Statistical Analysis Plan
EAL in the Mainstream Classroom
Evaluator (institution): University of York
Principal investigator(s): Louise Tracey

Template last updated: March 2018

PROJECT TITLE	**EAL in the Mainstream Classroom**
DEVELOPER (INSTITUTION)	Challenge Partners/ Hounslow Language Service/ Lampton School
EVALUATOR (INSTITUTION)	University of York, UK
PRINCIPAL INVESTIGATOR(S)	Louise Tracey, University of York, UK
TRIAL (CHIEF) STATISTICIAN	Jan R. Böhnke, Dundee Centre for Health and Related Research, School of Nursing and Health Sciences, University of Dundee, Dundee, j.r.boehnke@dundee.ac.uk
SAP AUTHOR(S)	Jan R. Boehnke, Louise Tracey
TRIAL REGISTRATION NUMBER	ISRCTN15266150
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/eal-in-the-mainstream-classroom/

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	07/05/2019	*Original Version*

Table of contents

Introduction

'EAL in the Mainstream Classroom' is a CPD programme for teachers aiming to support EAL pupils in the mainstream classroom, with a particular focus on academic language. It is designed to enhance teachers' language skills and enable them to provide more focused classroom provision for EAL pupils, thus reducing the need for specialist teachers and support staff for this cohort. This is particularly important as schools cannot provide dedicated specialist support for EAL pupils who are not new arrivals. The CPD aims to improve teachers' skills with language, both general and subject specific. It provides training in how teachers can plan lessons with EAL pupils' language skills in mind, develop specific resources relating to those skills, and differentiate between pupils with different language skills and varying prior experience of education. The training supports classroom teachers' use and understanding of grammar, core academic vocabulary, and spoken language, which are key to helping EAL pupils within a whole class context, and which are also likely to have benefits for children more broadly.

The training was delivered through regional Delivery Centres located in schools especially selected and trained by Challenge Partners for this purpose. The Delivery Centres provided 3 days' training and support across two academic terms in a group setting to mainstream classroom teachers, who were invited to start implementing the learning with their designated classes after the first training day, so that they could reflect on, and learn from, their practice in feedback sessions during the remaining two training days.

The primary research question is:

- How effective is the 'EAL in the Mainstream Classroom' programme in improving subject specific academic attainment when delivered to Key Stage 4 EAL pupils taking GCSE Science?

The secondary research questions are:

- How effective is the 'EAL in the Mainstream Classroom' programme in improving subject specific academic attainment in a second GCSE subject (History)?
- How effective is the 'EAL in the Mainstream Classroom' programme in improving Academic attainment in English (as measured by GCSE English Language) when delivered to Key Stage 4 EAL pupils?
- What is the impact of 'EAL in the Mainstream Classroom' when pupils receive the approach from more than one teacher in more than one subject area (ie. when pupils are taught by trained 'EAL in the Mainstream Classroom' teachers in both Science and History GCSE subjects)?; and
- What is the impact of 'EAL in the Mainstream Classroom' on non-EAL pupils within the same classrooms?

The research is also assessing the impact of the programme on pupils with differing baseline fluency levels and on EAL pupils eligible for Free School Meals (FSM).

Design overview

Trial type and number of arms		**Two-arm, cluster randomised**
Unit of randomisation		School
Stratification variable		Delivery Hub (based on geographical location ie. schools grouped according to their nearest local Delivery Hub and randomised within that group)
Primary outcome	Variable	GSCE Science results (KS4) (EAL pupils only)
	measure (instrument, scale)	GCSE Science Results (scored 1-9; lowest-highest)
Secondary outcome(s)	variable(s)	GCSE History results (KS4) (EAL pupils only)
	measure(s) (instrument, scale)	GCSE History results (scored 1-9; lowest-highest)
Secondary outcome(s)	variable(s)	GCSE English Language results (KS4) (EAL pupils only)
	measure(s) (instrument, scale)	GCSE English Language results (scored 1-9; lowest-highest)

This is a two-armed school-level randomised efficacy trial. Randomisation between schools was preferred because the risk of diffusion was considered to be quite high since (a) the programme focuses on an approach to lesson planning and teaching with EAL pupils in general; and (b) the secondary research question relates to potential impact of the programme on a second subject-area, which means that children could be allocated to different conditions across subjects if a within-school design was adopted. The two trial arms therefore consist of schools implementing the programme (intervention group) and schools that continue with their teaching as usual (control schools).

