Statistical Analysis Plan
Power of Pictures

Evaluator: UCL Institute of Education and BIT
Principal investigators: Jake Anders and Nikki Shure

PROJECT TITLE	Power of Pictures
DEVELOPER	Centre for Literacy in Primary Education
EVALUATOR	UCL Institute of Education and the Behavioural Insights Team (BIT)
PRINCIPAL INVESTIGATOR(S)	Jake Anders, Nikki Shure
TRIAL (CHIEF) STATISTICIAN	Jake Anders
SAP AUTHOR(S)	Jake Anders, Nikki Shure Quality assurance by Pantelis Solomon
TRIAL REGISTRATION NUMBER	ISRCTN15334278
EVALUATION PROTOCOL URL OR HYPERLINK	Link to evaluation protocol

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	16 October 2018	N/A

Table of contents

Introduction

The Power of Pictures (PoP) is a programme that uses the creation of picture books to develop primary school pupils' writing skills. Through its programme, PoP aims to enhance teachers' understanding of the power of picture books and increase teachers' comfort in and ability to teach using picture books to a range of ages. In the year prior to the trial, the PoP team has worked with 65 teachers in 40 schools.

The PoP 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 and the Royal Society for the Arts.[1] It consists of five evaluations: two in Key Stage 1 (Reception and Year 1) and three in Key Stage 2 (Year 5). Despite the unique aspects of these intervention models, there are many similarities in how they are delivered and what they hope to achieve.

PoP operates as follows. The PoP team selects picture book authors with whom they wish to work. This selection is based on the quality and relevance of a specific book from the author's work, which will be used for the intervention. The PoP team work with the author and the illustrator (always the same person) of this book throughout the course of the intervention. The author-illustrator is selected based on the quality of their text and their experience of talking about their process either through direct teaching, workshops at festivals or similar. They then take part in training with the PoP team before the intervention in the schools begins.

Schools that are selected to take part in the intervention must commit at least four weeks of their English curriculum to PoP. Typically, two teachers from a school are selected to lead on the intervention: one from Year 1 and one from Year 4 or 5.

The teachers selected from each school receive 2.5 days of training between October and March directly from the author-illustrator and the PoP team. The first day (in October) of training focuses on conveying meaning through drawing, reading pictures, creative approaches to using books and feedback. Throughout the entire process, the PoP team makes web resources available to the teachers, including sample teaching sequences.

Following the first day of training, the teachers are expected to do a 'gap task' with their pupils, where the pupils work with a picture book by the focus author/illustrator. This task is documented via photos, which are sent to the PoP team. During this time (in November), the schools attend a half day visit delivered by the PoP team and the author-illustrator, where they get to meet the author/illustrator and take part in a workshop focussed on character design and development of narrative.

The second day of training for the teachers, which occurs in January, focuses on the full writing process for a book, how to design characters, sequential storytelling, feedback on writing, publication and reflection. This training session is delivered by the same PoP team and author-illustrator as the first training session. The teachers then return back to their class for another six-week period and use the teaching sequences in more detail to focus on the innovation of pupil's own picture book stories, inspired by the workshop. This is again documented via photos sent to the PoP team.

The third and final half-day of training involves not only the teachers who have previously participated, but also the school's senior leadership team and the literacy coordinator. On this

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

day, reflection takes places as to the effectiveness of approaches and an action plan is formulated for how to continue incorporating picture books and illustration into schools' curriculum, based on successes shared.

The evaluation is structured as a two-armed school-level cluster randomised controlled trial involving 101 primary schools. 51 schools were allocated to receive the intervention and 50 to a business as usual control group. Recruitment occurred in Summer 2018 with the aim of starting the intervention with the cohort of pupils starting Year 5 in September 2018. The evaluation will look at the impact of the programme on writing attainment, measured by the Writing Assessment Measure (WAM), as its primary outcome.

