Statistical Analysis Plan
Speech Bubbles

Evaluator (institution): Behavioural Insights Team and UCL Institute of Education
Principal investigator(s): Pantelis Solomon

Template last updated: March 2018

PROJECT TITLE	Using the Speech Bubbles programme to improve pupil attainment in school
DEVELOPER (INSTITUTION)	London Bubble Theatre Company
EVALUATOR (INSTITUTION)	Behavioural Insights Team (BIT) & UCL Institute of Education (IoE)
PRINCIPAL INVESTIGATOR(S)	Pantelis Solomon
TRIAL (CHIEF) STATISTICIAN	Pantelis Solomon
SAP AUTHOR(S)	Pantelis Solomon and Kim Bohling Quality Assurance by Jake Anders and Nikki Shure
TRIAL REGISTRATION NUMBER	ISRCTN14448319
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Speech_Bubbles_protocol.pdf

Statistical Analysis Plan (SAP) version history

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

Table of contents

Table of Contents

Introduction

The Speech Bubbles intervention aims to improve children's reading, communication and social skills by providing them with weekly creative drama sessions. This is an intervention targeted at pupils with below expected communication and social skills. The model that will be tested comprises 24 weekly drama sessions for Year 1 and Year 2 pupils (5- to 7-year-olds) over the course of three terms. During the sessions, trained practitioners will encourage children to tell, act out and reflect on their own stories by creating a safe and playful environment, promoting children's communication, confidence and wellbeing.

The evaluation is designed as a two-armed individually randomised, randomised controlled trial involving 26 primary schools with a total of 1,009 pupils. 504 pupils were randomly allocated to receive the intervention and 505 pupils were randomly assigned to be in the control group and not receive any intervention. The randomisation was stratified at the school level such that in each school 50% of pupils were allocated to the treatment group. Recruitment occurred in the winter/spring of 2017/18 with the aim of starting the intervention with the September 2019 cohort of Year 1 and Year 2 pupils.

The evaluation has two primary outcomes:

- Reading attainment, measured by the Progress in Reading Assessment (PIRA) by Rising Stars.[1]
- Oral communication measured by the Renfrew Bus Story test.

There are two primary outcomes because oral communication was seen as an important primary outcome alongside the more standard reading attainment indicator.

Secondary outcomes will measure the programme's effect on social skills, as measured by the Social Skills Improvement System (SSiS)[1], and on creative self-efficacy, as measured by the ideation sub-measure of the writing self-efficacy measure.[2]

Design overview

Trial type and number of arms		Two-arm, individually randomised
Unit of randomisation		Pupil
Stratification variables (if applicable)		School, year
Primary outcome	variable	(1) Reading attainment, (2) Oral communication
	measure (instrument, scale)	(1) PIRA, score range 0-25, (2) Renfrew Bus Story
Secondary outcome(s)	variable(s)	(1) Social skills, (2) Creative self-efficacy
	measure(s) (instrument, scale)	(1) SSiS -- social skills sub-measure, score range 4-12, (2) Writing self-efficacy measure -- ideation sub-measure (3 questions), 3-point Likert scale, score range 3-9

This is an individually randomised controlled trial. The trial recruited 1009 children across 26

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

[2] 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

schools, with pupils randomly allocated to either the treatment arm (who will receive the programme) or the control group. The aim was to recruit 40 children within each school and assign to the treatment and control conditions at a 50:50 ratio. Pupils in the control group continue on a 'business as usual' basis.

In order to participate in the study schools needed to:

- be located in the North West England, South London and East London (for programme delivery purposes);
- be at least a two-form entry school (to reach the required sample size across a smaller number of schools)[3];
- have discussed participation with Speech Bubbles and signed a Memorandum of Understanding (MoU) detailing the conditions of participation (opt-out process, pupil data provision, endline assessment, participation in IPE activities etc.); and
- be able to refer 40 children into the study.

Schools with an average or above average share (14.1%[4]) of Free School Meal (FSM) children received priority in recruitment.

To enter the study, pupils needed to be in Years 1 or 2 at the time of intervention and be referred by their teachers. The referral process was based on guidance from the Speech Bubbles programme which targets the programme at children who:

- Lack confidence in communicating;
- Have difficulty organising thoughts and communicating them;
- Have poor attention and poor listening.

Randomisation followed recruitment of schools, including the signing of MoUs, which was concluded in February 2018. Randomisation was stratified at the school and year level (Years 1 and 2). This was conducted using Stata as follows:

1. If there are more than 40 children referred,[5] we contacted the schools and asked them to restrict the sample to 40 children.
2. Within each school, children were stratified into two blocks, based on their year level. Each student was assigned a randomly generated number within each block and half the children within each block were assigned to the treatment.

Given the degree of uncertainty about test-retest correlation for both primary outcomes, the original target was to recruit 25 schools. A total of 26 schools were approached to account for schools dropping out of the process.

Follow-up

The original recruitment target of 25 schools was exceeded as Speech Bubbles approached 26 schools and all were participating at the time of writing this SAP.