The timeline for the study is provided in Table 1 below.

Table 1. Milestones and assessments scheduled for the EAL in the Mainstream Classroom trial

Date	Measure
May/ June 2014	KS2 reading assessments (from NPD) – baseline
June-December 2017	Teacher baseline survey (online collection by evaluation team)
September-December 2017	Pupil UPNs, EAL status and fluency data collected from schools
September-December 2017*	Randomisation
January 2018	CPD Workshop 1
March-April 2018	CPD Workshop 2
June 2018	CPD Workshop 3
Spring-Summer Term 2018	Programme delivery in intervention schools
June/July 2018	Teacher post-test survey
September 2018-July 2019	Ongoing delivery of programme

June/July 2019	Teacher follow-up survey
June/July 2019	Students sit GCSE examinations
October 2019	GCSE results accessed from NPD

* Randomisation conducted in batches as recruitment was on-going.

At recruited schools, participating pupils were those Year 10 EAL pupils enrolled in Science and History GCSE classes in September 2017, who were taught by the 'EAL in the mainstream classroom' trained teachers who agreed to participate in the trial and their non-EAL peers in the same classes. Pupils also needed to have been in the English education system at the end of Key Stage 2, as this forms the pre-test for the trial (see below). EAL pupils are those defined as EAL using the Department for Education's binary designation (Yes/ No). Since 2016/ 2017 additional data has been collected for EAL pupils in the school census relating to their fluency level (see detail below). This fluency measure was collected from schools for the participating pupils at the beginning of the academic year 2017/2018

Group allocation was conducted at the school-level using minimisation. Minimisation uses algorithms to ensure balance at baseline and permits ongoing allocation so schools know which condition they have been assigned to soon after recruitment. "Delivery Centre' (i.e. geographical region) was the only stratification variable used. This was in order to take into account the capacity of the Delivery Centres to deliver training and to ensure comparability within each Delivery Centre region. Randomisation was conducted by the Evaluation team using MinimPy software (Saghaei & Saghaei, 2011) between 29th September and 18th December 2017[1]. Seventy-one schools were randomised: 33 control and 38 intervention, with 730 pupils in the control condition and 851 in the intervention.

Changes to Design

As detailed in the protocol (version 2), the initial intention was to recruit 100 schools. However, due to lower than anticipated recruitment numbers (70 schools) the decision was taken to extend the trial to include a second cohort (of 50 new schools). At the same time, the assumptions used in the sample size calculations were reviewed resulting in an increase in the overall total number of schools to be included in the final sample (an increase from 100 to 120 schools). However, on 26 October 2018 the plan to recruit a second cohort had to be abandoned, since recruitment to the trial remained too low. It was agreed that the schools which were originally recruited (in the interim termed 'Cohort I') would be taken forward as the analysis sample. In the following, this smaller version of the study ("Cohort I-only") will be described and changes to the published protocol are flagged up where necessary.

Sample size calculations overview

Table 2 presents under "Protocol" the sample size calculation for the published protocol (version 2), which presented an agreed sample size of $N = 120$ schools, with an average of 14 Year 10 EAL pupils (about 16% of the population; Strand, 2015) in the primary subject specialism (GCSE in science) to potentially detect an effect size of $MDES = .22$ (significance

[1] Randomisation occurred in batches as follows: 29/9/17, 38 schools; 6/10/17, 6 schools; 15/11/17, 8 schools; 20/11/17, 6 schools; 29/11/17, 3 schools; 14/12/17, 8 schools; 18/12/17, 2 schools.

level $p < .05$, statistical power of 0.80, two-sided test; calculated with the formula presented in EEF, 2013). Additional assumptions for this *MDES* were:

- Intra-class correlation $\rho = 0.19$
- Since no data are available for the correlation between the pre-test (KS2 SATs) and the Science test, a conservative estimate of explained variance was agreed both at the **pupil** and **school** level ($r = .50$ or $R^2 = .25$).

Based on this proportion of EAL pupils, recruited schools would need to have at least 107 pupils in Year 10 to cover for approximately 20% student dropout during the study. Both are likely to be conservative assumptions, although currently no empirical estimates are available. At protocol stage it was noted that this was an optimistic estimate, since recruitment already showed that the sample sizes within schools varied quite strongly and the design would have no room to compensate for further drop-out.