Design overview

Trial type and number of arms		Cluster randomised, two arms
Unit of randomisation		School
Stratification variables (if applicable)		Proportion of FSM-eligible students; proportion of EAL students
Primary outcome	variable	Writing attainment
	measure (instrument, scale)	Writing Assessment Measure (WAM) score
Secondary outcome(s)	variable(s)	Writing self-efficacy Ideation Reading attainment
	measure(s) (instrument, scale)	Writing Self-Efficacy Measure (WSEM) Ideation sub-measure of the WSEM Reading test performance in KS2 SATS

This is a clustered randomised controlled trial, with randomisation taking place at the school level. As the programme involves training teachers, the choice was between randomising at school or class level. While the fact that a single class is identified as the treatment or control group in each school suggests that class-level randomisation would be possible, there were concerns that this would have entailed significant risk of cross-contamination. The trial recruited 101 primary schools, with schools randomly allocated to either the treatment arm or the control group. Schools in the control group are expected to continue with 'business as usual', and will be offered the opportunity to take part in the programme following the completion of the study.

The project team advertised the trial and also approached schools through their existing networks. Where possible it aimed to recruit schools that have larger populations of individuals receiving Free School Meals (FSM) than the national average of 15.3 per cent of pupils aged 5-10.[2]

The eligibility criteria for schools to participate were:

- participating schools must be state-funded primary schools recruited from the following regions: London, NE England and East Sussex;

[2] Department for Education (2016). Schools, pupils and their characteristics: January 2016, SFR20/2016. London: England.

- schools had to agree to distribute study information sheets, data privacy information, and data processing objection forms to parents;
- schools had to agree that, if allocated to the control group, they would continue with 'business as usual' for the duration of the trial;
- schools had to return a signed Memorandum of Understanding, including committing to participate fully in the study – including the collection of outcome measures in summer 2019 – regardless of which trial arm they are assigned to;
- schools had to agree to allow time for each assessment phase and liaise with the evaluation team to find appropriate dates and times for assessments to take place; and
- schools had to agree that teachers in both trial arms cooperate with activities for the implementation and process evaluation, if requested.

Randomisation followed recruitment of schools, including the signing of Memoranda of Understanding (MoUs) and baseline data collection, in March-July 2018. Randomisation was stratified on school-level characteristics (proportion of FSM students and proportion of EAL students) to ensure balance between treatment and control groups on these characteristics. This was conducted using Stata. The randomisation followed the following process:

1. The schools were stratified into four blocks on the basis of proportion of FSM students (split across the median sample proportion) and proportion of EAL students (split across the median sample proportion).
2. Each school was assigned a randomly generated number (setting a stable seed for the random number generation).
3. The schools were sorted by block and random number.
4. Schools were assigned to the treatment arm and to the control arm in turn.

Follow-up

The recruitment target was set at 120 but unfortunately only 101 were fully recruited for randomisation, with a significant number dropping out shortly before randomisation due to concerns about updated data protection legislation that was being introduced at the time.

Sample size calculations overview

		Protocol		Randomisation[3]	
		OVERALL	FSM	OVERALL	FSM
MDES		0.18	0.29	0.21	0.26
Pre-test/ post-test correlations	level 1 (pupil)	0.50	0.50	0.50	0.50
	level 2 (class)	0.50	0.50	0.50	0.50
	level 3 (school)	N/A	N/A	N/A	N/A
	level 2 (class)	0.15	0.15	0.15	0.15

[3] See important notes below regarding cluster size assumptions.

Intracluster correlations (ICCs)	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		23	3	22	6
Number of schools	intervention	60	60	51	51
	control	60	60	50	50
	total	120	120	101	101
Number of pupils	intervention	1380	180	1122	306
	control	1380	180	1100	300
	total	2760	360	2266	618

Protocol MDES calculations were based on the following assumptions:

- **Randomisation performed at the school-level.** However, as we are only testing outcomes in one class per school, for analysis purposes it makes sense to think of class-level clustering. All children in a class will be in the same trial arm, a requirement of this trial given we are testing the effect of teacher training.
- **Number of children per cluster is 23.** This is an estimate of the average number of children in each class reported in the protocol (25) with a 10% adjustment for attrition.
- **An intracluster correlation coefficient (ICC) of 0.15.** This defines how alike individual children are within each class. The ICC increases the more individuals within the clusters resemble one another. An ICC of 0.15 is based on EEF's guidance on ICCs. In the absence of ICC data for our outcomes of interest we use this guidance, specifically for the reading fine points score, and, given uncertainty about the geographical spread of participating schools, we use the highest regional ICC (which happens to be Inner London) to the nearest two decimal places.
- **Power: 80%; Significance level: 5%.** These are standard assumptions.

