English Mastery
Statistical Analysis Plan

Evaluator: The National Centre for Social Research
Principal investigators: Valdeep Gill, Robert Wishart

Education Endowment Foundation

PROJECT TITLE	English Mastery
DEVELOPER (INSTITUTION)	Ark
EVALUATOR (INSTITUTION)	The National Centre for Social Research
PRINCIPAL INVESTIGATOR(S)	Valdeep Gill, Robert Wishart (formerly Priya Khambhaita, Martina Vojtkova)
PROTOCOL AUTHOR(S)	Anysia Nguyen, Josep Espasa, Robert Wishart, Enes Duysak, Ben Stocker
TRIAL DESIGN	Two-arm cluster randomised design using minimisation (adaptive randomisation technique)
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Years 7-8, Key Stage 3
NUMBER OF SCHOOLS	97
NUMBER OF PUPILS	5,820
PRIMARY OUTCOME MEASURE AND SOURCE	Pupil attainment in English (GL Progress Test in English)
SECONDARY OUTCOME MEASURE AND SOURCE	Hours spent teaching Year 7 English per week; Hours spent on non-teaching Year 7 English per week[1]; English Spelling, Punctuation and Grammar (SPAG)[2]; and reading comprehension[3].

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	09/11/19	N/A
2.0	26.11.20	Changes to the evaluation due to the Covid-19 pandemic. These changes include: • A possible two-level multilevel model for the primary analysis. This will be carried out if the decision rule to implement the preferred three-level multilevel model is not met. • A new Covid-19 optimal compliance measure.

Please note that the contents reflect the statistical analysis plan at time of writing (November 2020). Research activities may be subject to change as the Covid-19 pandemic evolves.

[1] From NatCen bespoke survey
[2] From GL Progress Test in English
[3] From GL Progress Test in English

Contents

Introduction

This analysis plan sets out the approach for the evaluation of the two-arm cluster-randomised controlled efficacy trial of English Mastery (EM). Due to the Covid-19 pandemic in March 2020, the evaluation has been subject to some changes. The analysis plan describes evaluation design and delivery as intended prior to the outbreak of the pandemic as well as highlights any changes to them.

The English Mastery intervention provides teachers with subject-specific training, curriculum materials and ongoing in-school support and coaching for two years from the start of Year 7 to the end of Year 8.

All non-academically selective, state secondary schools in England were eligible for the trial, except for:

- Schools with less than 4 hours of English per week in Year 7 or 8;
- Schools classified as grade 4 according to OFSTED school classification;
- Schools that had previously partnered with English Mastery.

The justification for establishing these eligibility criteria is that, according to English Mastery, a minimum of 4 hours of English per week was needed to implement the programme. Similarly, it was established that low performing schools (according to their OFSTED classification) were deemed not in a suitable condition to successfully implement the programme. The English Mastery team identified and recruited eligible schools, prioritising recruiting schools with high numbers of FSM pupils.

All English teachers delivering the programme receive an induction training session before the start of the school year. Each school is required to nominate a member of their English department to become the English Mastery ambassador and lead the intervention in their school. The English Mastery Ambassador and a member of the Senior Leadership Team receive an additional day of training that focuses on the leadership and implementation of the programme. This training is followed by optional termly subject mastery webinar sessions[4], termly in-school visits with ambassadors (which include coaching and bespoke CPD sessions), and termly Assessing for Mastery days.

In parallel to the training, teachers receive lesson plans and resources. English departments are granted access to the English Mastery Library, where co-planning guidance, lesson-by-lesson resources and units of work can be downloaded. In addition, there are at least three school visits per year to ensure the programme is implemented effectively. Schools that require further support with implementation and integration of English Mastery in their curriculum or timetable may receive additional visits.[5] This is implemented through teaching observations and discussions with members of staff that are at different levels of hierarchy within the English department and the wider schools that are involved in delivering the intervention.

English Mastery aims to implement a cumulative approach to curriculum design including bespoke assessment to improve English attainment of Year 8 pupils. It is composed of two pathways: (1) the traditional curriculum for children reaching age-related expectations, and (2) the foundation curriculum for those working below expectations. All pupils receive the same dosage and study the same topics, but the foundation curriculum is adapted to be accessible to lower attaining pupils (for example, using abridged texts). If pupils meet age-related expectations in two successive assessments, then it is recommended that they

[4] These sessions allow teachers and ambassadors to share their knowledge and get feedback on their own experience.
[5] This is not the typical approach for English Mastery, but were introduced to monitor fidelity during the evaluation.

graduate from foundation to traditional. This would ensure that the English Mastery programme would improve English attainment for all pupils, including disadvantaged pupils, at KS3 but also potentially at GCSE. By providing training and resources the theory of change hypothesised that English Mastery would reduce teacher workload.

The evaluation will be conducted as a two-arm cluster-randomised controlled trial with allocation at school level. Due to the length of the intervention (two years), it was important to let schools knows as soon as possible their allocation group so that they could organise their timetable and staffing. Consequently, schools were allocated to intervention or control using minimisation, an adaptive random allocation method.

The primary outcome of interest is English attainment as measured by the overall raw score on the GL Progress Test in English (level 13) at the end of Year 8.[6] There are four secondary outcomes, measuring: progress in (1) spelling, punctuation, grammar (2) reading comprehension, (3) teacher workload in class, (4) teacher workload outside of class. We may also conduct a further long-term follow-up analysis using GCSE attainment.[7]

Specifically, the evaluation of English Mastery aims to answer the following research questions:

Primary research question:

- RQ1. What is the impact of a two-year English Mastery programme on the overall English attainment of participating Year 8 pupils in England?

Secondary research questions:

- RQ2. How does the impact of a two-year English Mastery programme on the overall English attainment of participating Year 8 pupils in England differ by FSM status?
- RQ3. What is the impact of a two-year English Mastery programme on (i) the English Spelling, Punctuation and Grammar (SPAG) and (ii) reading comprehension attainment of participating Year 8 pupils in England?
- RQ4. What are the effects of the programme on teacher workload?
- RQ5. What are the longer-term effects of the programme on (i) KS4 English attainment and (ii) Attainment 8 for the participating pupils?

Exploratory research questions:

- RQ6. How does the impact of a two-year English Mastery programme differ by number of hours of English lessons?
- RQ7: How does the impact of a two-year English Mastery programme differ by Key Stage 2 prior attainment in English?
- RQ8: How does the impact of a two-year English Mastery programme differ by gender?

Analysis will investigate the following primary hypothesis on an intention-to-treat basis.

Primary analysis:

- H1: Participating in English Mastery improves Year 8 pupils' English attainment, as measured in GL's Progress Test in English.