At 26 October 2018 it was decided to end recruitment for this study due to the continuing recruitment problems and the sample size calculation for the *MDES* was revisited as well. The realised sample size for "Cohort I" leads to a minimal detectable effect size of *MDES* = .31.[2] Both the EEF and the evaluation team conducted further analyses taking the variation in school sizes into account and came to an agreement that the associated uncertainty in MDES was acceptable.

Since the study is focused on a specific sub-population (EAL pupils), performing further subgroup analyses to a high quality standard is difficult. Due to the small numbers in the primary population (EAL pupils) a subgroup analysis for FSM status can only provide broadly indicative results regarding the effect in this specific subgroup. To our knowledge no estimate of the share of year 10 pupils who fulfil both FSM and EAL criteria exists. We therefore expected conservatively about 20% of the EAL pupils per school to fulfil both criteria. For our 71 recruited schools with three FSM+EAL pupils each, the minimal detectable effect size for this sub-population is estimated at *MDES* = .38.

Table 2. Minimal Detectable Effect Sizes based on protocol (version 2) and trial after end of recruitment and discontinuation of cohort II efforts.

		Protocol		Randomisation	
		OVERALL	FSM[1]	OVERALL	FSM[1]
MDES		.22	.31	.31	.38
Pre-test/ post-test correlations	level 1 (pupil)	.50	.50	.50	.50
	level 2 (class)	--	--	--	--
	level 3 (school)	.50	.50	--	--

[2] Since the inception of the project, the EEF guidance on statistical analyses has changed. Originally it was envisaged to control for school-level variation in achievement in the analyses. This would have boosted the trial's statistical power, but would have made generalisation of the results more difficult (since they would have been conditional on between-school variation in prior achievement). When the changes to the size of the cohort were agreed, it was also agreed in line with the statistical guidance to control for between-school variation in the analysis. The minimal detectable effect size (MDES) with the original assumption of a between-school correlation of $r = .50$ and with the realised sample size in cohort I would have been 0.27.

Intra-cluster correlations (ICCs)	level 2 (class)	--	--	--	--
	level 3 (school)	.19	.19	.19	.19
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		two-sided	two-sided	two-sided	two-sided
Average cluster size		14	2.8	22.27	4.45
Number of schools	intervention	60	60	38[2]	38
	control	60	60	33[2]	33
	Total	120	120	71	71
Number of pupils	intervention	840	168	851	169[3]
	control	840	168	730	147[3]
	Total	1680	336	1581	316[3]

Notes:
[1]Extrapolated based on the assumption of 20% FSM in sample.
[2]As the randomisation was performed in batches, this resulted in an allocation that slightly favoured the intervention group. It was originally planned to recruit more schools which would have corrected this imbalance through minimisation.
[3]Numbers rounded since based on averages.

Outcome Measures

Primary Outcome

The primary outcome is the KS4 GCSE Science Score. Pupils are taking either Combined Science (Double Award) or three separate Science subject GCSE awards (Triple Award) in Summer 2019. These are scored across the range from 1-9 (1 being the lowest awarded score, and 9 the highest). Two scores will be provided for Double Science and three for Triple Science assessments. For the analysis of all questions regarding the primary outcome "Science GCSE", we will use the average of the scores provided, i.e. over two scores for those pupils taking Double Science and over three scores for those pupils taking Triple Science. While we agree with the EEF's caution regarding the use of composite scores in analyses, we argue that, in this case, the use of this composite is justifiable.[3] None of the research questions refer to specific sub-scores of the science exam. In addition, given that the programme aims to embed academic language across the different science subjects, and that there is a difficulty in separating the sciences given the current assessment system, we suggest that an intervention effect should be detectable across all (two/three) subjects. The pupils' test results will be available via the NPD at the ONS.

[3] Responses received from NPD on the availability of the data have so far been not entirely clear. It could be that the average is the only data that is returned.

Secondary Outcomes

The secondary outcomes are:

- the KS4 GCSE History Score for those pupils who are taking History GCSE. The test is scored from 1-9 .
- The KS4 GCSE English Language Score for all pupils (this being a compulsory subject at the end of KS4). The test is scored from 1-9 (as above).

