by ($1-0.6^2$) (Borm *et al.* 2007) to take advantage of the pre-post correlation, and then deflating for the design effect of ~10.15 (accounting for variable cluster sizes), we obtained an effective sample size (for an individually randomised trial) of 628.

[1] http://www.cambridgeassessment.org.uk/Images/181034-exploring-the-value-of-gcse-prediction-matrices-based-upon-attainment-at-key-stage-2.pdf

With 80% power, this sample size gave us a MDES of approximately 0.22 in the analysis (calculated in Stata v15).

FSM

Data from the Sixth Form Colleges Association (2018 key facts and figures[2]) indicate that 21% of pupils attending sixth form colleges and 16-19 academies are 'disadvantaged', defined as "those who were eligible for free school meals at any point in the previous six years or have been looked after by their local authority". We could find no published data relating to FSM prevalence in our specific trial population (i.e. learners resitting GCSE maths). However, we may expect that it is slightly higher than the average of 21%. In the absence of evidence to support this, however, we conservatively assumed a percentage of 21%. If we recruit 80 setting with an average of 60 pupils, we might therefore expect to have at least 856 FSM pupils in the analysis, assuming 15% attrition. Assuming a learner-level pre-post test correlation of 0.6 and a setting-level ICC of 0.17, this sample size would have 80% power to detect an effect size of 0.25. This calculation also accounts for variable cluster sizes as described above; however, the variation would likely be small and so, rounding to 2 decimal places, the MDES is the same when assuming equal cluster sizes at analysis.

AT RANDOMISATION

In total 88 settings were randomised into the trial, with 44 in each of the two arms (intervention and control). However, some settings did not provide pupil level data until after randomisation, as the pupils had not yet started the academic year, and subsequently it was only provided by 78 settings (41 intervention and 37 control). As predicted, there was a large variation in expected cluster size, and as such any setting which had more than 80 pupils in their cohort resitting GCSE maths were asked to provide a list of class sizes (35 settings, 39.8% of those randomised). From this, the average class size for each setting was used to determine how many classes should be selected from that setting to have as close to 80 students as possible. This number of classes was then randomly selected from the corresponding settings. This was performed in Stata v15. As such, there were 3816 pupils, 2148 in the intervention arm, and 1668 in the control arm – cluster size now ranging from 3 to 108. This still exceeds the maximum of 80 students per setting in some instances, however the classes were selected on expected number of pupils. It is also worth noting that the lowest cluster is smaller than was pre-specified, however, pupils within the classes had the choice to not participate after the classes were selected.

As detailed earlier, for the ten settings which withdrew prior to providing pupil level data, we have estimated the numbers from each setting. These numbers are been estimated to be the same as the provided data for randomisation, number of pupils resitting in 2018-2019. For settings where there were more than 80 pupils, an estimate of 80 was used. This provides an additional 518 control pupils, and an additional 152 intervention pupils; giving a total sample size of 4486 (2300 intervention, 2186 control).

From this, we found there was a mean cluster size of 51.0 ranging from 3 to 108. Accounting for 15% attrition at analysis, the mean cluster size is 43.3 (SD 22.3). This allows us to calculate the cv of 0.52, and thus gives a design effect of 10.14; assuming the ICC of 0.17 as before. With a total sample size of 4486, assuming an ICC of 0.17, a pre-post (pupil-level) correlation

[2] https://sfcawebsite.s3.amazonaws.com/uploads/document/Sixth-form-colleges-Key-facts-and-figures-2018.pdf?t=1545390007

of 0.6, and 15% attrition at the pupil level, we would have 80% power to detect an effect size of 0.23, accounting for the unequal number of pupils in each arm.

As the FSM status is not yet known (there is no exact figure for the FSM group at randomisation), we shall assume it to be 21% of those randomised, so there would be a sample of 942. This would allow us to detect a MDES of 0.26, with 80% power under the pre-specified conditions.