Secondary analysis:

[6] https://www.gl-assessment.co.uk/support/pte-product-support/pte-test-level-age-guide/
[7] As measured by GCSE English Language (KS4_APELANG), GCSE English Literature (KS4_APELIT) and Attainment 8 (KS4_ATT8) results available from the NPD in 2023

- H2: Participating in English Mastery improves Year 8 pupils' spelling, punctuation and grammar, as well as reading comprehension, as measured by the GL subscales of PTE.
- H3: Receiving training and lessons materials reduces teacher workload in and out of class, as measured by a survey of English Mastery teachers.
- H4: Participating in English Mastery improves long-term attainment as measured by GCSE results.

Sub-group analysis:

- H5: English Mastery will have a different (higher or lower) impact on pupils ever eligible for Free School Meals (FSM) compared with those ineligible.
- H6: English Mastery will have a different (higher or lower) impact on pupils of different genders.
- H7: English Mastery will have a different (higher or lower) impact on pupils with different Key Stage 2 prior attainment in English[8].
- H8: English Mastery will have a different (higher or lower) impact on pupils in schools receiving different number of hours of English lessons per week[9].

Design overview

Trial design, including number of arms		Two-arm cluster randomised design using minimisation (adaptive randomisation technique)
Unit of randomisation		School
Stratification variables (if applicable)		Variables used at minimisation: -School attainment (Proportion of pupils achieving strong 9 to 5 in GCSE English and Maths) -School FSM (percentage of pupils eligible for FSM in school) - School size (total number of pupils)
Primary outcome	variable	KS3 English attainment
	measure (instrument, scale, source)	GL's Progress Test in English (PTE 13)
Secondary outcome(s)	variable(s)	Spelling, grammar and punctuation attainment Reading comprehension attainment Teacher workload Long-term outcome: KS4 Attainment
	measure(s) (instrument, scale, source)	GL Progress Test in English (PTE) SPAG subscale GL PTE reading comprehension subscale Teacher survey[10] GCSE English Language, GCSE English Literature and Attainment 8 from NPD
Baseline for primary outcome	**variable**	KS2 English attainment
	measure (instrument, scale, source)	KS2 SATS results from NPD

[8] As measured by KS2_ENGTOTMRK from NPD
[9] As measured by the teacher survey
[10] The survey will be collected at baseline, end of year 1 and end of year 2, using a bespoke teacher survey (for more details see Appendix 1)

Randomisation

Schools were allocated to the intervention or control group using minimisation. Minimisation was chosen as English Mastery is an intensive programme that has implications on staffing and timetables. Therefore, it was important to inform schools of their allocation as soon as possible, to facilitate the delivery of the programme during the school year. Minimisation has the flexibility of allocating schools in several stages, rather than waiting for the recruitment process to finish. Minimisation started at the end of March and ended in July 2019. Four rounds of minimisation took place during that period, as shown in Table 1.

Table 1 Allocation throughout recruitment

Batch	Number of schools allocated	Schools recruited until
First	37	22/03/2019
Second	28	01/05/2019
Third	20	07/06/2019
Fourth	12	28/06/2019

Minimisation is based on a different principle from other types of randomisation and has been demonstrated to be a valid alternative to ordinary randomisation methods (Treasure & MacRae, 1998). The first school is allocated to intervention or control using simple randomisation. All subsequent schools are given a probability of being assigned to each group depending on how that would affect the balance between the groups on a pre-defined set of observed school characteristics (minimisation variables).[11]

The 'minimisation' algorithm used three prognostic factors and a 50:50 allocation ratio between the intervention and control groups. The minimisation aimed to achieve balance across groups on the following factors:

- Mean proportion of pupils achieving strong 9 to 5 in both English and Mathematics GCSEs (schools' average between academic years 17/18 and 16/17).
- Proportion of pupils ever eligible for Free School Meals at any time during the past 6 years (measured at academic year 17/18)
- Total number of pupils in school in the latest academic year (measured at academic year 17/18)

According to prior research, each of these factors had an explanatory value for our outcome of interest. By implementing minimisation, we aimed to ensure that samples in intervention and control groups would be relatively balanced on those school characteristics. These factors received equal weight in the randomisation process.

The minimisation algorithm requires minimisation variables to be categorical. The variables were therefore recoded in the following way for the first and second batch of schools:

- The proportion of pupils achieving strong 9 to 5 in both English and Mathematics GCSEs was classified into three categories: schools whose proportion was below 31%, those between 31% and 75%, and those above 75%. The cut-offs for these categories corresponded to the 25th and 75th percentiles of the variable in the whole census of secondary schools in England;
- The factor for the proportion of pupils ever eligible for Free School Meals (FSM) was categorised into those schools with less than 29% of pupils in this group and those with more than 29%. This proportion was taken from computing the mean of the proportion of eligible FSMs in all secondary schools in England;

[11] A detailed explanation of the algorithm can be found in Pocock & Simon (1975).

- Similarly, the total number of pupils in school in the latest academic year was categorised into schools with less than 500 pupils in the last academic year and those with 500 pupils or more.

After the allocation of the second batch of schools was performed, a slight imbalance on the variable for school level prior attainment variable was observed.[12] Simulations performed at that stage indicated that adding a further category to the factor could potentially help reduce the imbalance. Thus, a modified version of the original variable was used in the third and fourth batches (recruited schools 56 to 97). The modified variable had four categories (below 31%; between 31% and 40.5%; between 40.5% and 75%; above 75%). The addition of this category should not impact the impact analysis.

In total, 49 schools were allocated to the intervention and 48 schools to control. Table 2 shows the balance obtained after the allocation on the original (continuous) variables. This indicates that after minimisation there are relatively small differences in the characteristics of schools.

Table 2 Balance achieved after allocation using continuous variables

Minimisation variable	Mean		SD	Difference	Effect Size
	Intervention	Control			
Proportion of pupils achieving 9-5 in GCSE English and Maths	35.4	34.1	11.95	1.3	0.11
Proportion of pupils eligible for FSM	34.4	35.8	14.63	-1.4	-0.10
Total number of pupils	912	905	402	7	0.02

The minimisation algorithm was implemented using the 'Minrand' R package (Jin et al, 2019). A seed for the random number generator was used to ensure replicability.

Sample size calculations overview

Table 3 outlines the sample calculations for this study. The sample size calculations were calculated in PowerUp! (Dong & Maynard, 2013) accounting for the three-level structure of this data.