These two outcomes were chosen to assess the potential impact on a second intervention subject (i.e. GCSE History) and in another subject where the intervention was not delivered, but because of the focus on academic language would be expected to have an impact (GCSE English Language). The pupils' GCSE test results are available via the NPD at the ONS.[4]

Other measures

Additional demographic data will be downloaded from the NPD such as Gender, EAL status, and FSM (EVERFSM_6_P). EAL status is needed for inclusion in the analysis; FSM is needed for the subgroup analysis; and Gender will be used for missing data imputation if needed.

The analysis will use KS2 SATs Reading outcomes (KS2_READMRK; raw scores) as a pre-test since these assessments are high in contextual validity and have previously been highly correlated with attainment at the end Key Stage 4. The data are available from the NPD.

While the NPD provides a classification of pupils' EAL status, it does not provide a gradated assessment of students' verbal fluency. Therefore, a measure of fluency (using the EAL fluency descriptors from the school census) was at baseline for subgroup analysis of differential effects of the intervention based on students' pre-intervention fluency levels. From September 2016 all schools are required to return this information, recording level of fluency (graded from A-E, A being new to English and E being fluent), annually for all EAL pupils. This data will be used for a secondary analysis on whether there is a differential impact of the intervention on students with different verbal fluency levels at the start of year 10. As the assessment is performed by the schools themselves, we received this data directly from all participating schools. Data was transferred via encrypted, password protected spreadsheets directly from schools in line with Data Protection regulations and Data Sharing Agreements made between schools and the University of York.

Analysis

All analyses will be conducted by Jan R. Boehnke (University of Dundee). All data will be presented descriptively with means, standard deviations, and medians for quantitative outcomes and category frequencies for categorical data. All statistical analyses will be reported for complete cases as well as corrected for missing data and drop-outs (which we expect to be low in relative frequency for all NPD data we use). Bootstrapped confidence intervals are used to judge the statistical significance of the intervention effect. Bootstrapped confidence intervals take into account violations of normality assumptions in model error distributions and since they are recommended to evaluate the variation in mediation and

[4] At the moment the secondary outcome KS4 GCSE History is to be collected at all 71 schools for 606 pupils in the control and for 633 in the intervention group.

cross-level interaction terms in our analyses (e.g. Pituch et al., 2008), we apply them consistently throughout.

Primary outcome analysis

The primary research question of this trial is: How effective is the 'EAL in the Mainstream Classroom' programme in improving subject specific academic attainment when delivered to Key Stage 4 EAL pupils taking GCSE Science? For this analysis only EAL pupils will be included and the impact evaluation will use a mixed effects model in which students are nested within schools. This makes it possible to separate within-school variation in the outcome from between-school variation.

The analysis will be intent-to-treat, which means that schools will be treated according to the condition they were allocated (control, intervention), not that which they actually received. The analyses will use cluster-bootstrapped confidence intervals (e.g., Huang, 2018) to account for potential violations of distributional assumptions: From each school a random sample of the same size as its actual sample is drawn (with replacement) and across these school-wise bootstrap samples, the mixed model is then estimated.[5] This process is repeated *b* = 1000 times and for a 95%-confidence interval the statistical estimates are saved and their top and bottom 2.5%-quantiles are identified. The average of the bootstrapped values will be treated as the point estimate and will be reported in all coefficient tables. No *p*-values will be reported.

This study was planned for a single primary outcome, the KS4 Science GCSE. In accordance with the power analysis, pre-test data from the Key Stage 2 (KS2) SATs Reading raw scores (KS2_READMRK, possible range 0-120) will be used as pupil-level covariate without random variation across schools. An individual student i's KS4SCI result in a specific school will be modelled as depending on school j's random school-level intercept (μ_0j) and a pupil-level error term (ε_ij). Each school's random intercept (μ_0j) will depend on an overall intercept (average performance; γ_00); each school's level on the stratification variable which controls for the Delivery Hub (DEL); the intervention to which the school was randomised (CP, testing for the intervention effect) and contain a school-level error term (u_00):

$$KS4SCI_{ij} = \mu_{0j} + \mu_{1j} KS2_{ij} + \varepsilon_{ij} \quad (1)$$

$$\mu_{0j} = \gamma_{00} + \gamma_{01} DEL_{0j} + \gamma_{02} CP_{0j} + u_{00} \quad (2)$$

$$\mu_{1j} = \gamma_{10} \quad (3)$$

$$\mu_{2j} = \gamma_{20} \quad (4)$$

The analysis will be performed in the R environment (R Core Team, 2018; most up to date version at point of analysis); specifically the R-package `lme4` (Bates, Mächler, Bolker, & Walker, 2015) will be used with the corresponding formula expression in the command `lmer()`:

```
KS4SCI ~ KS2 + DEL + CP + (1 | school)
```

The intervention is evaluated as potentially effective in this trial when the average bootstrapped point estimate for the coefficient of the intervention effect (γ_{02}) reflects that

[5] E.g. if there were observations 1,2,3,4,5 in a school, one resample could be [1,2,2,5,4] and another [1,5,1,1,3].

intervention schools achieve on average higher scores on *KS4SCI* and the 95%-bootstrap confidence interval of this coefficient does not include 0.