When looking just at the 78 who provided pupils details, we have a mean cluster size of 48.9 (range still 3-108), which is 41.6 (SD 22.3) when adjusting for 15% attrition, a cv of 0.54, and thus giving a design effect of 10.05. With a total sample size of 3816, assuming an ICC of 0.17, a pre-post (pupil-level) correlation of 0.6, and 15% attrition at the pupil level, we would have 80% power to detect an effect size of 0.25, accounting for the unequal number of pupils in each arm.

As the FSM status is not yet known (there is no exact figure for the FSM group at randomisation), we shall assume it to be 21% of those randomised, so there would be a sample of 801. This would allow us to detect a MDES of 0.29, with 80% power under the pre-specified conditions.

Planned Analysis

The statistical analysis proposed follows the most recent EEF guidance (https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf, accessed 25 October 2019). Analysis will be conducted in Stata v15 (or later), using the principles of intention to treat, where data are available, including all settings and pupils in the groups they were allocated to, irrespective of whether or not they received the intervention. The statistical significance will be assessed using two-sided tests at a 5% significance level. Effect estimates will be presented alongside 95% confidence intervals (CIs), and p-values.

Primary analysis

The primary outcome of this study is the GCSE maths score at the resit. Since settings may sit the exam for whichever exam board they choose (Edexcel, OCR, AQA, etc) and the scale for these exam boards differ, the scores will not be directly comparable. As such, the raw marks will be converted to a 'standard' (z) score for analysis. That is, for pupil i, sitting exam board b, their z-score (z_{ib}) is:

$$z_{ib} = \frac{x_{ib} - \mu_b}{\sigma_b}$$

Where x_{ib} is the pupil's raw score, μ_b is the mean score of those who sat the exam with board b, and σ_b is the standard deviation of the population sitting the exam with board b. Where possible μ_b and σ_b will be obtained from the exam boards, else a sample mean and standard deviation will be used from the observed data. Both the raw mark and the exam board used will be provided by the settings. The standard scores will be summarised descriptively by trial arm. The raw marks will be presented within each exam board, by trial arm.

The difference in maths attainment between pupils in the intervention arm and the control arm will be compared using a mixed-effect linear regression at the pupil-level with the standardised maths GCSE mark of the most recent resit used as the response variable.

There are two exam sittings in an academic year, in November and May. Pupils who sit and pass the exam in the first sitting (November 2019) would not be expected to continue with the program, and as such this will be their most recent resit mark. For pupils who sit and fail the exam in November and go on to resit again in May 2020, or who do not sit the November exam, the May 2020 result will be the most recent. Therefore, there will be a variation in dosage of 5Rs, which will be explored in the subgroup and sensitivity analysis, see later. However, the most recent resit mark will be the one that is the response variable in the primary analysis model. The only time this will not be applicable is if a student sits and passes the exam in November but chooses to resit again in May for an even higher mark. In this case, we will only take account of their November sitting, as this trial is primarily interested in students achieving a pass grade who have previously only failed; we will not consider their result from the May sitting.

Group allocation, KS2 maths score and the minimisation factors (type of setting, number of resits in previous year at the setting, and participation in BMP) will be included as fixed effects in the model. Here, number of pupils resitting at each setting in the previous academic year will be included as a continuous variable, rather than in the dichotomised way used for minimisation (<173, >=173), to reduce the risk of confounding remaining (Altman, 2006). Setting will be included as a random effect to account for the clustering (Wears, 2002). The results will be presented as the adjusted mean difference in scores between the two groups with an associated 95% CI and p-value. In the model below β_2, the fixed effect parameter for the group allocation, represented by I_{A_k}, is the parameter of interest.

Pupil-level fixed effects:

- KS2 maths score (continuous)

Setting-level fixed effects:

- Allocation (2 levels; intervention or control)
- Type of setting (3 levels; Future Education College, School Sixth Form, Sixth Form College)
- Number of pupils who resat maths in 2018/2019 (continuous)
- Participating in Basic Maths Premium (yes/no; binary)

Adjustment will be made for clustering at the setting level by including school as a random effect.