The power calculations are based on the following assumptions:

- 80% statistical power
- A statistical significance level of 95% for a two-tailed test
- Explanatory power of baseline scores (R-squared) of 0.35 at pupil and 0.10 at school level. Based on figures for GL PTE and KS2 scores for the subject of English (Allen et al, 2018)
- Intra-cluster correlations (ICCs) of 0.10 and 0.15 for school- and class-level respectively. EEF guidance on ICCs (EEF, 2015) indicates that ICCs of 0.14 and 0.19 can be expected for English in KS2 and KS4 respectively. However, evidence

[12] The difference after the allocation of the second batch of schools was of 2.28 percentage points (~0.2 standardised differences), compared to the final 1.27 (reported in the table at the end of the section).

from other trials (such as Boylan et al, 2015) indicates that ICCs could be substantially higher. In addition, our assumptions are informed by Demack (2019) which provides evidence around ICCs that can be expected when accounting for clustering of pupils within classes and classes within schools.

- 60 children per school are randomly selected for follow-up testing

Table 3 Sample size calculations

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.19	0.21	0.20	0.22
Pre-test/ post-test correlations[13]	level 1 (pupil)	0.59	0.59	0.59	0.59
	level 2 (class)	0.00	0.00	0.00	0.00
	level 3 (school)	0.32	0.32	0.32	0.32
Intracluster correlations (ICCs)	level 2 (class)	0.15	0.15	0.15	0.15
	level 3 (school)	0.10	0.10	0.10	0.10
Alpha		0.05	0.05	0.05	0.05
Power		0.80	0.80	0.80	0.80
One-sided or two-sided?		2	2	2	2
Average cluster size		60	15*	60	15*
Number of schools	intervention	50	50	49	49
	control	50	50	48	48
	total	100	100	97	97
Number of pupils	intervention	3,000	750*	2,940	735*
	control	3,000	750*	2,880	720*
	total	6,000	1,500*	5,820	1,455*

* Recruitment of this project prioritised schools with a higher proportion of FSM pupils, therefore the estimated number of FSM pupils (25%) is anticipated to be higher than the national average (14% of the total enrolment for this age group in state funded schools) (DfE, 2018). This figure is taken from the percentage of Year 7 pupils in the intervention group classified as 'FSM' at baseline in the 'Multiplicative reasoning professional development programme' trial (Boylan, et al., 2015).

At the protocol stage, developers intended to recruit 110 schools to have an available sample size of 100 schools after an estimated attrition of 10%. Therefore, we expected our study to be powered to detect an effect of 0.19 standard deviations.[14]

[13] Our sample size calculations include estimates of the proportion of variance explained through the included covariates at each of these levels (R-squared). We have converted these into pre- post-test correlations by taking the square root of the R-squared value. School level variance is assumed to be explained by the inclusion of the stratification variables used at randomisation in the primary analysis model.

[14] These assumptions are in line with those set out in Torgerson and Torgerson (2013) for scenarios when the same pre-test and post-test are used.

In total, 97 schools were recruited and randomised: 49 in the intervention and 48 in the control group.[15] Thus, we now estimate the study to be powered to detect an effect of 0.20 standard deviations.[16]
As the trial lasts for two years, further attrition at school and pupil level during the trial is to be expected. Therefore, we assume a school-level attrition of ten percent and twenty percent at pupil level.

To minimise attrition at pupil level, all pupils will be randomly ordered within schools. The first 60 pupils in each school will be selected for follow-up testing. The remaining pupils will be selected in order (based of the random ordering used to select the first 60 pupils) if any of the first 60 randomly selected pupils cannot sit the test (for example, if they are off sick or left school).

Analysis

The evaluation of English Mastery aims to evaluate its impact on the English attainment of Year 8 pupils in England and how it differs by FSM eligibility, gender, prior attainment and number of English lessons received each week.

Primary outcome analysis

The trial was designed as a two-armed, three-level randomised controlled efficacy trial. Pupils (level 1) are nested in classes (level 2) within schools (level 3). The primary analysis will use a multi-level model to account for the hierarchical data structure.

The primary analysis will estimate the intervention's impact on enrolled Year 8 pupils' English attainment, as measured by the GL's Progress test in English (raw scores), using an intention-to-treat approach. The test will be administered in June/July 2021 to all pupils that have been randomly sampled from participating schools.

Following EEF guidance, evidence of efficacy and reported effect sizes will be obtained from a baseline-adjusted model, in which the dependent variable is the raw score of the GL Progress Test in English. The effect size will be estimated through a multi-level linear model containing a dummy indicator of treatment allocation, a derived variable for pupils' KS2 attainment in English and the variables used at minimisation. The KS2 attainment variable will be derived by combining the raw scores from the reading test and Grammar, Punctuation and Spelling test, sourced from the National Pupil Database (KS2_READMARK and KS2_GPSMARK respectively). The model will also contain random intercepts for the variance at school and class level to account for the clustering of pupils within classes and classes within schools.

The basic form of the model is:

$$PTE_{ijk} = \beta_0 + \beta_1 Baseline_{ijk} + \beta_2 Intervention_k + \beta_3 attainment_k + \beta_4 FSM_k + \beta_5 size_k + v_k + u_{jk} + e_{ijk}$$

Where pupils (i) are clustered in classes (j) within schools (k). The intervention effect is estimated by β_2. $attainment_k$, FSM_k , $and\ size_k$ represent the categorical variables used at minimisation[17]. The term v_k is a school-level random effect, u_{jk} a class-level random effect, and e_{ijk} the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. In line with the EEF analysis guidance, other additional covariates will not be considered at this stage. The three-level multilevel model will be followed if class level data is missing for less than 5% of pupils with outcome data.

[15] At this stage we were able to use a harmonic mean (6.47) for the number of pupils in classes, rather than arithmetic mean (8) to calculate our sample sizes
[16] At the time of writing this SAP, we are aware of some drop-outs, so MDES in the final analysis will be higher. However, for the SAP we follow EEF guidance and report power calculations based on the number of schools randomised.
[17] We will use the revised versions of these categorical variables used in the third and fourth minimisation batches.

If the decision rule to implement the preferred three-level multilevel model is not met, a two-armed, two-level randomised controlled efficacy trial, where pupils (level 1) are nested in schools (level 2), will be used. The primary analysis will use a multi-level model to account for the hierarchical data structure.

The primary analysis will estimate the intervention's impact on enrolled Year 8 pupils' English attainment, as measured by the GL's Progress test in English (raw scores), using an intention-to-treat approach. The test will be administered in June/July 2021 to all pupils that have been randomly sampled from participating schools.

Following EEF guidance, evidence of efficacy and reported effect sizes will be obtained from a baseline-adjusted model, in which the dependent variable is the raw score of the GL Progress Test in English. The effect size will be estimated through a multi-level linear model containing a dummy indicator of treatment allocation, a derived variable for pupils' KS2 attainment in English and the variables used at minimisation. The KS2 attainment variable will be derived by combining the raw scores from the reading test and Grammar, Punctuation and Spelling test, sourced from the National Pupil Database (KS2_READMARK and KS2_GPSMARK respectively). The model will also contain random intercept for the variance at school level to account for the clustering of pupils within schools.