Secondary outcome analysis

The secondary outcomes are: How effective is the 'EAL in the Mainstream Classroom' programme in improving subject specific academic attainment in a second GCSE subject (History), and in GCSE English Language when delivered to Key Stage 4 EAL pupils? The analytic approach will use exactly the same procedure and model as for the primary outcome, with two differences:

- instead of KS4Sci the secondary outcome variables will be used as dependent variables (GCSE History Score; GCSE English Language Score; see above); and
- the pupil-level control variable for the type of science class (double/ triple; see above) will not be used in this analysis since it is not relevant to control the dependent variables' scores for this difference.

As before, the intervention will be evaluated as having shown a potential effect on a secondary outcome when the 95%-bootstrap confidence interval of the coefficient (γ_02; see formula 2 above) does not include 0. This result cannot be used to gauge the efficacy of the intervention and is reported purely for exploratory purposes to evaluate whether there are potential positive or negative effects of the intervention on curriculum outcomes which would need further research.

Subgroup analyses

Two subgroup analyses will be conducted that were identified by the EEF and project partners: the impact of the programme on pupils eligible for Free School Meals; and the impact of the programme on EAL pupils of different fluency levels at the start of the programme. The trial was not powered for either subgroup analysis, therefore these analyses only provide descriptive insights into potential subgroup effects.

The research will assess the impact of the programme on pupils eligible for Free School Meals (EverFSM). First, the results for the primary outcome will be presented descriptively for the EAL+FSM pupils only. Second, to evaluate whether there is a differential effect for FSM students, the mixed effects model described for the primary outcome will be extended by adding FSM and an interaction term between FSM and the intervention variable (CP) and run on all EAL students. The intervention will be evaluated as showing a subgroup effect for FSM when the bootstrapped 95%-confidence interval for the coefficient for the interaction term does not include 0. As before, this analysis is purely exploratory and does not estimate the efficacy of the intervention itself.

The previous analytic strategy is extended to include a pupil-level covariate for FSM which has a random effect across schools and this variation is predicted by the intervention variable:

$$\boldsymbol{KS4SCI_{ij} = \mu_{0j} + \mu_{1j} KS2_{ij} + \mu_{2j} FSM + \varepsilon_{ij}}$$

$$\boldsymbol{\mu_{0j} = \gamma_{00} + \gamma_{01} DEL_{0j} + \gamma_{02} CP_{0j} + u_{00}}$$

$$\boldsymbol{\mu_{1j} = \gamma_{10}}$$

$$\mu_{2j} = \gamma_{20} + \gamma_{21} CP_{0j} + u_{20}$$

The analysis will be performed in the R environment with the corresponding formula expression in the command lmer():

```
KS4SCI ~ KS2 + FSM + DEL + CP + CP:FSM + (1 + FSM | school)
```

The intervention will be evaluated as having shown a potential interaction with the specified subgroup variable when the 95%-bootstrap confidence interval of (γ_21) does not include 0.

The approach for testing the effect for different baseline fluency levels is similar. Again, only EAL pupil data will be analysed. The fluency measure will be entered effect-coded into the regression equation. Since the programme is targeted at the mid-fluency levels (ie. not A's and E's), fluency level B will be used as a reference category. The four dummy variables entered in the equation therefore allow assessing effects due to the baseline fluency levels (fluA to fluE in the R representation below) as well as whether there were specific effects of the programme depending on the baseline fluency scores (cross-level interaction terms CP:fluA to CP:fluE in the R representation below). Allowing the strength of the fluency coefficients to vary across schools (i.e. assuming random effects) will be explored but, due to the high number of categories and comparatively small number of schools and respondents, these models are not likely to converge or be based on sufficient coverage (i.e. some of the estimated coefficients would be based only on data from some of the schools). If at least 80% of schools in both intervention arms report EAL students in all five categories, the coverage will be assumed as sufficient and the results from random effects models will be reported. Otherwise a fixed effect analysis will be conducted.