Randomisation MDES calculations were based on the following additional assumptions, required due to the organisational need to confirm final class after randomisation (where it was not possible to confirm these in advance of randomisation then schools provided class lists across the relevant year group):

- **Number of children per cluster is 22.** This is an estimate of the average (arithmetic mean) number of children in each class (25) based on the number of children per class reported among schools that do not reorganise classes between years and, hence, were able to confirm final class lists prior to randomisation (81% of the total) with a 10% adjustment for attrition (90% of 25 if 22.5, so we have rounded down to the nearest whole number to be conservative).
- **Number of children eligible for FSM per cluster is 6.** This is an estimate of the average number of children eligible for FSM in each class (7) based on the number of

children per class reported among schools that do not reorganise classes between years and, hence, were able to confirm final class lists prior to randomisation (81% of the total) with a 10% adjustment for attrition. Note that this is double the estimate used for power calculations in the protocol, hence the estimated FSM sample size has increased substantially.

As we will use data on pupils' performance in the Year 1 phonics screening check (PSC) available from the National Pupil Database (NPD), consistent with EEF policy to use an administrative measure rather than an additional pre-test where possible) as a pre-test, the predictive power of this has also been factored into our sample size calculations. An appropriate pre-test/post-test correlation assumption could not be estimated empirically for this trial, since correlation data between the pre- and post-tests used in this trial were unavailable. This is because the PSC has only been in place since 2012, and our post-test (the WAM) is an even newer measure.

EEF guidance suggests that a pre- and post-test correlation of 0.7 in education research is common;[4] however, we see this as too optimistic in this case. The 21-day test-retest correlation coefficient of the WAM is reported to be 0.82,[5] but the time elapsed between the pre- and post-test in this trial is much longer, and we will not be using the WAM itself as a baseline. Our pre-test (score in Year 1 phonics screening check) has less variance than would be ideal, due to a degree of bunching between the pass (32) and highest available mark (40). Nevertheless, given its closer temporal proximity to the post-test point, we believe it is likely to explain more variance in our post-test than earlier measures also available in the NPD (which would have to be measured at the Early Years Foundation Stage).

While there is no direct measure of the pre-test/post-test correlation between the WAM and the phonics screening check available, a value of 0.52[6] has been estimated using Year 1 phonics screening check scores and Progress in International Reading Literacy Study (PIRLS) scores[7] (taken in Year 5, the same year as the WAM will be administered). Given the similar time period between pre-test and post-test administration, and the related domain, we believe this estimate is likely to approximate the value that will be observed in this trial. Based on this, we assumed that 25 per cent of post-test variance at both pupil- and school-level is explained by the pre-test (equivalent to pre-test/post-test correlation of 0.5).

These assumptions suggested a requirement of 113 schools to achieve an MDES of 0.20. Based on discussions with the PoP team at the set-up meetings, we agreed on a sample size of 120 schools. They confirmed that recruitment of 120 schools and intervention delivery to 60 treatment schools are reasonable and achievable numbers given their capacity.

[4] Torgerson, C. & Torgerson, D. (2013). Randomised trials in education: An introductory handbook. EEF.
[5] Dunsmuir, S., Kyriacou, M., Batuwitage, S., Hinson, E., Ingram, V. & O'Sullivan, S. (2015) An evaluation of the Writing Assessment Measure (WAM) for children's narrative writing. Assessing Writing 23(2015) 1-18.
[6] No guidance is available from this analysis on explanatory power at the pupil-level and school-level.
[7] Department for Education (2017). Progress in International Reading Literacy Study (PIRLS): National Report for England. December 2017.

