Statistical Analysis Plan
First Thing Music

Evaluator (institution): BIT and UCL Institute of Education

Principal investigator(s): Pantelis Solomon

Template last updated: March 2018

PROJECT TITLE	**Using the First Thing Music programme to improve pupil attainment in schools**
DEVELOPER (INSTITUTION)	Lindsay Ibbotson and Tees Valley Music Service
EVALUATOR (INSTITUTION)	Behavioural Insights Team & UCL Institute of Education
PRINCIPAL INVESTIGATOR(S)	Pantelis Solomon
TRIAL (CHIEF) STATISTICIAN	Pantelis Solomon
SAP AUTHOR(S)	Kim Bohling, Florentyna Farghly QA by Jake Anders and Nikki Shure
TRIAL REGISTRATION NUMBER	ISRCTN14035536
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/First_Thing_Music_protocol.pdf

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	11 January 2019	

Table of contents

Introduction

The First Thing Music (FTM) programme aims to improve children's reading and social skills by providing them with daily music sessions. Students will learn the basics of music through daily singing and musical games with teachers who will be trained by music practitioners. The model that will be tested in this programme is comprised of daily 15-minute music sessions for Year 1 pupils (5-6 year olds) over the course of three terms.

The FTM evaluation is part of a broader programme of work entitled 'Learning about Culture', which aims to improve the evidence base around arts-based education programmes. This is coordinated by the Education Endowment Foundation (EEF) and the Royal Society for the Encouragement of Arts, Manufactures and Commerce (RSA).[1] It consists of five programmes: two in Key Stage 1 and three in Key Stage 2.

The evaluation is designed as a two-armed classroom-level cluster randomised controlled trial involving 65 primary schools with a total of 123 classes. 61 classes were randomly assigned to receive the intervention and 62 classes were randomly assigned to be in the control group and not receive any intervention. Recruitment occurred in Winter/Spring 2017/18 with the aim of starting the intervention with the cohort of pupils starting Year 1 in September 2018. The primary outcome measure of the evaluation will be impact of the programme on reading attainment, measured by the Progress in Reading Assessment (PIRA) by Rising Stars.[1] Secondary outcomes will measure the programme's effect on social skills, as measured by the social skills sub-scale of the Social Skills Improvement System (SSiS)[2]; and on creative self-efficacy, as measured by the ideation sub-measure of the writing self-efficacy measure.[3]

Design overview

Trial type and number of arms		**Two-arm, clustered randomised**
Unit of randomisation		Classroom
Stratification variables (if applicable)		School
Primary outcome	variable	Reading attainment
	measure (instrument, scale)	PIRA, score range 0-25
Secondary outcome(s)	variable(s)	(1) Social skills (2) Creative self-efficacy
	measure(s) (instrument, scale)	(1) Social Skills Improvement System (SSiS) -- social skills sub-measure, 46 items each scored 0-3, total raw score range 0-138

[1] https://www.thersa.org/globalassets/pdfs/reports/rsa-learning-about-culture-report.pdf

[2] https://www.pearsonclinical.com/education/products/100000322/social-skills-improvement-system-ssis-rating-scales.html

[3] Bruning, R., Dempsey, M., Kauffman, D., McKim, C. & Zumbrunn, S. (2013) Examining Dimensions of Self-Efficacy for Writing. Journal of Educational Psychology, 105(1), 25-38

		(2) Writing Self-Efficacy measure -- ideation sub-measure (3 questions), 3-point Likert scale; score range 3-9

Follow-up

The original recruitment target of 120 classes was exceeded as First Thing Music approached additional schools.

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	**FSM**	**OVERALL**	**FSM**
MDES		0.20	0.23	0.191	0.23
Pre-test/ post-test correlations	level 1 (pupil)	0.61	0.61	0.61	0.61
	level 2 (class)	0.61	0.61	0.61	0.61
	level 3 (school)	N/A	N/A	N/A	N/A
Intracluster correlations (ICCs)	level 2 (class)	0.19	0.19	0.19	0.19
	level 3 (school)	N/A	N/A	N/A	N/A
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 [4]		22	5	20	5
Number of classes	intervention	56	56	61	61
	control	56	56	62	62
	total	112	112	123	123
Number of pupils	intervention	1534	384	1509	343
	control	1534	384	1549	338
	total	3068	768	3058	681

Number of forms	**Schools**	**Intervention**	**Control**
1	19	10	9
2	34	34	34
3	12	19	17