The corresponding formula expression for the random effects model in the command lmer() in R is:

```
KS4SCI ~ KS2 + fluA +fluC + fluD + fluE + DEL + CP +
      CP:fluA + CP:fluC + CP:fluD + CP:fluE +
      (1 + fluA +fluC + fluD + fluE | school)
```

If the confidence interval of a coefficient does not include zero it will be evaluated as having a potential (differential) effect. In that case the interaction will also be visually displayed to help interpretation.

No other subgroup analyses will be performed since none were defined in the protocol.

Additional analyses

Three additional analyses will be performed.

First, the impact of 'EAL in the Mainstream Classroom' when pupils receive the approach from more than one teacher in more than one subject area (i.e. when pupils are taught by trained 'EAL in the Mainstream Classroom' teachers in both Science and History GCSE subjects) will be analysed. The analytic approach for the primary and secondary outcomes relating to the impact of the programme on Science, History and English Language GCSEs will be extended by adding a pupil-level predictor to the equation that captures how much pupils were exposed

to the programme. For this we will code, at pupil-level, how many teachers of their classes attended the training. We will re-run the same model as for the secondary outcome analysis with GCSE KS4 English Language as the dependent variable and add this exposure score as a pupil-level covariate. Adding this coefficient allows us to test whether being taught by multiple teachers trained in the programme has an additional effect over and above programme provision in itself. If the bootstrapped confidence interval of the covariate's coefficient does not include zero it will be evaluated as having a potential additional effect and the direction of the effect will be interpreted.

Second, we will test whether an EAL student's choice regarding double/triple science introduces heterogeneity in the programme's effect due to possibly different profiles of students taking these options. For this we will code on pupil-level, whether they chose double or triple science. We will re-run the same model as for the primary outcome analysis and the added coefficient for double/triple science allowing us to test whether the choice of double or triple science has an additional effect over and above the programme impact by itself. If the bootstrapped confidence interval of the covariate's coefficient does not include zero it will be evaluated as having a potential additional effect and the direction of the effect will be interpreted.

Third, we will test the impact of the 'EAL in the Mainstream Classroom' programme on non-EAL pupils within the same classrooms. We will (a) use all available student data and (b) add a cross-level interaction effect between EAL/non-EAL and the programme. This coefficient assesses both whether there is an overall effect of the intervention and whether this effect is differential between the two pupil groups (e.g., higher averages in EAL population). If the confidence interval of the coefficient does not include zero it will be evaluated as having a potential differential effect. In that case the interaction will also be visually displayed to help interpretation.

Imbalance at baseline

KS2 results will be presented with means, standard deviations, and medians; the control and intervention group distribution will be displayed via density plots; imbalance will be evaluated in standardised mean differences for KS2. Category frequencies for gender, EverFSM and Verbal Fluency will be presented and imbalance will be evaluated via effect size (Faul et al., 2007) for cross-tabulations of intervention group. If imbalance is detected for any of the three categorical variables, a robustness analysis will be run for primary and secondary outcome analyses, including variables showing imbalance as student-level covariates (KS2 is included as a student-level covariate in all analyses as per power analysis).

Missing data

The amount of missing data will be documented for each variable individually as well as for the patterns of missing values which occur. Further, the relative frequency of students with any missing data will also be presented by school and by treatment arm. To evaluate the impact of missing data on the robustness of findings from the ITT analyses of the primary outcome, sensitivity analyses will be run to evaluate the robustness of the results if either > 5%

missing data for the primary outcome analysis are encountered (i.e. 5% of cases would have to be deleted listwise for that analysis). We will employ a fully conditional specification approach (FCS) (Enders, Keller & Levy, 2018) to impute missing values in the sensitivity analysis.

The proportion of missing observations for each imputed variable will be reported based on the full sample as well as percentages of students with any observed missings per school. To identify potential patterning of missing data across variables, we will explore these patterns visually via pattern plots; and based on the final imputation model, estimates for potential relationships between observed variables and missingness at pre- and post-assessments will be documented (see below).