Model equation:

$$Y_{ik} = \beta_0 + \beta_1 x_{ik} + \beta_2 I_{A_k} + \beta_3 I_{BMP_k} + \beta_4 I_{T1_k} + \beta_5 I_{T2_k} + \beta_6 I_{T3_k} + \beta_7 y_j + u_k + \epsilon_{ik}$$

Y_{ik}= response (GSCE resit standardised score) of the ith member of the kth cluster (setting), i=1,…, n_k, k=1,…, m.

n_k = size of kth cluster (setting)

m = number of clusters (settings)

x_{ik}= baseline score (KS2 maths score) for ith member of the kth cluster (setting)

I_{A_k}= 0/1 indicator variable for group allocation of the kth cluster (setting)

I_{BMP_k}= 0/1 indicator variable for participation of the k[th] cluster (setting) in BMP

I_{T1_k}= 0/1 indicator variable for Future Education College of the k[th] cluster (setting)

I_{T2_k}= 0/1 indicator variable for School Sixth Form of the k[th] cluster (setting)

I_{T3_k}= 0/1 indicator variable for Sixth Form College of the k[th] cluster (setting)

y_k= number of pupils resitting GCSE maths in academic year 2018/2019 in the k[th] cluster (setting)

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$= fixed effect parameters

u_k = random effect for k[th] cluster (setting)

ϵ_{ik}= residual error term for i[th] student of the k[th] cluster (setting)

Model assumptions will be checked as follows: the normality of the standardised residuals will be checked using a qq plot. No formal statistical tests will be undertaken, only a visual inspection of the plots. If the model assumptions are in doubt, a sensitivity analysis will be conducted in which transformations of the outcome and/or covariate data will be tried to improve the model fit.

Secondary outcome analysis

Achievement of grade 4 or above at resit (i.e. achieving a pass)

Mixed-effects logistic regression will be used to compare the likelihood of students achieving a pass on their most recent sitting between the two groups, as a binary variable. The model will be adjusted in the same way as the primary outcome. The baseline measure of prior attainment will be KS2 maths score, as it was in the primary analysis.

Student attendance at exam session

Within each exam sitting there are three papers that are to be sat by the students; each student will attend between 0 and 3 exam sessions. The number of sessions that are attended by students in their most recent sitting will be reported in a dichotomised fashion; all 3, or 0-2, and reported descriptively by trial arm. The results will be compared using a mixed-effects logistic regression model, adjusted as in the primary analysis.

Student attitudes towards maths – Adapted ATMI

The Adapted ATMI will be used to assess attitudes towards maths. The language was adjusted where needed (North American origin) and, with agreement of the originator (Martha Tapia), it was shortened. Questions that seemed less relevant/applicable to 5Rs were removed and, to maximise the validity of the shortened instruments, this resulted in the removal of the whole Motivation and Value subscales – for justification see 5Rs protocol (https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/5Rs_post-16_GCSE_Resit_maths_EP.pdf). This meant that the Self-confidence and Enjoyment subscales remained, as they were deemed most suitable; this results in 25 questions.

The questions in the ATMI have a 5-point Likert response scale: 1=strongly disagree to 5=strongly agree, with 3 being neutral. The score is calculated by summing the responses, with any negative questions being reverse scored. In the version used in this evaluation,

there are 15 positively worded questions, and 10 negatively worded questions. The resulting total score is between 25 and 125, where a higher score indicates a more positive attitude.

The total ATMI score will be analysed in the exact same way as the primary outcome. The subscale scores for Self-confidence and Enjoyment will be summarised descriptively by arm, but no formal comparison will be undertaken.

Subgroup analyses

Two subgroup analyses are planned, as described below. These analyses are underpowered, and as such are exploratory only.

FSM

To explore the effect of the intervention on pupils who are identified as eligible for FSM, according to the EVERFSM_6_P variable from the National Pupil Database, both the FSM status and an interaction term between FSM status and allocation will be included additionally in the primary analysis model. The primary analysis will also be repeated in the FSM subgroup.