Analysis

Primary outcome analysis

Our primary analysis will focus on the Writing Assessment Measure score, and will be performed using Stata.[8] All continuous variables will be used in their 'raw' form (in line with EEF guidance) as there is no clear reason to transform the data.

Outcome variables will be regressed using an ordinary least squares (OLS) model on treatment arm indicators, strata indicators (based on proportion of the class eligible for FSM, proportion of the class identified as EAL, and whether the school was randomised as part of the first or second batch), and pre-test phonics screening check score (further details below).

As noted by EEF guidance, in a model that does not account for clustering, when this is a feature introduced by the experimental design, "the point estimates will be accurate, but the standard errors will be downward biased" (EEF, 2018, p.3)[9]. However, we can account for the potential effects of the experimental design in this respect by calculating standard errors taking into account clustering (Angrist & Pischke, 2009)[10] at the school level which allow for correlation of pupil outcomes within schools. We prefer this to use of a hierarchical linear model which makes additional assumptions about the school-level effects that may not be justified.

The estimated impacts will be intention-to-treat (ITT) effects and will be reported with 95% confidence intervals. Intra-cluster correlations will also be reported. We will estimate the following model:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \boldsymbol{\gamma}' \boldsymbol{X}_j + \varepsilon_{ij}$$

where individual i is nested in school j, Y_{ij} is the Writing Assessment Measure (WAM) score, $PreTest_{ij}$ is the value of the phonics screening check score (using the NPD variable PHONICS_MARK) used as a pre-test, $Treat$ is our school-level treatment indicator, $\boldsymbol{X}$ is a vector of stratification variables, and ε is an error term. Errors will be clustered at school-level (j).

Our primary intention-to-treat outcome will be recovered from the estimate of β_1 when this model is estimated on the full sample at randomisation. This model will not be altered depending on the significance of any variables included (i.e. all variables will be retained in the model regardless of whether they are statistically significant) including the vector of blocking variables ($\boldsymbol{X}_j$). Example syntax for this primary analysis model is reported in the analysis syntax appendix.

Secondary outcome analysis

We will conduct three secondary outcome analyses:

- Writing Self-Efficacy: Same as the primary outcome analysis except replace Y_{ij} with the Writing Self-Efficacy Measure score and $PreTest_{ij}$ with assessment of pupils' Personal, Social and Emotional Development skills from the EY Foundation Stage

[8] 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.
[9] EEF (2018) Statistical analysis guidance for EEF evaluations.
[10] Angrist, J. D., & Pischke, J. S. (2009). Mostly Harmless Econometrics. Princeton: Princeton University Press.

Profile (aggregated scores from NPD FSP_PSE_G06, FSP_PSE_G07 and FSP_PSE_G08).

- Ideation: Same as the primary outcome analysis except replace Y_{ij} with the Ideation sub-score from the Writing Self-Efficacy Measure and $PreTest_{ij}$ with assessment of pupils' Personal, Social and Emotional Development skills from the EY Foundation Stage Profile (aggregated scores from NPD FSP_PSE_G06, FSP_PSE_G07 and FSP_PSE_G08).
- KS2 grammar, punctuation and spelling test attainment: Same as the primary outcome analysis except replace Y_{ij} with the KS2 grammar, punctuation and spelling test. Note that the results for this outcome will not be available until 2020, which is after the trial concludes; therefore, this analysis will not be included in the initial report. The results from this outcome are planned to be included in a separate report reflecting on all the projects from this round of funding to be published at a later stage.

Interim analyses

No interim analyses are planned.

Subgroup analyses

Following EEF guidance, we will first test for an interaction of the treatment and FSM status (using the NPD variable EVERFSM_6_P) using the following model:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 FSMEver_{ij} + \beta_3 Treat_j * FSMEver_{ij} + \beta_4 PreTest_{ij} + \boldsymbol{\gamma'} \boldsymbol{X}_j + \nu_{ij}$$

where individual i is nested in school j, Y_{ij} is the Writing Assessment Measure (WAM) score, $PreTest_{ij}$ is the value of the phonics screening check score used as a pre-test, $Treat$ is our school-level treatment indicator, $FSMEver$ is an indicator of FSM eligibility (EVERFSM_6_P), $Treat * FSMEver$ is an interaction between these two terms, $\boldsymbol{X}$ is a vector of stratification variables, and ν is an error term. Errors will be clustered at school-level (j). Example syntax for this interaction model is reported in the analysis syntax appendix.