Assumptions prior to randomisation

- **The intra-cluster correlation (ICC) is estimated to be 0.19.** As randomisation was conducted at the classroom level, we are only accounting for class-level ICC. Estimating ICC values for class-level randomisations is difficult as there is less guidance available relative to school-level randomisation. Other EEF trials that used class-level randomisation have found the estimated ICC values when performing

[4] Cluster sizes take into account an estimated attrition rate of 20% and have been rounded to nearest integer.

sample size calculations were overly optimistic.[5,6] At the time of writing the trial protocol, guidance provided by the EEF suggested that a school-level ICC value for a reading outcome measure in KS1 would be estimated at 0.11 for schools in the North East.[7] We adjusted this upwards to 0.19 to provide a margin of error commensurate with the experiences of prior EEF studies.[8]

- **There is an average of 27.4 pupils per class using ONS statistics from 2016.[9]**
- **20 per cent of children in each school will opt-out or be unable to participate in the collection of an endline outcome measure (attrition due to changing school, inability to complete assessment etc.).** This estimate is based on the 15% standard post-randomisation attrition rate in EEF studies[10], plus an additional allowance for children being opted-out of the study (5%).
- **Test-retest correlation of 0.61.** As we will use Early Years Foundation Stage Profile (EYFSP) scores as a baseline when analysing our primary outcome measure, the predictive power of this baseline will also factor into our sample size calculations. We estimate this value using unpublished analysis from the Fisher Family Trust (FFT) of the test-retest correlation coefficient of Early Years Foundation Stage Profile (EYFSP) score and PIRA assessments collected at the end of year 1 for a prior EEF trial (ABRA: Online Reading Support).[11]

Analysis

The analysis plan is described in the sections that follow. All analyses will be carried out using the statistical software Stata[12] (see Appendix 1 for the prospective Stata syntax). The estimated impacts from all primary and secondary analyses will be 'intention to treat' (ITT) effects and will be reported with 95% confidence intervals.

Primary outcome analysis

Outcome

[5] Foreign Language Learning in Primary Schools, a trial testing an intervention on English literacy involving Year 3 and 4 children, estimated an ICC of 0.05 when performing power calculations, but found it to be 0.13 when post-hoc analysis was performed (see pg 23). https://educationendowmentfoundation.org.uk/public/files/EEF_Project_Report_FLL.pdf

[6] Grammar for Writing, a trial testing an intervention on writing involving Year 6 children, estimated an ICC of 0.19 when performing power calculations, but found it to be 0.32 when post-hoc analysis was performed (see pg 26). https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Campaigns/Evaluation_Reports/EEF_Project_Report_GrammarForWriting.pdf

[7] Document previously found at: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf

[8] This was estimated by inflating the school level ICC value expected by 70%, as per change between expected and observed ICC values in the Grammar For Writing trial.

[9] www.gov.uk/government/uploads/system/uploads/attachment_data/file/552342/SFR20_2016_Main_Text.pdf

[10] Based on the EEF allowing projects to recruit 15% extra schools to account for likely attrition. See: Preventing Attrition: Pack for projects (date unknown). Retrieved from https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Attrition_pack.pdf

[11] https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ABRA.pdf

[12] The precise version used will be out of our control as this analysis will be conducted on the ONS Secure Research Service. We will use the most recent version available.

The primary outcome measure will be the Progress in Reading Assessment (PIRA) by Rising Stars.[13] PIRA is a standardised assessment of pupils' reading attainment and profile of reading skills. It measures reading ability in the following areas: phonics, literal comprehension, and reading for meaning. This is a standardised and well-known test, which has been used in a number of prior EEF evaluations.[14, 15] Endline PIRA assessments will be conducted during May and June 2019 by trained research assistants (RAs) who will be blind to trial arm assignment. Rising Stars, the publisher of PIRA, will mark the assessments.

Our baseline covariate will be the child's EYFSP aggregate score for four learning goals:

1) understanding (FSP_COM_G02);
2) speaking (FSP_COM_G03);
3) reading (FSP_LIT_G09); and
4) writing (FSP_LIT_G10).

These goals were selected as they are most closely linked to reading, our primary outcome measure. Past research found that neither the total EYFSP score nor the score for personal, social and emotional development correlated well with later attainment, but the scores for Communication Language and Literacy do correlate strongly with later attainment.[16] For each goal, teachers judge whether the pupil is meeting, exceeding, or not yet meeting the expected level of development at the end of the EYFS. Each grade will be assigned point scores as follows:

- Not yet meeting expectation (emerging) – 1 point
- Meeting expectation (expected) – 2 points
- Exceeding expectation (exceeding) – 3 points
- Not assessed (A) – coded as "missing"[17]

The aggregate score will range from 4 to 12.