TIMING OF RESITS

The effect of the intervention on those resitting in May versus those resitting in November will be assessed. It is theorised that those who resist in May 2020 may perform better, due to receiving the 5Rs intervention for a full academic year, as opposed to those who resit in November 2019 who received the intervention for a few months. However, those who enter the November resit may have done so due to either higher ability, or settings' policies. To assess the effect the primary analysis model will be run with the inclusion of an interaction term between time of resit and group allocation. The effect of the timing of the resit is further explored within the sensitivity analysis.

Additional analyses

SENSITIVITY ANALYSES

TIME OF RESIT

The effect of the intervention on those resitting in May versus those resitting in November will be further assessed, as it is theorised that there will be a greater benefit to those who resit in May, as they will receive more time with the 5Rs program. The primary analysis will be repeated including timing of resit (November or May) and whether this was the pupils first or second resit attempt as fixed effects in the model.

The primary analysis will be repeated in just those whose result was from the November 2019 resit, and similarly for those where the result was from the May 2020 resit, with the inclusion of first or second attempt as a fixed effect.

Additionally, the primary analysis will be repeated for only those pupils who sat the foundation tier paper (i.e. excluding those who sat the higher tier paper). It is not expected that a high proportion of students will be sitting the higher paper – 5 to 10%[3].

TEACHER IMPACT

As information on the class/teacher will be collected to explore the compliance aspect of this intervention, the impact of class/teacher level clustering will also be explored here. The primary analysis will be repeated with the inclusion of an additional variable term for the random effect associated with the class.

Longitudinal follow-up analyses

No longitudinal analyses are currently planned.

Imbalance at baseline

Baseline student and setting characteristics and measures of prior attainment will be summarised descriptively both as randomised and as included in the primary analysis (if differing). These data will include type of setting, number of pupils resitting in previous academic year, participation in BMP, and % FSM at the setting level, as well as previous KS2 result at the pupil level. Continuous measures will be reported as a mean, standard deviation (SD), median, minimum and maximum, while categorical data will be reported as a count and percentage. No formal statistical comparisons will be undertaken (Senn, 1994), except for a comparison of the difference in prior attainment (KS2 maths attainment) between the groups, reported as the Hedge's g effect size, with a 95% CI.

Missing data

The amount of missing data will be summarised for covariates in the primary analysis model and the outcome variable (resit GSCE score). Where possible, reasons for missing data will be explored and provided. The baseline characteristics of those included in the primary analysis model will be compared with those not included due to missingness of the outcome or covariate data.

If more than 5% of the pupils cannot be included in the primary analysis model potential predictors of missingness will be explored using a mixed-effects logistic regression model with the presence of GSCE raw resit mark as a binary outcome, including all baseline variables as fixed effects, and setting as a random effect. Any factors deemed to be significant predictors of missingness will be discussed in the final report.

Should the assumption of missing at random hold, that is, there are no variables that predict the missingness of the primary outcome, the impact of missing data on the primary analysis will be explored using multiple imputation by chained equations. This will be done in Stata v15 (or later). A 'burn-in' of 20 will be used, to allow the iterations to converge to a stationary distribution, and 30 imputed datasets will be created. Each of these imputed datasets will then be used to rerun the primary analysis, and Rubin's rule will be used to combine the imputed estimates. Should these results differ from that found in the primary analysis, it could be concluded that the missingness is not at random.

[3] Figures based on the November 2019 GCSE resit sitting, provided by the developer team.

Where the missing at random assumption does not hold, and missingness of the primary outcome can be attributed to one or more of the variables include in the mixed-effects logistic model, the primary analysis model will be re-run including these variables.

Compliance

A Complier Average Causal Effect (CACE) analysis for the primary outcome will be considered to account for engagement with the intervention. Compliance will be defined at the teacher level, rather than the setting level – teachers are associated with classes within the settings. Compliance is defined as attending at least the first two of the three training sessions (or equivalent if the setting joined the program late). This information will be provided by the developers, who will be running the training sessions and keeping a register of attendance. A Two Stage Least Squares (2SLS) instrumental variable (IV) approach with group allocation as the IV will be used. The correlation between instrument (I_{A_k}) and the endogenous variable will be reported, along with the F-statistic, for the first stage.