The basic form of the model is:

$$PTE_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 attainment_j + \beta_4 FSM_j + \beta_5 size_j + u_j + e_{ij}$$

Where pupils (i) are clustered within schools (j). The intervention effect is estimated by β_2. $attainment_k$, FSM_k , $and\ size_k$ represent the categorical variables used at minimisation[18]. The term u_j is a school-level random effect, and e_{ij} the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. In line with the EEF analysis guidance, other additional covariates will not be considered at this stage.

It is important to note that using a two-level model, rather than a three-level model, may have possible implications for this evaluation. In a review of existing evidence, Demack (2019) showed that ignoring the class-level matters if class-level clustering is present. This is especially apparent in English secondary schools, where within-school pupil segregation is a common feature. Failing to account for this class-level clustering may mean we underestimate our standard errors, bring hidden bias into the design, and make a false-positive finding more likely.

At the time of writing this analysis plan, we are aware that some participating schools are planning on using GL PTE tests to assess their children in Year 7 and / or 8. Even if pupils have sat an earlier version of the test (such as GLE PTE 12), they may have an advantage over peers who have not sat a GL test before. Therefore, our fully adjusted sensitivity analysis (described in additional analyses) will contain a covariate that controls for whether the schools have used any GL PTE test previously as this is may affect how well children perform in the follow-up test. All analysis will be implemented in Stata 16 SE-64.

Secondary outcome analysis

There will be four secondary outcome analyses. Two analyses will explore the impact of the intervention on subscales of the GL PTE test: (1) spelling, grammar and punctuation, and (2) reading comprehension. The remaining two analyses will explore how the intervention affected teacher workload in and outside of class.

[18] We will use the revised versions of these categorical variables used in the third and fourth minimisation batches.

The secondary analysis for the pupil outcomes will be estimated using a similar multi-level model as described in the primary outcome analysis section.

The basic form of the model using the three-level multi-level model will be followed if class level data is missing for less than 5% of pupils. The model is as follows:

$$PTE_{subscale_{ijk}} = \beta_0 + \beta_1 Baseline_{ijk} + \beta_2 Intervention_k + \beta_3 attainment_k + \beta_4 FSM_k + \beta_5 size_k + v_k + u_{jk} + e_{ijk}$$

The basic form of the model using the two-level multi-level model will be followed if class level data is missing for more than 5% of pupils. The model is as follows:

$$PTE_{subscale_{ij}} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 attainment_j + \beta_4 FSM_j + \beta_5 size_j + u_j + e_{ij}$$

Teacher workload will be derived from a bespoke survey of teachers administered at the start of year 7 (round one), the end of year 7 (round two, cancelled due to Covid-19 – see more below) and the end of year 8 (round three) (see Appendix 1).[19] The survey contains questions concerning:

1. The overall number of hours spent on teaching and all teaching related activities in schools weekly
2. Hours spent teaching Year 7 English weekly
3. Hours spent on different non-teaching tasks involving Year 7 English:
    - Preparing lessons
    - Developing resources
    - Administering and marking work
    - General non-curriculum related administrative work relating to Year 7 English
    - Co-planning and dialogue with colleagues within school about Year 7 English teaching
    - Coordinating with people or organisations outside of school in relation to Year 7 programmes or interventions which aim to improve ability in English
4. The workload compared to the same time last year (e.g. increased, decreased, stayed the same)

The secondary analysis will use the following items:

- Hours spent teaching Year 7 English weekly (question two), and
- Total hours spent on non-teaching tasks involving Year 7 English weekly (question three)

Descriptive statistics will be conducted as part of the analysis to investigate changes over time in the hours spent teaching Year 7 English. For question two, we will derive change score to capture the overall change in workload over the two-year period (survey round one and survey round three). Round 2 of the Teacher Workload Survey was originally planned to take place in May 2020. This would have allowed us to capture the incremental change in teacher workload in year 1 (survey round 1 and survey round 2 measures), and the incremental change in teacher workload in year 2 (survey round 2 to survey round 3 measures). However, schools and teachers were under a great deal of pressure following the changes to schooling as a result of the Covid-19 pandemic. The Wave 2 survey did not collect workload information in recognition that teachers' day-to-day duties were changed and atypical during this time (for example provision of home learning in place of in-school teaching). We will use a two-independent-sample t-test to compare differences between

[19] Note that round one survey is implemented after English Mastery has been initiated. Therefore, round 1 cannot count as a true baseline measure.

intervention and control groups as the teacher workload will be a continuous variable measured in hours.

Changes in teacher workload will be estimated using a multilevel linear regression with teachers (level 1) nested within schools (level 2). This analysis will be conducted on an intention-to-treat basis.

The basic model would be:

$$Workload\_in\_class_{jkt} = \beta_0 + \beta_1 workload_{jkt-1} + \beta_2 Intervention_k + \beta_3 attainment_k + \beta_4 FSM_k + \beta_5 size_k + v_k + e_{jk}$$

In the equation above, (j) represents teacher level outcomes, and (k) represents school level outcomes, (t) represents the survey round. The intervention effect is estimated by β_2, while β_3, β_4 and β_5 represent the variables used at minimisation, v_k is a school-level random effect and e_{jk} the teacher-level error term.

For the survey question three, we will sum answers of all sub-components to create a composite score illustrating the total number of hours spent on non-teaching activities relating to Year 7 English. Change scores will be derived in a similar fashion as above for the descriptive analysis. As change scores are continuous, we will use two-independent-sample t-tests to compare differences between intervention and control groups.

We will also conduct a regression using a similar analytical approach as the one above:

$$Workload\_out\_class_{jkt} = \beta_0 + \beta_1 workload_{jkt-1} + \beta_2 Intervention_k + \beta_2 attainment_k + \beta_3 FSM_k + \beta_4 size_k + v_k + e_{jk}$$

We will report Hedges' g estimates for all the secondary analyses described above. Results will be interpreted cautiously in terms of their generalisability.

The process evaluation will focus on questions one, four and the individual components of question three.

The secondary analysis models assume that most teachers surveyed at all time points are the same. However, it may be that between time points, some teachers have changed roles within schools, or left schools entirely. If a majority of teachers differ across time points, this will mean a lower correlation between baseline and follow-up, thereby making the change variable less meaningful. We will report the response rates to each survey and report the proportion of teachers responding to surveys at both time points.

We will also run an adjusted sensitivity analysis which will include a dummy variable for whether teachers have changed between time points in the model to compensate for possible teacher turnover and response rate.[20]

Subgroup analyses

The subgroup analyses will explore the following hypothesis:

- H5: English Mastery will have a different (higher or lower) impact on pupils ever eligible for Free School Meals (FSM) compared with those ineligible[21].
- H6: English Mastery will have a different (higher or lower) impact on pupils of different genders.