With this approach to aggregating the scores, we acknowledge that we are making an assumption that the distance between meeting and not meeting expectations is similar in both directions on multiple learning goals. However, given that more granular baseline data is not available, we think this is the best way to utilise this data as a baseline measurement, as it provides an indication as to whether the pupil is generally at, above, or below expectations on the range of learning goals most closely associated with our outcome

[13] https://www.risingstars-uk.com/Series/Rising-Stars-Pira-Tests

[14] McNally, S. (2016). *Evaluation Protocol: An Evaluation of Teaching Assistant-Based Small Group Support for Literacy*. London, United Kingdom: Education Endowment Foundation. Retrieved from https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Digital_-_Small_Group_Support_for_Literacy.pdf.

[15] McNally, S., Ruiz-Valenzuela, J., & Rolfe, H. (2016). *ABRA: Online Reading Support.* London, United Kingdom: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ABRA.pdf

[16] Snowling, M. J., Hulme, C., Bailey, A. M., Stothard, S. E., & Lindsay, G. (2011). Better communication research project: language and literacy attainment of pupils during early years and through KS2: does teacher assessment at five provide a valid measure of children's current and future educational attainments?. London: Department for Education.

[17] According to the EYFS Assessment and Reporting guidelines, a child is not assessed due to one of the following: long periods of absence (e.g. prolonged illness), attendance of provision for an insufficient amount of time for the teacher to make an adequate assessment, an exemption. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748449/2019_early_years_foundation_stage_assessment_and_reporting_arrangements.pdf

measure.

Analysis

Primary analysis will be intention-to-treat (ITT), in which we test the hypothesis that participating in the programme has an effect on reading attainment. Analysis will use raw PIRA total scores (0-25) and will be carried out using an ordinary least squares (OLS) linear model:

$$Y_i = \alpha + \beta Treat_i + \gamma X_i + \theta School_i + \epsilon_i$$

where:

- Y_i is the raw PIRA score for student *i*;
- $Treat_i$ is a binary indicator for the treatment assignment (1 if the class is assigned to treatment; 0 if not)
- X_i is a vector of baseline attainment measured through aggregated EYFSP learning goal scores for student *i*
- $School_i$ is a vector of school-level fixed effects
- ϵ_i is the error term clustered at the class level (standard errors are corrected for clustering)

Given the assumptions about the baseline measure, we will conduct exploratory analysis using a more flexible specification of the model the same model as above (for example, include a quadratic term for baseline attainment) in order to assess whether the relationship between EYFSP and PIRA scores is non-linear.

Secondary outcome analysis

The secondary analysis will measure the impact of the intervention on the pupils' social skills and creative self-efficacy.

Social skills outcome

Social skills will be assessed at endline using the Social Skills sub-scale of the Social Skills Improvement System (SSiS).[18] The SSiS Social Skills scale assesses pupils' skills across the following sub-scales: communication, cooperation, assertion, responsibility, empathy, engagement and self-control.

SSiS is standardised and has been used in prior EEF evaluations.[19] We chose to use SSiS, over an equally popular instrument, the Strengths and Difficulties Questionnaire (SDQ) because it is more thorough and in-depth than SDQ. The questionnaires will be delivered to teachers electronically. As with all measures of social skills at this age, this must be completed by the child's teacher and thus cannot be blind to trial arm assignment.

The sub-scale contains 46 items on which teachers rate the frequency with which they observe the pupil demonstrating the behaviour; the frequency rating is then translated into point scores (Never=0, Seldom=1, Often=2, Always=3). Aggregate scores will range from 0-138.

In the analysis, we will use a baseline covariate consisting of EYFSP scores aggregated

[18] https://www.pearsonclinical.com/education/products/100000322/social-skills-improvement-system-ssis-rating-scales.html

[19] Centre for Effective Education, Queen's University Belfast. (2016). *Evaluation Protocol: Zippy's Friends.* London, United Kingdom: Education Endowment Foundation. Retrieved from: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_Character_Zippys_Friends_protocol.pdf.

across the following learning goals:

1) self-confidence and awareness (FSP_PSE_G06);
2) managing feelings and behaviour (FSP_PSE_G07); and
3) making relationships (FSP_PSE_G08).