If a significant interaction is found (i.e. the absolute value of the point estimate of β_3 divided by the school-level clustered standard error is greater than 1.96), we will conduct a specific sub-group analysis for those who have ever been registered for Free School Meals (FSM) in the National Pupil Database (identified using the variable EVERFSM_6_P) using the same model as our primary analysis.

This sub-group was identified in the trial protocol and FSM-eligible pupils are a key subgroup to be analysed in all EEF trials. The subgroup analysis will be conducted for both the primary and secondary outcomes.

Additional analyses

No additional statistical analyses are planned.

Imbalance at baseline

We will check for balance of analysed sample for the following characteristics:

- pre-test phonics screening check score; and
- proportion ever eligible for Free School Meals.

We will do this by reporting means and standard deviations for the treatment and control group and calculating absolute standardised differences (Imbens & Rubin, 2015)[11] (i.e. the absolute value of the mean difference divided by the sample standard deviation)[12] between the treatment and control groups and these will be presented in the report. These provide a simple, scale-free measure of differences that is easy to interpret.

[11] Imbens, G. M. and D. B. Rubin (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York, NY, Cambridge University Press.

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

Missing data

We will describe and summarise the extent of missing data in the primary and secondary outcomes, and in the model associated with the analysis. Reasons for missing data will also be described.

For all models we will implement a missing data strategy if more than 5% of data in the model is missing or if more than 10% of data for a single school is missing. The strategy will be followed separately for each instance of model and variable for which the threshold is exceeded:

- We will first assess whether the missing data is missing at random (MAR), since this is a pre-requisite for missing data modelling to produce meaningful results. To do this we will create an indicator variable for each variable in the impact model specifying whether the data is missing or not. We will then use 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 we will proceed to the appropriate next step of this strategy.

- For situations for which the MAR assumption appears to hold and only the outcome variable in the model is missing, we will re-estimate the treatment effect using our pre-specified model with the addition of the covariates found to be statistically significantly predictive of missingness of the outcome.

- For situations for which the 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 using a fully conditional specification, implemented using Stata MI to create 20 imputed data sets. We will re-estimate the treatment effect using each dataset and take the average and estimate standard error using Rubin's combination rules.[13]

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.

Compliance

The following criteria have been defined in the trial protocol as variables that can be used to assess dosage of the intervention. This draws principally on attendance data collected from the project team.

The fidelity of this intervention will be measured at the teacher level, which reflects the intervention delivery method. A school will be considered to have complied if and only if the following three conditions are met:

- Participating teacher attends all training sessions (2.5 days);

[13] Rubin, D. (2004). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

- Participating teacher delivers at least 50% of lessons in both block one and two (10/20); and
- Participating teacher's class needs to attend the half day author off-site workshop.

We will use Complier Average Causal Effect (CACE)[14] analysis to estimate intervention effects on treated children. We will estimate the CACE using two stage least squares (2SLS) regression by estimating a (first stage) model of compliance, as follows:

$$Comply_j = \alpha + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \boldsymbol{\gamma' X}_j + \xi_{ij}$$

where $Comply$ is the binary compliance variable defined above, and ξ is an error term. The predicted values of $Comply$ from the first stage are used in the estimation of a (structural) model of our outcome measure Y_{ij}. In other respects, the specification remains the same as the primary outcome ITT model. This second stage model is specified as follow:

$$Y_{ij} = \alpha + \beta_1 \widehat{Comply}_j + \beta_2 PreTest_{ij} + \boldsymbol{\gamma' X}_j + \omega_{ij}$$

where $\widehat{Comply}_j$ are the predicted values of treatment receipt derived from the first stage model, and ω is an error term. Our primary outcome of interest will be β_1, which should recover the effect of the intervention among compliers. We will conduct this analysis using the ivregress functionality of Stata to make necessary adjustments to standard errors (which will also be clustered at school level) due to the instrumental variables approach. Example syntax for this CACE model is reported in the analysis syntax appendix.