The first stage model is as follows:

$$C_j = \beta_0 + \beta_1 x_{ijk} + \beta_2 I_{A_k} + \beta_3 I_{BMP_k} + \beta_4 I_{T1_k} + \beta_5 I_{T2_k} + \beta_6 I_{T3_k} + \beta_7 y_k + \mu_{1k} + \epsilon_{1ijk}$$

Where:

- C_j = 0/1 variable indicating compliance or not for teacher j
- x_{ijk}= baseline score (KS2 maths score) for ith member of the kth cluster (setting), with teacher j
- I_{A_k}= 0/1 indicator variable for group allocation of the kth cluster (setting)
- I_{BMP_k}= 0/1 indicator variable for participation of the kth cluster (setting) in BMP
- I_{T1_k}= 0/1 indicator variable for Future Education College of the kth cluster (setting)
- I_{T2_k}= 0/1 indicator variable for School Sixth Form of the kth cluster (setting)
- I_{T3_k}= 0/1 indicator variable for Sixth Form College of the kth cluster (setting)
- y_k= number of pupils resitting GCSE maths in academic year 2018/2019 in the kth cluster (setting)
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$= fixed effect parameters
- μ_{1k} = random effect for kth cluster (setting)
- ϵ_{1ijk} = error term for student i, with teacher j, at setting k

The estimated values of C_j from this model,$\widehat{C_j}$, will then be used in the second stage model:

$$Y_{ijk} = \alpha_0 + \alpha_1 x_{ijk} + \alpha_2 \widehat{C_j} + \alpha_3 I_{BMP_k} + \alpha_4 I_{T1_k} + \alpha_5 I_{T2_k} + \alpha_6 I_{T3_k} + \alpha_7 y_k + \mu_{2k} + \epsilon_{2ijk}$$

Where:

- Y_{ijk}= response (GSCE resit standardised score) of the ith member of the kth cluster (setting), i=1,…, n_i, k=1,…, m, associated with teacher j
- n_k = size of kth cluster (setting)
- m = number of clusters (settings)
- x_{ijk}, I_{BMP_k}, I_{T1_k}, I_{T2_k}, I_{T3_k}, y_k, defined as above
- $\alpha_0, \dots, \alpha_7$= fixed effect parameters
- u_{2k} = random effect for kth cluster (setting)

- ϵ_{2ijk}= error term for student i, with teacher j, at setting k

The parameter of interest here will be α_2.

Intra-cluster correlations (ICCs)

The intra-cluster correlation coefficient (ICC) will be summarised for standardised GCSE maths score at setting level, extracted from the primary analysis model; this will be reported with the 95% CI. Similarly, the associated ICC for standardised GCSE maths score and teacher will be reported with a 95% CI.

The empty variance components model:

$$\boldsymbol{Y_{ijk} = \beta_0 + \eta_k + \epsilon_{ik}}$$

Will be used, where $\boldsymbol{Y_{ik}}$ is the standardised score GSCE resit of the ith pupil of the kth setting, with teacher j, $\boldsymbol{\eta_k}$ the setting-level random effect, and $\boldsymbol{\epsilon_{ik}}$ is the error term for pupil i and setting k. The ICC will then be calculated using the following formula:

$$\boldsymbol{ICC = \frac{var(\eta_k)}{var(\eta_k) + var(\epsilon_{ik})}}$$

Effect size calculation

Effect sizes (ES) will be expressed in terms of Hedges' g, calculated based on the adjusted mean difference between the intervention and control group (AMD) (controlling for prior attainment and the minimisation factors) from the multi-level model and the pooled unconditional variance obtained from the mixed model adjusting for trial arm and setting level clustering only (Var_{pooled}); obtained by the sum of the between- and within-cluster variance.

$$ES = \frac{AMD}{\sqrt{Var_{pooled}}}$$

The 95% CI will be calculated by dividing the 95% CI values for AMD, by the $\sqrt{Var_{pooled}}$ .The parameters used will be reported in the final report.