The aggregate score will range from 0-9. As previously stated, we believe that aggregating the measures is the best way to utilise this data as a baseline measurement to generally indicate whether the pupil is at, above, or below expectations on the learning goals most closely associated with the outcome measure.

Social skills analysis
Analysis will follow the model specified for primary analysis, substituting the appropriate secondary outcome measure and baseline measure.

Secondary analysis will be ITT, in which we test the hypothesis that participating in the programme has an effect on student social skills. Analysis will use raw SSiS social skills sub-scale scores (0-138) and will be carried out using an ordinary least squares (OLS) linear model:

$$Y_i = \alpha + \beta Treat_i + \gamma X_i + \theta School_i + \epsilon_i$$

where:

- Y_i is the raw SSiS social skills sub-scale score for student *i*
- $Treat_i$ is a binary indicator for the treatment assignment (1 if the class is assigned to treatment; 0 if not)
- X_i is a vector of baseline attainment measured through aggregated EYFSP learning goal scores for student *i*
- $School_i$ is a vector of school fixed effects
- ϵ_i is the error term clustered at the class level

Creative self-efficacy outcome
Creative self-efficacy will be measured using an adapted version of the ideation sub-measure of the Writing Self-efficacy measure. The sub-measure has three items, which can each be scored with 1-3 points. Each of the 3 scores will be added together and final possible scores will range from 3-9.

In the analysis, we will use a baseline covariate consisting of EYFSP scores aggregated across the following learning goals:

1. exploring and using media and materials (FSP_EXP_G16);
2. being imaginative (FSP_EXP_G17).

The aggregate EYFSP score will range from 2-6. As previously stated, we believe that aggregating the measures is the best way to utilise this data as a baseline measurement to generally indicate whether the pupil is at, above, or below expectations on the learning goals most closely associated with the outcome measure.

Creative self-efficacy analysis
Analysis will follow the model specified for primary analysis, substituting the appropriate secondary outcome measure and baseline measure.

Secondary analysis will be ITT, in which we test the hypothesis that participating in the programme has an effect on student creative self-efficacy. Analysis will use the writing self-efficacy measure raw scores (3-9) and will be carried out using an ordinary least squares (OLS) linear model:

$$Y_i = \alpha + \beta Treat_i + \gamma X_i + \theta School_i + \epsilon_i$$

where:

- Y_i is the raw writing self-efficacy measure score for student *i*
- $Treat_i$ is a binary indicator for the treatment assignment of classroom (1 if the class is assigned to treatment; 0 if not)
- X_i is a vector of baseline attainment measured through aggregated EYFSP learning goal scores for student *i*
- $School_i$ is a vector of school fixed effects
- ϵ_i is the error term clustered at the class level

Interim analyses

No interim analyses are planned.

Sub-group analyses

We will conduct analysis on the primary and secondary outcomes for the sub-group of pupils who have ever been registered for free school meals (FSM) in the NPD (using the EVERFSM_6_P variable), using the same models as specified above, with the addition of an interaction between treatment assignment and FSM status, to assess whether there is a significant difference in the treatment effect between FSM students and others. The model we will use for this analysis is as follows:

$$Y_i = \alpha + \beta_1 \text{ Treat }_i + \beta_2 FSM_i + \beta_3 FSM_i \times \text{Treat}_i + \gamma X_i + \theta School_i + \epsilon_i$$

where:

- Y_i is the primary or secondary outcome specified above for student *i*
- $Treat_i$ is a binary indicator for the treatment assignment (1 if the class is assigned to treatment; 0 if not)
- FSM_i is a binary indicator for student *i*'s EVERFSM_6_P status (1 if the student has been recorded as eligible for free school meals; 0 if not)
- X_i is a vector of baseline attainment specified in the corresponding model above
- $School_i$ is a vector of school fixed effects
- ϵ_i is the error term clustered at the class level

If a significant interaction is found, we will estimate a separate model on the restricted sample of only EVERFSM pupils using the model specified in our primary and/or secondary analysis.

Additional analyses

No additional statistical analyses are planned.

Imbalance at baseline

We will assess imbalance at baseline, and for the sub-sample of those analysed, by calculating the following values in each case and cross-tabulating by treatment arm:

- For aggregate EYFSP scores utilised in primary analysis (understanding, speaking, reading, and writing), we will report the means and standard deviations for the treatment and control group and calculate absolute standardised differences (i.e. the absolute value of the mean difference divided by the sample standard deviation)[20] between the treatment and control groups and these will be presented in the report.
- Count and % EVERFSM

Missing data

We will describe and summarise the extent of missing data in the primary outcomes, and in the model associated with the analysis. Reasons for missing data will also be described. The most likely causes of missing data are the withdrawal by participants from data processing, withdrawal of the school from the study, a student leaving the school, and a student being absent on the day(s) of data collection.