Intra-cluster correlations (ICCs)

- Describe the model that will be used to estimate the ICCs at pre- and post-test, for each level at which they will be computed (state which level they are computed at).

In order to estimate the intra-cluster correlation (ICC) of the pre-and post-tests at school-level we will employ an empty variance components model, as follows:

$$\boldsymbol{Y_{ij} = \alpha + \eta_j + \varepsilon_{ij}}$$

where individual $\boldsymbol{i}$ is nested in school $\boldsymbol{j}$, $\boldsymbol{Y_{ij}}$ is the Writing Assessment Measure (WAM) score for the purpose of calculating the post-test ICC or the value of the phonics screening check score for the purpose of calculating the pre-test ICC, $\boldsymbol{\eta_j}$ is a school-level random effect, and $\boldsymbol{\nu_{ij}}$ is an individual-level error term. The school-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:

$$\boldsymbol{\rho = \frac{var(\eta_j)}{var(\eta_j)+var(\varepsilon_{ij})}}$$

Effect size calculation

Hedges' g effect size will be calculated as follows:

[14] Gerber AS, Green DP. (2012). Field Experiments: Design, analysis and interpretation. WW Norton and Company, New York.

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

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:

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

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:

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

where n_1 is the sample size in the control group and n_2 is the sample size in the treatment group.

If calculating $J(n_1 + n_2 + 2)$ proves computationally intractable[15] using the above method, we will instead use the following approximation:

$$J(n_1 + n_2 + 2) \approx (1 - \frac{3}{4(n_1 + n_2) - 9})$$

Ninety-five per cent confidence intervals (95% CIs) of the effect size will be estimated by inputting the upper and lower confidence limits of $\widehat{\beta_1}$ from the regression model into the effect size formula.

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

[15] 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.

Appendix: Analysis Syntax

In this appendix, we provide indicative analysis syntax to implement the models specified in the Statistical Analysis Plan using Stata. Eventual syntax may have small changes (e.g. variable name changes) that do not affect the syntax's implementation of the models specified above. Variables are as specified in the statistical analysis plan.

Primary intention-to-treat (ITT) analysis:

```
regress wam i.treat phonics_score i.block, vce(cluster school_id)
```

is a linear regression model estimated on individual-level full randomised sample data where `wam` is the Writing Assessment Test (WAM) raw score (corresponding to Y in the regression equation), `treat` is a binary treatment variable (corresponding to $Treat$ in the regression equation), `phonics_score` is the phonics screening check score (corresponding to $PreTest$ in the regression equation), `block` is a categorical stratification variable (corresponding to X_j in the regression equation), and `school_id` is a school identifier (corresponding to j in the regression equation).

CACE analysis:

```
ivregress 2sls wam phonics_score i.block (comply = treat), vce (cluster school_id)
```

is an instrumental variables (two stage least squares) regression model estimated on individual-level full randomised sample data where `wam` is the Writing Assessment Test (WAM) raw score (corresponding to Y in the regression equation), `treat` is a binary treatment variable (corresponding to $Treat$ in the regression equation), `phonics_score` is the phonics screening check score (corresponding to $PreTest$ in the regression equation), `comply` is a binary indicator of school compliance defined in the evaluation protocol, `block` is a categorical stratification variable (corresponding to X_j in the regression equation), and `school_id` is a school identifier (corresponding to j in the regression equation).

Sub-group analysis:

```
regress wam i.treat i.EVERFSM_6_P treat#EVERFSM_6_P phonics_score i.block,
vce(cluster school_id)
```

is a linear regression model estimated on individual-level full randomised sample data where `wam` is the Writing Assessment Test (WAM) 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 $FSMEver$ in the regression equation), `phonics_score` is the phonics screening check score (corresponding to $PreTest$ in the regression equation), `block` is a categorical stratification variable (corresponding to X_j in the regression equation), and `school_id` is a school identifier (corresponding to j in the regression equation).