The stand-alone software *Blimp* (Enders, Keller & Levy, 2018) will be used for this analytic step and the following variables will be entered into the algorithm:

- Gender, EverFSM, Verbal Fluency, and the KS2 result ("baseline data"; independent of whether they have missing data or not);
- The primary and secondary outcome variables ("follow-up"; which are likely to have missing data); our preference would be to use all scores individually in the imputation model and average after the imputation (i.e. using "passive imputation"), but since responses received from NPD on the availability of the data have so far been not entirely clear (see FN3), this aspect cannot be finalised at this stage.
- Additionally two dummy variables which code whether baseline data is missing (yes/ no) or only follow-up data (yes/ no; see below).

The chosen approach allows for multilevel imputation with random effects for level-1 variables (KS2, Gender, EverFSM), which will be tried in the first instance since it has been shown to be the preferred approach (Enders, Keller & Levy, 2018). But since this can lead to convergence problems even in well-structured data sets (many cases, few variables – such as in our case), we will potentially reduce this to random intercept models if a reduction in complexity was needed.

Interval-scaled variables will be modelled with linear regressions; ordinal variables with ordinal regression models; and dichotomous variables with logistic regressions. FCS allows us to simultaneously optimise the imputation of a multivariate data set with missing data on a number of variables. The algorithm will be set to run with Blimp's Gibb's sampler with four chains, for 5000 burn-in iterations and taking five imputed data sets every 2000 iterations ("thinning"; i.e. 4x5=20 imputed data sets). These settings are more conservative than standard settings to accommodate the size and complexity of the imputation model. The convergence of the imputation will be tested via inspection of the potential scale reduction factors (PSR; Gelman & Rubin, 1992). Blimp calculates PSR for fixed effects, between- and within-cluster residual variance parameters, and the thresholds for ordinal variables (more than two categories) and the convergence of all these will be evaluated and described for the imputed data sets.

To link imputations and bootstraps, the 20 imputed data sets will be read into R and draw 50 bootstrap samples from each (again resulting in $b = 1000$ bootstrap samples), which also allows for the assessment of whether, and how much, impact specific imputation runs have on the results.

At the moment, generalised versions of FCS and joint modelling are the two predominantly used approaches to missing data imputation, especially in the case of multilevel data. Both approaches have the advantage that neither defines a specific model for the missingness mechanism. In cases such as this, with very few variables and virtually no information about the specific assessment context, both allow researchers to maximise available data. The methodological literature dealing with these approaches suggests that they both work well for a variety of cases similar to the clustered trial analyses discussed in this study and that there is currently no specific reason to prefer one approach over the other (e.g., Mistler & Enders, 2017; Enders, Hayes & Du, 2019). We are employing an FCS approach since we are expecting missing data on all variables used in the analysis, which can be dealt with slightly more flexibly in this approach, which is based on the definition of individual equations. It further builds only on very basic tenets of the missing-at-random assumption, i.e. that conditional on observed variables, data are missing at random. Even in the presence of data that are not missing at random, taking the observed relationships into account is preferable to an analysis based on listwise deletion only (e.g., van Ginkel et al., in press, p. 5). To approximate the most basic of missingness processes we included two dummies which will condition predictions of the MICE procedure on whether any data for a respondent is missing at baseline or whether any data is missing at follow-up (the prediction equations for these two variables will be reported based on the imputed results to explore potential processes leading to missing data). The advantages and limitations of this approach are discussed in detail in van Ginkel et al. (in press).

Compliance

To assess compliance, we will use attendance at training as a proxy measure. Given that there are three training sessions this will be on a scale of "0" (no training attended, 0/3 sessions) to "1" (all three sessions attended, 3/3 sessions). To analyse the effect of compliance on the primary outcome of EAL pupils, we will use the approach outlined by Steele et al. (2007). In this approach, two multilevel models are estimated; one is the same as the primary outcome analysis (eg. 1-4), except that the intervention variable is replaced by the compliance measure; in the other, the compliance measure is predicted by the intervention group and additional variables (see below). The random effects across these equations are allowed to correlate in the simultaneous estimation procedure to capture characteristics that influence either dependent variable:

- Equation system 1 will be the same as equations 1-4 with the respective compliance variable replacing the intervention variable (*CP*);
- Equation system 2 will have the respective compliance variable as the dependent variable (treating it as a student-level variable since both may vary within one school), predicted by the intervention (*CP*) and other variables specified in eq. 1-4 above; additionally the schools' averages of pre-treatment KS2, FSM and Fluency (taking all students into account) will be used as predictors, to proxy potential selection effects due to better performing schools and/or pupils being drawn from higher socioeconomic status backgrounds.