In line with EEF guidelines, any imputation will be restricted to the primary analysis and will only be carried out when more than 5% of the data is missing for a given variable. We will first use a logistic regression to test whether this missing status can be predicted from the following variables: all variables in the analysis model plus eligibility for FSM (and proportion eligible for FSM in the school), and EAL status (and proportion EAL in the school). Where predictability is confirmed (i.e. if the estimated coefficient on any of the explanatory variables in the model is significantly different from zero at the 5 percent significance level) we will proceed to the appropriate next step of this strategy.

For situations for which the missing at random (MAR) assumption appears to hold and any variable other than the outcome variable in the model is missing, we will use all variables in the analysis model plus eligibility for FSM (and proportion eligible for FSM in the school), and EAL status (and proportion EAL in the school) to estimate a Multiple Imputation (MI) model. Multiple imputation (MI) will be carried out using the Markov chain Monte Carlo (MCMC) method to predict the missing values prior to the analysis of treatment effects. We will then estimate the treatment effect using the imputed data in the model associated with the primary analysis and compare our result with the primary analysis (conducted on complete cases only).

Analysis using the multiply imputed dataset will be used as a sensitivity analysis i.e. we will base confirmation of the effectiveness of the treatment on complete case analysis only but assess the sensitivity of the estimate to missingness using the estimates from the multiply imputed dataset. If the complete case analysis model implies effectiveness but the imputed estimate does not we must assume that the missing data is missing not at random to such an extent as to invalidate our conclusion of effectiveness, which we would state in the reporting of the evaluation.

Missing outcome data

Observations with missing outcome data will be dropped from the analysis and a complete case analysis will be run.

[20] Standardised differences are practically the same as effect sizes but are conceptually different, since they are not attempting to quantify an effect.

Compliance

We will estimate the treatment effect across all three outcome measures for compliers using a Complier Average Causal Effect (CACE) analysis using a classroom-level measure of compliance with the intervention. Compliance will be defined at the class level with respect to a teacher missing no more than two training sessions and having delivered 80% of all scheduled daily First Thing Music sessions. The delivery team will ask schools to conduct a register in order to document the number of sessions. Training attendance will be recorded centrally.

The instrument that we will use is treatment assignment, which is assumed to influence whether the class participates in the programme but not the outcome variable in its own right. It is important to note that we do not know the true minimal amount of compliance needed to generate a treatment effect, so the cut-off chosen for minimal compliance is our best estimate, which was defined in coordination with the delivery organisation. This analysis is likely to generate treatment effects that exceed those generated by ITT (unless the treatment is detrimental).

The CACE estimation will use a two-stage least squares (2SLS) approach[21]:

$$Comply_i = \gamma_0 + \gamma_1 Treat_i + \delta School_i + \zeta X_i + \mu_i \quad (1)$$

$$Y_i = \beta_0 + \beta_1 \hat{Comply_i} + \theta School_i + \phi X_i + \epsilon_i \quad (2)$$

where:

- $Treat_i$ is a binary indicator for the treatment assignment (1 if the class is assigned to treatment and 0 if the class is assigned to control);
- $Comply_i$ is a binary indicator for whether student *i*'s teacher met the minimal compliance threshold;
- $School_i$ is a vector of school fixed effects
- X_i is a vector of baseline attainment measured through aggregated EYFSP raw scores for student *I*, as specified in the corresponding primary and secondary analysis models;
- μ_i are the errors in the first stage;
- ϵ_i are the errors in the second stage;
- $\widehat{Comply_i}$ are the predicted levels of compliance with the programme from (1); and
- Y_i is the raw PIRA score for student *i*

Intra-cluster correlations (ICCs)

We will estimate the intra-cluster correlation of the baseline, primary outcome measures, and secondary outcome measures at the classroom-level by estimating a variance components model, as follows:

$$Y_i = \alpha + \gamma_i + \epsilon_i$$

where:

- Y_i is the aggregate EYFSP baseline score specified in the analysis for pre-test ICC and one of the specified outcome measures (PIRA, SSiS, Creative Self-Efficacy) for post-ICC;

[21] See, for instance, Gerber A.S. and Green D.P. (2012). Field Experiments. New York: W. W. Norton & Company.