Binary outcome measure will be reported as the risk ratio (RR) and difference in percentages, for example (% of pass in intervention group) – (% of pass in the control group), in line with the EEF guidance.

Actual Analysis (COVID-19 affected)

The analysis that was planned for this trial, as described above, will not be undertaken as a result of the way the GCSE scores will be awarded in response to the COVID-19 pandemic. GCSE exams in Summer 2020 were cancelled (for first-time *and* resit exams) and teachers were instead asked to submit a grade for students for each of their GSCE subjects, based

on previous performance and attainment in mock exams, coursework, etc. These GCSE scores were then to be standardised by the exam boards according to prior performance of the setting. It was hypothesised that any effect the 5Rs program might have had, maybe lost, or at least significantly diluted, by this process and therefore not be detectable. As such, it was decided that use of results from the Summer 2020 GCSEs (May) was not feasible and so they were not collected, meaning the planned analysis could not be undertaken. Subsequently, the decision was made to award the GCSE scores primarily based on the teachers assessed grade, thus avoiding the standardisation progress. It is still hypothesised that these results may dilute any possible effect from the 5Rs programme, and as such the planned analysis will not be undertaken.

The primary outcome, GCSE score, and two of the three secondary outcomes, achievement of a grade four or higher and attendance at exams sessions, are affected by the way awarding of the GCSE score and as such were unavailable. The final secondary outcome, Student attitude towards maths, will not be collected as this was to be administered by the Settings in person, and it is believed that any response to the survey will be a very unrepresentative sample as it will have to be distributed by email or post.

However, data for the November exam sitting have already been provided by the settings; this was unaffected by the pandemic, and as such can still be used. Responses from 76 of the 78 settings (March 2020) indicated that 1337 pupils resat their maths GCSE in November 2019 (35% of the 3816 pupils for whom details were provided) – 652 from the intervention settings and 685 from the control settings.

The November resit data will be written up descriptively only. Continuous measures will be reported as a mean, standard deviation (SD), median, minimum and maximum, while categorical data will be reported as a count and percentage. No formal statistical comparisons will be undertaken.

The number of pupils resitting in each arm will be presented, and the proportion of pupils from each setting.

The number of papers the pupils sat will detailed, both as a continuous measure, but also the number and proportion who sat 3, 2, 1 and 0 (where they should have sat; not including those who were not due to sit the resit) will be given by arm.

As the GCSE raw marks will differ by exams board, these will still be standardised as previously described. The standardised score will be summarised by trial arm, alongside the number of students who passed. The number of pupils who sat each exams board will be detailed by trial arm. As a pupil could sit either the foundation paper (capped at a Grade 4) or the higher paper, the number of pupils sitting both will be broken down by trial arm, and exam board.

Additionally, baseline setting characteristics will be summarised for all settings as randomised and those who entered pupils in for a resit in November (i.e. those whose data will be presented descriptively), including type of setting, number of pupils resitting in the previous academic year, participation in BMP, and % of FSM at the setting level.

References

Allen, R., Jerrim, J., Parameshwaran, M., & Thompson, D. (2018). *Properties of commercial tests in the EEF database.* EEF Research Paper No. 001.

Altman, D. G., Royston P. 2006. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080. doi:10.1136/bmj.332.7549.1080

Borm, G. F., Fransen, J. & Lemmens, W. A. 2007. A simple sample size formula for analysis of covariance in randomized clinical trials. J Clin Epidemiol, 60, 1234-8.

Eldridge, S.M., Ashby, D. and Kerry, S., Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol, 2006. 35(5): p. 1292-300

Senn, S. 1994. Testing for baseline balance in clinical trials. Statistics in medicine, 13, 1715-1726.

Tapia, M., & Marsh, G. (2000). Attitudes Toward Mathematics Instrument: An Investigation with Middle School Students. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Bowling Green, KY, November 15-17, 2000).

Wears, R. L. 2002. Advanced statistics: statistical methods for analyzing cluster and cluster randomized data. Acad Emerg Med, 9, 330-41.