The intention is to treat the compliance variable as categorical. The analysis will be conducted using MLwiN (Charlton et al., 2019) and R2MLwiN (Zhang et al., 2016). The procedure results in an estimate of the relationship between the compliance variable and the outcome (*KS4SCI*), corrected for potential selection effects as captured by the second equation. The results of both equations and random effects will be reported and the results discussed. We will also

evaluate how plausible it is that one or more variables serve as an 'instrument' (i.e. predicting compliance, but not science outcome; Steele et al., 2007). If the bootstrap confidence interval for a compliance variable does not include "0", the result will be interpreted as indication for an effect that is due to dose rather than ITT group assignment.

Intra-cluster correlations (ICCs)

ICCs will be calculated at school level via variance components from a mixed model without any predictors (see primary outcome analysis). Where necessary these computations will be conducted via a generalised version of the model (e.g., gender). Confidence intervals will be calculated via clustered bootstrap (see section on analysis).

Effect size calculation

Effect sizes will be calculated based on the total variance in the models. For descriptive differences Hedge's *g* will be used as suggested by the EEF guidance. For the results from our mixed model analyses, the estimated intervention effect as represented by the intervention's coefficient in a model (Effect) will be divided by the total variance in the model (see definition of error terms above):

$$ES = \frac{Effect}{\sqrt{u_{00}^2 + \varepsilon_{ij}^2}}$$

Confidence intervals will be bootstrapped. Here, **Effect** is the coefficient from the estimated model (e.g., γ_02 in the analysis of the primary outcome; formula 2).

References

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1), 40–49. https://doi.org/10.1002/mpr.329

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software; Vol 1, Issue 1 (2015). https://doi.org/10.18637/jss.v067.i01

Charlton, C., Rasbash, J., Browne, W.J., Healy, M., & Cameron, B. (2019) MLwiN Version 3.03. Centre for Multilevel Modelling, University of Bristol.

Education Endowment Foundation, 'Pre-testing in EEF evaluations'. Dated 31.10.2013

Enders, C.K., Hayes, T., & Du, H. (2019). A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint

Model Imputation with Random Covariance Matrices. Multivariate behavioral Research, 53, 695-713.

Enders, C.K., Keller, B.T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. Psychological Methods, 23, 298-317.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research methods, 39, 175-191.

Gelman, A., & Rubin, D.B. (1992). Inferences from iterative simulation using multiple sequences. Statistical Science, 7, 457-472.

Heymans, M.W., van Buuren, S., Knol, D.L., van Mechelen, W., & de Vet, H.C.W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Medical Research Methodology, 7:33.

Huang, F. L. (2018). Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. Educational and Psychological Measurement, 78(2), 297–318. https://doi.org/10.1177/0013164416678980

Kasim, A., Xiao, Z. M., Higgins, S., & De Troyer, E. (2017). eefAnalytics: Analysing Education Trials (Version 1.0.6). Last access 08.12.2018 from https://cran.r-project.org/web/packages/eefAnalytics/index.html

Mistler, S.A., & Enders, C.K. (2017). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. Journal of Educational and Behavioral Statistics, 42, 432-466.

Pituch, K.A., Stapleton, L.M., & Kang, J.Y. (2006) A Comparison of Single Sample and Bootstrap Methods to Assess Mediation in Cluster Randomized Trials, *Multivariate Behavioral Research*, *41*, 367-400

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.1). Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Saghaei, M., & Saghaei, S. (2011). Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials. Journal of Biomedical Science and Engineering, 4, 734–739.

Schomaker, M., & Heumann, C. (2007). Model selection and model averaging after multiple imputation. Computational Statistics & Data Analysis, 71, 758-770.

Steele, F., Vignoles, A. & Jenkins, A. (2007). The effect of school resources on pupil attainment: A multilevel simultaneously equation modelling approach. Journal of the Royal Statistical Society, Series A, 170, 801-824.

van Ginkel, J.R., Linting, M., Rippe, R.C.A., & van der Voort, A. (in press): Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, DOI: 10.1080/00223891.2018.1530680

Zhang, Z., Parker, R.M.A., Charlton, C.M.J., Leckie, G., & Browne, W.J. (2016). R2MLwiN: A Package to Run MLwiN from within R. Journal of Statistical Software, 72, 1-43.