[20] Whilst multiple imputation was considered, it is unlikely that suitable auxiliary variables will exist to robustly implement imputation.

[21] The NPD-derived variable for being ever eligible to FSM in the last six years will be used: EVERFSM_6_P

- H7: English Mastery will have a different (higher or lower) impact on pupils with different Key Stage 2 prior attainment in English[22].
- H8: English Mastery will have a different (higher or lower) impact on pupils in schools with different number of hours of English lessons per week[23]

English Mastery uses an alternative pathway for lower performing pupils and we therefore hypothesise that the intervention may be more effective for these pupils, reducing attainment gaps. This includes pupils ever eligible for Free School Meals and boys, who are both less likely to reach the expected standard in English at KS2.

English Mastery was originally designed as an intervention to be delivered in schools with six hours a week of English lessons each week but was restructured for the evaluation to include schools with four hours of English lessons each week. We therefore also hypothesise that the intervention will be more effective in schools with more than four hours of English lessons each week.

Subgroup impacts on the primary outcome will be estimated for pupils ever eligible for FSM. This will involve the re-estimation of the model described in the primary outcome section with the addition of the pupil-level FSM indicator and an interaction term combining pupil FSM eligibility and treatment allocation. Following EEF guidance (2018), a separate model will also be estimated and reported for each subgroup.

The model specification using a three level multi-level model, which is used if class level data is missing for less than 5% of pupils, is as follows:

$$PTE_{ijk} = \beta_0 + \beta_1 Baseline_{ijk} + \beta_2 Intervention_k + \beta_3 FSM\_pupil_{ijk} + \beta_4 Intervention * FSM_{jk} + \beta_5 attainment_k + \beta_6 FSM_k + \beta_7 size_k + v_k + u_{jk} + e_{ijk}$$

The model specification using a two level multi-level model, which is used if class level data is missing for more than 5% of pupils, is as follows:

$$PTE_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 FSM\_pupil_{ij} + \beta_4 Intervention_j * FSM_j + \beta_5 attainment_j + \beta_6 FSM_j + \beta_7 size_j + u_j + e_{ij}$$

Similar interaction models will be estimated for each of the following subgroups:

- Gender
- Number of hours of English lessons
- KS2 prior attainment in English.

Results of subgroup analyses should be considered as indicative only.

Additional analyses

A range of sensitivity analyses will also be carried out as additional analyses to explore the robustness of the main primary outcome findings, with findings for all models transparently reported. If a sensitivity analysis finds any substantively different findings to the main analysis, this will be acknowledged. The following analyses will be carried out:

- An adjusted model, adding a wider range of explanatory variables to control for potential imbalance at baseline: pupil's FSM status, pupil's gender and term of birth,

[22] As measured by KS2_READMARK and KS2_GPSMARK scores from NPD. The pass marks are 28 in Reading English, and 36 in Grammar, punctuation and spelling (DfE, 2019). Therefore, lower achievers will be those scoring below 64 in the combined score of prior attainment in English.
[23] As measured by the teacher survey at round 1. We will take a meaningful cut-off of four hours as this is the minimum number of hours suggested by EM. Should more than 80% of the schools teach more than four hours, we will use the median as a cut-off.

school setting[24], school type, a dummy variable indicating if the school has used a GL PTE test with this cohort of pupils, a dummy variable indicating if the school is also implementing Maths Mastery,[25] and a dummy variable indicating whether teachers have changed since the start of the trial.

- A single-level OLS regression model including cluster-robust standard errors. The variables included will be the same as the primary analysis model: baseline attainment, treatment allocation and minimisation variables.

Longitudinal follow-up analyses

If long term follow-up analysis is conducted, it will explore three outcomes: GCSE English Language, GCSE English Literature and Attainment 8.[26] These outcomes will be sourced from the NPD in 2023. The first two models will assess whether English Mastery has had an impact in the long-term on English, whilst the last analysis will assess its long-term impact on other subjects (including English).[27]

All three analyses will use a multi-level model using the same specification as the primary analysis.[28] No additional analysis will be conducted on these outcomes. The results will be published as an addendum to the final report, if it is conducted. The decision to proceed with this analysis is conditional on the evaluation finding that no significant issues with implementation or experimental effects are observed (e.g. control group contamination, important differential attrition).

Imbalance at baseline

Minimisation, if conducted correctly, should result in there being no important difference between intervention and control groups in the main determinants of our outcomes of interest. We will explore the potential for imbalance first through an inspection of the descriptive statistics of baseline and minimisation variables, comparing intervention and control groups for all units as randomised and as analysed. The former informs whether randomisation was successful at obtaining a balanced sample, while the latter provides evidence of whether attrition might have introduced imbalance. These baseline characteristics and strata variables will be summarised by intervention and control group.

At school level, the comparison will include:

- School attainment (Percentage of pupils achieving strong 9 to 5 in both English and Mathematics GCSEs)
- School's proportion of pupils ever eligible for FSM in the past six years
- School size (total number of pupils in school)
- OFSTED rating
- Urban-rural indicator
- School type

[24] A binary indicator of whether schools stream or set their classes was collected as part of our school recruitment process.

[25] A small number of schools (both in the intervention and control group) are known to be implementing "Maths Mastery". In theory, the implementation of English Mastery may be strengthened in schools also implementing Maths Mastery for intervention schools. In control schools, there may be a higher risk that the techniques used in Maths Mastery are transferred to other subjects, such as English.

[26] This is subject to EEF funding long-term follow-up analysis.

[27] Attainment 8 (KS4_ATT8) measures the achievement of a pupil across 8 Key Stage 4 qualifications including mathematics, English, three further qualifications that count in the English Baccalaureate (EBacc) measure and three further qualifications that can be GCSE qualifications (including EBacc subjects) or any other non-GCSE qualifications on the DfE approved list. For more detail on how these are calculated see: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/561021/Progress_8_and_Attainment_8_how_measures_are_calculated.pdf

[28] Currently, class level indicators are only going to be collected for pupils completing follow-up testing in Year 8. Using a three-level model would either restrict the analysis sample to these pupils, or further data collection would be required. Alternatively, the long-term follow-up analysis could use a two-level model with pupils nested in schools.

At pupil level, the following baseline comparisons will be presented:

- Ever eligible for FSM
- Gender
- KS2 English attainment

Imbalance on baseline covariates between the intervention and control groups in the sample as randomised and as analysed, will be presented as Hedge's g effect sizes. An effect size greater than 0.05 will be considered as an indication of possible imbalance.

Missing data

For the primary analysis, it is possible that there may be loss to follow-up due to moves and other external factors influencing participation in follow-up testing and this risk may be exacerbated by the length of the intervention (two years).