- γ_i is the classroom-level random-effect; and
- ϵ_i is the individual-level error term

The classroom-level random effect is assumed to be normally distributed and uncorrelated with the individual-level errors.

The ICC itself will be estimated from this model using the following equation:

$$\rho = (var(\gamma_i))/(var(\gamma_i) + var(\epsilon_i))$$

Effect size calculation

Hedges' g effect size will be calculated as follows:

$$g = J(n_1 + n_2 + 2)\frac{\overline{x_1} - \overline{x_2}}{\hat{s*}}$$

where our conditional estimate of $\overline{x_1} - \overline{x_2}$ is recovered from $\beta_1$in the primary ITT analysis model;

$\widehat{s*}$ is estimated from the analysis sample as follows:

$$\hat{s*} = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

where $n_1$is the sample size in the control group, n_2 is the sample size in the treatment group, s_1 is the standard deviation of the control group, and s_2 is the standard deviation of the treatment group (all estimates of standard deviation used are unconditional, in line with the EEF's analysis guidance to maximise comparability with other trials);

and $J(n_1 + n_2 + 2)$ is calculated as follows:

$$J(n_1 + n_2 + 2) = \frac{\Gamma((n_1 + n_2 + 2)/2)}{(\sqrt{((n_1 + n_2 + 2)/2)}\Gamma((n_1 + n_2 + 2 - 1)/2))}$$

If calculating this proves computationally intractable[22] using the above method, we will instead use the following approximation:

$$J(n_1 + n_2 + 2) \approx (1 - 3/(4(n_1 + n_2) - 9)))$$

[22] The output of the gamma (Γ) function in the Hedges' g correction factor (J) becomes large quickly, making this method of computation intractable where $n_1 + n_2$ is not small. As such, it can quickly become intractable. Thankfully, the approximate method tends towards the fully correction factor quickly. As such, where the computational intractability is an issue, the approximate method is appropriate. In any event, the correction factor is likely to be small in this trial.

Ninety-five per cent confidence intervals (95% CIs) of the effect size will be estimated by inputting the upper and lower confidence limits from the regression model into the effect size formula.

All of these parameters will be made available in the report.

Appendix: Analysis Syntax

Provided below is prospective analysis syntax that executes the models specified in this Statistical Analysis plan using Stata. The syntax used in the actual analysis may be slightly different (e.g. variable name differences), but changes will not affect the execution of the models specified in this SAP.

Primary intention-to-treat (ITT) analysis:

```
regress pira i.treat eyfsp_pira i.school_id, vce(cluster class_id)
```

is a linear regression model estimated on individual-level full randomised sample data where *pira* is the Progress in Reading Assessment (PIRA) raw score (corresponding to Y in the regression equation), *treat* is a binary treatment variable (corresponding to $Treat$ in the regression equation), *eyfsp_pira* is the aggregate EYFSP score for the learning goals specified for the primary analysis (corresponding to Xin the regression equation), *school* is a categorical stratification variable (corresponding to $school$ in the regression equation), and *class_id* is a class identifier.

CACE analysis:

```
ivregress 2sls pira eyfsp_pira i.school_id (comply = treat), vce (cluster class_id)
```

is an instrumental variables (two stage least squares) regression model estimated on individual-level full randomised sample data where *pira* is the PIRA test raw score (corresponding to Y in the regression equation), *treat* is a binary treatment variable (corresponding to $Treat$ in the regression equation), *eyfsp_pira* is the aggregate EYFSP score for the learning goals specified for the primary analysis (corresponding to Xin the regression equation), *comply* is a binary indicator of class-level compliance defined in the evaluation protocol, *school_id* is a categorical stratification variable (corresponding to $School$ in the regression equation), and *class_id* is a class identifier.

Sub-group analysis:

```
regress pira i.treat i.EVERFSM_6_P treat#EVERFSM_6_P eyfsp_pira i.school_id, vce(cluster
class_id)
```

is a linear regression model estimated on individual-level full randomised sample data where *pira* is the PIRA test raw score (corresponding to Y in the regression equation), *treat* is a binary treatment variable (corresponding to $Treat$ in the regression equation), EVERFSM_6_P is an indicator of whether an individual has ever been eligible for free school meals (corresponding to FSM in the regression equation), *eyfsp_pira* is the aggregate EYFSP score for the learning goals specified for the primary analysis (corresponding to Xin the regression equation), *school_id* is a categorical stratification variable (corresponding to $School$ in the regression equation), and *class_id* is a school identifier.