Baseline data will be sourced from NPD. Very low levels of attrition may occur if the pupils cannot be linked, or if they are missing at baseline.[29] Given the available information on the retention of schools to the trial, we anticipate moderate attrition at school level.

We will explore the extent and pattern of missingness for the primary outcome if data is missing for greater than five percent of pupils. The number of pupils with missing outcome data will be reported by intervention allocation. To explore if there is a pattern to missing data, we will estimate a logistic regression with a binary indicator of the outcome being missing as the dependent variable, with a suite independent variables, including all the characteristics explored for baseline balance (free school meal eligibility, gender, baseline attainment) and minimisation variables. If covariates are not associated with missing outcome data, multiple imputation will not be feasible, and the implications discussed in the evaluation report.

However, if covariates predict missing outcome data, we will conduct multiple imputation (MI) to infer the likely results of those lost to follow-up as a sensitivity analysis. The MI analysis will draw on all available information from the trial data. To account for clustering of pupils within settings, data will be imputed separately for each cluster, using the by option. The first 200 iterations of the imputation will not be used ('burn in').[30] This ensures that iterations have converged to a stationary distribution. In total, 100 imputed datasets will be created.[31]

Compliance

Whilst intention-to-treat (ITT) analysis is informative to policymakers about the effects of an offer of treatment, it is not informative about the impact of an intervention on those who receive it. Consequently, the trial analysts propose conducting analysis of non-compliance at the school level across the two-year period, using yearly compliance assessments.

There are several potential areas for non-compliance issues in this trial. Non-compliance could arise because of:

- Staff not attending induction training
- English Mastery Ambassadors not attending the termly Assessing for Mastery training days.
- Teachers not delivering lessons using materials provided
- English Mastery Ambassadors not delivering some of the English Mastery curriculum

[29] However, as we use unique identifiers (e.g. names and date of birth), to link NPD data to our pupils, a small loss can occur due to erroneous identifier data supplied by schools.
[30] We will assess the Monte-Carlo Markov Chain (MCMC) diagnostic plots to assess if a stable distribution has been reached. If a larger burn-in period is required, the analysis will be adjusted accordingly.

- Co-planning time not being allocated on regular basis
- Pupils not sitting the termly English assessment in standardised conditions.
- Control schools delivering the intervention or elements of the intervention
- Covid-19 pandemic

The following table provides compliance definition and thresholds for the measures we will use to analyse compliance:

Item	Definition	Compliance threshold	COVID-19 compliance threshold compliance threshold
1. Induction training attendance	Teachers delivering English Mastery attend the English Mastery induction training.	80% of teachers delivering English Mastery attend induction training. (In year 2 of the programme, the percentage trained will be calculated based on whether teachers attended in either year 1 or year 2).	80% of teachers delivering English Mastery attend induction training.
2. Assessing for Mastery training attendance	English Mastery Ambassadors attend the termly Assessing for Mastery training days.	English Mastery Ambassador attends at least two out of three Assessing for Mastery training days per year.	English Mastery Ambassador attends termly Assessing for Mastery training. These are equivalent to the MyMastery+ standardisation zoom sessions
3. Co-planning time allocation	Co-planning time is allocated on a regular basis.	Co-planning is allocated fortnightly or more regularly for at least 10 weeks per term.	Co-planning time is allocated on a regular basis (12 sessions in 2020/21).
4. Standardised conditions for assessment	Pupils sit the termly English Mastery assessments in standardised conditions.	Pupils sit at least two out of three English Mastery assessments per year in standardised conditions.	Pupils complete the English Mastery quiz (one quiz per term)
5. English Mastery curriculum delivery	Schools teach all components of the English Mastery curriculum.	School teach the Literary Heritage curriculum component for at least 100 mins+ per week for at least 10 weeks per term, and the Mastery Writing component for at least 50 mins per week for at least 10 weeks per term.	The school follows English Mastery's revised recommended structure (four lessons per week and with recommended content)
6. English Mastery Ambassador involvement	The English Mastery Ambassador teaches the English Mastery curriculum.	English Mastery Ambassador teaches at least two lessons per week for a minimum of 10 weeks per term.	The English Mastery Ambassador teaches two lessons per week for 27 weeks of the year [70%].

These items were identified by English Mastery as fundamental elements of the intervention. Their assessment by English Mastery forms part of a wider assessment of compliance and fidelity conducted by English Mastery links, via termly school visits with each school participating in the programme. During these school visits, English Mastery links observe and discuss with teachers to assess each compliance components.[32] Compliance

[32] Respondents could change their answers and behaviours due to being observed, but we judge this unlikely as a good rapport between English Mastery and schools will exist by then, and NatCen and English Mastery would explain to teachers that their answers are only for research and will not affect their roles or schools.

judgements for each year will be based on English Mastery's assessment of compliance across all three terms each year.[33]

For the compliance analysis, three measures of compliance will be constructed according to the extent of fidelity of implementation of key elements of the English Mastery Intervention as outlined in the "compliance threshold" column above.

1) **optimal compliance**, where a school would be judged to be fully compliant if and only if it meets or exceeds the compliance threshold on all 6 items in both year 1 and 2 (and judged non-compliant otherwise); and

2) **partial compliance**, where a school would be judged to be partially compliant if it meets or exceeds the compliance threshold on a minimum of 3 out of 6 items in both year 1 and year 2 (and judged non-compliant otherwise).

3) Covid-19 optimal compliance, a new measure that has been designed to account for the challenges schools may encounter to full EM implementation as a result of the Covid-19 pandemic and resulting school pressures. This measure will capture whether the critical components of the English Mastery's adapted programme have been met. A school would be judged to be fully compliant if and only if it meets the overall compliance threshold on all 6 items. Further details of these critical components can be found in the table above.

In each case, the compliance measure will be binary, with each intervention school having a value of either zero or one on each of these measures. We assume that none of the control group schools can feasibly implement more than one or two of the six English Mastery elements outlined above as part of their business as usual approach to teaching and assessments. Given that this is insufficient to reach the proposed thresholds for a control school to be considered non-compliant, we will assume one-sided non-compliance in our analysis[34].

To analyse the impact of English Mastery in the presence of non-compliance, a two-stage least squares (2SLS) instrumental-variable (IV) regression will be implemented. Compliance will be instrumented by intervention allocation (Angrist & Imbens, 1995). The first stage equation will take the following form:

$$Comply_k = \propto + \beta_1 Treat_k + \varepsilon_{ik}$$

The predicted values from the first stage equation, $\widehat{Comply_J}$, will then be used in the estimation of the second stage equation, as follows:

$$Y_{ik} = \propto + \beta_1 \widehat{Comply}_k + \beta_3 Baseline_{ik} + +\beta_4 attainment_k + \beta_5 FSM_k + \beta_6 size_k + \omega_{ik}$$

IV regression will be conducted in Stata using the **ivregress** command. It is not possible to account for clustering at level two (classes) and at level three (schools) using this approach. However, cluster robust standard errors will be estimated with the **cluster** option to account for clustering of pupils within schools. Tests for endogeneity will be conducted to test if treatment allocation is a suitable instrument (Wooldridge, 1995) and following EEF guidance

[33] Please note, as school visits will be taking place at different times for different schools over the course of each term, English Mastery's compliance assessment will be to some extent based on reasonable expectation of compliance based on best available evidence at the time of the school visit. For example, if a school visit is taking place in week 5 of the term, compliance with the co-planning element of the intervention will be judged based on existing evidence on co-planning practice to date. E.g., English Mastery will use the school visit to assess whether co-planning is well established, regularly occurring and whether there is reasonable expectation that this will continue. If the school has not started co-planning at time of the school visit, English Mastery will conduct a follow-up with the school later in the term to review progress towards and achievement against the termly compliance threshold.

[34] This is in line with a substantive assessment by English Mastery that implementation of teaching approaches resembling one or two of the six English Mastery elements would be insufficient to deem the school to be implementing English Mastery in any meaningful way.

(EEF, 2018) the F-statistics and p-value will be reported. If compliance is exogenous, the analysis will be conducted using multi-level linear regression.

Intra-cluster correlations (ICCs)

The intra-cluster correlations (ICCs) will be calculated directly from the primary analysis model, using the unadjusted variance estimates for each level of clustering. This will be estimated using a multi-level model, either using a three-level or two-level structure, dependent on the same decision rule as the primary analysis (less than 5% of class identifiers missing for a three-level analysis). Pupils (level 1) are nested within classes (level 2) within schools (level 3) in the three-level structure, or pupils (level 1) within schools (level 2) for the two-level structure. The dependent variable will be attainment in English, with a dummy variable for treatment allocation as the sole independent variable, with random intercepts for classes and schools.

In three-level models, two indices are necessary to characterise the relations between the variances associated with the three levels. These are the school level ICC ρ_S, and the class level ICC ρ_C. The formulae are as follow:

$$\boldsymbol{\rho_S} = \frac{\boldsymbol{\sigma^2_{BS}}}{\boldsymbol{\sigma^2_{BS}} + \boldsymbol{\sigma^2_{BC}} + \boldsymbol{\sigma^2_{WC}}} = \frac{\boldsymbol{\sigma^2_{BS}}}{\boldsymbol{\sigma^2_{WT}}}$$

$$\boldsymbol{\rho_C} = \frac{\boldsymbol{\sigma^2_{BC}}}{\boldsymbol{\sigma^2_{BS}} + \boldsymbol{\sigma^2_{BC}} + \boldsymbol{\sigma^2_{WC}}} = \frac{\boldsymbol{\sigma^2_{BC}}}{\boldsymbol{\sigma^2_{WT}}}$$

In these formulae $\boldsymbol{\sigma^2_{BS}}$ represents the between-school variance, $\boldsymbol{\sigma^2_{BC}}$ the between-class variance, $\boldsymbol{\sigma^2_{WC}}$ the within-class variance and $\boldsymbol{\sigma^2_{WT}}$ the sum of the variance at all levels.

In the two-level model, the ICC, ρ, will be estimated with the post-test estimation command *estat icc* in Stata 16.1 SE-64 using the following formula:

$$\rho = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_W} = \frac{\sigma^2_B}{\sigma^2_T}$$

Where σ^2_B, is the between-school variance, σ^2_W is the within-school variance. Values of ρ range from zero to one, where values closer to zero implies that the within-cluster variance is much greater than the between cluster variance.

Effect size calculation

As outlined throughout this SAP, we may estimate the effect size either from a three-level or two-level model depending on the outcome of the decision rule. We present both sets of effect size formulae here.

Three-level

The estimates of the impact will be reported using as Hedges' g. Hedges (2011) constructed formulae for effect sizes for three-level cluster randomised trials.

The point estimate, g_{WT}, is calculated as the difference between adjusted group means $\bar{Y}^T_{adj}$ and $\bar{Y}^C_{adj}$, scaled by the unconditional total standard deviation within-treatment groups S_{WT}, and adjusted to account for school and class-level clustering, as follows:

$$g_{WT} = J \times \left(\frac{\bar{Y}^T_{adj} - \bar{Y}^C_{adj}}{S_{WT}}\right)\sqrt{1 - \frac{2(p_U - 1)\rho_S + 2(n_U - 1)\rho_C}{N - 2}}$$

Where *J* is the bias correction to estimate Hedges' *g* from Cohen's *d*, given by:

$$J = 1 - \left(\frac{3}{4(n_T + n_C - 2) - 1}\right)$$

The standard deviation is the square root of the estimated pooled variance, S^2_{WT}, calculated as:

$$S^2_{WT} = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}\sum_{k=1}^{n_{ij}^T}\left(Y_{ijk}^T - \bar{Y}^T\right)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}\sum_{k=1}^{n_{ij}^C}\left(Y_{ijk}^C - \bar{Y}^C\right)^2}{N - 2}$$

In these formulae, the subscripts i, j and k represent pupils, classes and schools respectively.

The school intra-cluster correlation ρ_S, and the class intra-cluster-correlation ρ_C, are given by the formulae in the previous section above. The remaining terms are calculated as follows:

$$p_U = \frac{N^C\sum_{i=1}^{m^T}\left(\sum_{j=1}^{p_i^T} n_{ij}^T\right)^2}{NN^T} + \frac{N^T\sum_{i=1}^{m^C}\left(\sum_{j=1}^{p_i^C} n_{ij}^C\right)^2}{NN^C}$$

$$n_U = \frac{N^C\sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}\left(n_{ij}^T\right)^2}{NN^T} + \frac{N^T\sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}\left(n_{ij}^C\right)^2}{NN^C}$$

$$N = N^T + N^C = \sum_{i=1}^{m^T}\sum_{j=1}^{P_i^T} n_{ij}^T + \sum_{i=1}^{m^C}\sum_{j=1}^{P_i^C} n_{ij}^C$$

The 95% confidence intervals will be calculated as follows:

$$g_{WT} - 1.96v_g \leq \delta_T \leq g_{WT} + 1.96v_g$$

The variance of the effect size estimate, v_g, can be conservatively approximated by:

$$v_{\{g_{WT}\}} = \frac{(1 + (p_U - 1)\rho_s + (n_U - 1)\rho_C)}{\widetilde{N}} + \frac{g^2_{WT}}{2(M^T + M^C - 2)}$$

Where M^T and M^C the number of schools in the intervention and control groups respectively. Finally, $\widetilde{N}$ is given by:

$$\widetilde{N} = \frac{N^T N^C}{N^T + N^C}$$

Two-level:

The Hedge's *g* effect size will be estimated following Hedges' (2007) formula for the effect size d_t for designs with unequal sample sizes. The effect size, g_t is estimated as follows:

$$g_t = J \times \left(\frac{\bar{Y}^T - \bar{Y}^C}{S_T}\right)$$

Where $\bar{Y}^T$ and $\bar{Y}^C$ are the grand means of the intervention and control groups.

The correction factor J is defined as:

$$J = 1 - \left(\frac{3}{4(N^T + N^C - 2) - 1}\right)$$

The pooled standard deviation, S_T is defined as:

$$S_T = \sqrt{\frac{\sum_{i=1}^{m^T}\sum_{j=1}^{n_i^T}(Y_{ij}^T - \bar{Y}^T)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{n_i^C}(Y_{ij}^C - \bar{Y}^C)^2}{N-2}}$$

In order to calculate confidence intervals, we will estimate the variance term as follows:

$$V\{g_t\} = \left(\frac{N_T + N_C}{N_T N_C}\right) + \frac{{g_t}^2}{2(N-2)}$$

The confidence intervals will be calculated as follows:

$$g_t - 1.96\, V\{g_t\} \leq \delta T \leq g_t + 1.96\, V\{g_t\}$$

References

Allen, R., Jerrim, J., Parameshwaran, M. & Thompson, D., 2018. *Properties of commercial tests in the EEF database,* London: Enducation Endowment Foundation.

Angrist, J. D. & Imbens, G. W., 1995. Two-Stage Least Squares estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association,* 90(430), pp. 431-442.

Boylan, M. et al., 2015. *Multiplicative reasoning professional development programme: evaluation Technical report & detail of evaluation findings,* s.l.: Department for Education.

Cohen, J. & Cohen, P., 1983. *Applied multiple regression/correlation analysis for the behavioral sciences.* s.l.:s.n.

Demack, S., 2019. *Does the classroom level matter in the design of educational trials? A theoretical & empirical review,* s.l.: Enducation Endowment Foundation.

DfE, 2018. *Schools, Pupils and their Characteristics: January 2018 - National tables.* [Online]
Available at: https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2018

DfE, 2019. *2019 scaled scores at key stage 2.* [Online]
Available at: https://www.gov.uk/government/publications/2019-scaled-scores-at-key-stage-2

Dong, N. & Maynard, R., 2013. PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs.. *Journal of Research on Educational Effectiveness,,* pp. 24-67.

EEF, 2018. *Statistical analysis for EEF evaluations,* s.l.: Education Endowment Foundation.

Hedges, L. V., 2007. Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics,* 32(4), pp. 341-370.

Hedges, L. V., 2011. Effect Sizes in Three-Level Cluster-Randomized Experiments. *Journal of Educational and Behavioral Statistics,* pp. 346-380.

Jin, M., Polis, A. & Hartzel, J., 2019. Algorithms for minimization randomization and the implementation with an R package.. *Communications in Statistics - Simulation and Computation.*

Pocock, S. & Simon, R., 1975. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics,* 31(1), pp. 103-115.

Richburg-Hayes, L., Rebeck-Black, A. & Bloom, H., 2007. 'Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions'.. *Educational Evaluation and Policy Analysis.*

Torgerson, C. J. & Torgerson, D. J., 2013. *Randomised Trials in Education: An Introductory Handbook.* s.l.:EEF.

Treasure, T. & MacRae, K. D., 1998. Minimisation: the platinum standard for trials?. *BMJ,* pp. 362-363.

Wooldridge, J. M., 1995. Score diagnostics for linear models estimated by two stage least squares. In: P. C. Maddala, B. Phillips & T. N. Srinivasan, eds. *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao.* Oxford: Blackwell, pp. 66-87.

Appendix 1: Teacher workload survey

Teacher Workload

Intro 1
The next set of questions is about the time you spend on different tasks.
Some questions will ask for an estimate of time spent in hours in your most recent week that you were working (that is, Monday to Sunday that was not shortened by illness, religious breaks or public holidays).

Q1a. Teachhrs
In the most recent week that you were working, approximately how many hours did you spend in total on **teaching and all teaching related activities in this school?**

This would include planning lessons, marking, covering for absence, interacting with other teachers, participating in staff meetings, pastoral care, professional development training, parents' evenings and other activities.
Please include tasks that took place during weekends, evenings or other out-of-school hours.
Round up to the nearest half hour.
As an example, three and a half hours would be recorded as 3.5.
Please record a 0 (zero) if you spent no time teaching in in the most recent week that you were working."

_____hours

Q1b. Y7hrs
Thinking about the same week, how many hours did you spend on **teaching Year 7 English**? Please only count actual teaching time.

Round up to the nearest half hour. As an example, three and a half hours would be recorded as 3.5 below.
Please record a 0 (zero) if you spent no time on Year 7 teaching in the most recent week that you were working.

_____ hours

Intro 2
The next set of questions are about the time you spend on different tasks involving Year 7 English **other than teaching.**

Q2a. Otheract
Thinking about the same week how many hours did you spend on the following activities for **year 7 English other than teaching** at your school?

Please exclude all time spent teaching Year 7. Please include tasks that took place during weekends, evenings or other out-of-school hours.
Again, round up to the nearest half hour. Please record a 0 (zero) if you spent no time on a listed activity.

1. Preparing Year 7 lesson plans (excluding developing resources)	_____ hours
2. Developing student facing lesson resources such as worksheets and presentations	_____ hours
3. Administering and marking Year 7 students' work	_____ hours
4. General non-curriculum related administrative work relating to Year 7 English (including paperwork or emails, registers, displays, communications with parents and other staff)	_____ hours
5. Co-planning and dialogue with colleagues within this school about Year 7 English teaching	_____ hours
6. Coordinating with people or organisations outside of school in relation to Year 7 programmes or interventions which aim to improve ability in English (including English Mastery)	_____ hours

Q2b.EMworkload

Compared to this time last year, do you think that your workload has increased, decreased or stayed the same?

Please consider any impact on planning, resource creation, meetings, and formative assessments but exclude time spent at English Mastery induction training.

1. Increased
2. Decreased
3. Stayed the same
4. Don't know
5. Not applicable - I didn't work at the school last year

Q2bi. Increase (if EMworkload=1)
And compared to this time last year, by how many hours has your workload increased in a typical working week?

Round up to the nearest half hour.

0.5…35

{Ask if surveyflag =1 and if EMworkload =2}

Q2bii. Decrease (if EMworkload=2)
And compared to this time last year, by how many hours has your workload decreased in a typical working week?

Round up to the nearest half hour.

0.5…35