[3] With an exception of one pre-agreed school

[4] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf

[5] Schools were discouraged from doing this, and asked to prioritise referring those students they believe would most benefit from the intervention. The targeting of these students is how the intervention is used more generally.

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.20	0.51	0.17	0.29
Pre-test/ post-test correlations	level 1 (pupil)	0.30	0.30	0.30	0.30
	level 2 (class)	NA	NA	NA	NA
	level 3 (school)	NA	NA	NA	NA
Intracluster correlations (ICCs)	level 2 (class)	NA	NA	NA	NA
	level 3 (school)	NA	NA	NA	NA
Alpha		0.05	0.05	0.025	0.025
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		NA	NA	NA	NA
Number of schools	intervention	23	23	26	26
	control	23	23	26	26
	total	23	23	26	26
Number of pupils	intervention	460	65	504	165
	control	460	65	505	179
	total	920	130	1009	344

Protocol MDES calculations were based on the following assumptions:

- **Randomisation will be performed at an individual level.** This means that referred pupils were randomly allocated to either the treatment or the control group.
- **Number of treatments:** There are two trial arms (treatment and control) with 40 children in each school split equally into control and treatment groups.
- **Attrition:** We have assumed a 20% attrition rate for the endline outcome measure for various reasons (e.g. attrition due to changing school, prolonged absence, inability to engage with the endline assessments). This estimate is based on the 15% standard post-randomisation attrition rate in EEF studies,[6] plus an additional allowance for children whose parents objected to their data being used for the study (5%). This reduced the minimum number of children per arm within the school for the purposes of sample size calculations to 17.
- **Alpha and Power:** We assumed 80% statistical power and 5% significance level at the trial protocol and 2.5% at randomisation. This is because at the time of drafting the trial protocol, it was not yet confirmed that the Renfrew Bus Story would be used as a primary outcome. This was due to concerns over whether pupils targeted for participation in the trial would be able to engage with a research assistant (RA) when the assessment was administered. A pilot was conducted in June 2018 and confirmed that students were sufficiently engaged with the RA conducting the

[6] 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

assessment and it was therefore decided to include the Renfrew Bus Story as a primary outcome measure. Although the delivery of the assessment did not prove challenging, the marking did. We sought help from speech and language specialists who helped resolve marking inconsistencies and agreed to be involved in RA training prior to endline data collection.[7] Based on EEF statistical analysis guidelines when using dual primary measures, we are applying a Bonferroni correction which reduces the alpha to 2.5%.[8]

- **Test-retest correlation:** The baseline achievement measure used is the Early Years Foundation Stage Profile (EYFSP). The only estimate for the test-retest correlation between the EYFSP and the reading assessment PIRA was 0.61. This was based on unpublished analysis from the Fisher Family Trust (FFT) conducted at the end of year 1 for a prior EEF trial (ABRA: Online Reading Support).[9] However, given that our study targets a specific population, we opted for a conservative estimate of 0.3. For the Renfrew Bus Story assessment of oral communication skills, we do not have any information on its correlation with EYFSP and therefore we conducted power calculations for a range of values between 0-0.8.
- **Free School Meals:** In order to estimate the MDES for FSM students we assumed that the FSM sub-group is 14.1 per cent of the total sample (based on data from DfE statistics,[10] and maintained the expected test-retest correlation coefficient value of 0.30.

Analysis

The analysis plan is described in the sections that follow. All analyses will be carried out using the statistical software Stata[11] (see Appendix 1 for the prospective Stata syntax).

Primary outcome analysis

The evaluation has two primary outcome measures:

- Reading attainment measured by the PIRA by Rising Stars.
- Oral communication measured by the Renfrew Bus Story test, which is short standardised test that assesses narrative aspects of oral language.

The estimated impacts will be intention to treat (ITT) effects. As we are testing two primary outcome measures, we will apply a Bonferroni correction, thus reporting with 97.5% confidence intervals.

Reading attainment

The primary outcome measure for reading will be the PIRA by Rising Stars.[12] 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

[7] Although the delivery of the assessment did not prove challenging, the marking did. We sought help from a speech and language specialists who helped resolve marking inconsistencies and agreed to be involved in RA training prior to endline data collection.

[8] Statistical analysis guidance for EEF evaluations (March 2018). Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf

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

[10]https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf

[11]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.

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

prior EEF evaluations.[13] [14] Endline PIRA assessments will be conducted during May - June 2019 by trained RAs who will be blind to trial arm assignment. Rising Stars, the publisher of PIRA, will mark the assessments.

As different versions of the PIRA test will be used for the Year 1 and 2 cohorts, raw scores will be standardised to have a mean of zero and a standard deviation of one prior to combining cohorts for the purpose of analysis.

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 and oral communication, our primary outcome measures. 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.[15] 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"[16]

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 measure.

The analysis will use standardised PIRA scores across the two classes and will be carried out using an ordinary least squares (OLS) linear model:

---

[13] 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.

[14] 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

[15] 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.

[16] 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

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

where

- Y_i is the standardised PIRA score for student *i*
- $Treat_i$ is a binary indicator for the treatment assignment for student *i* (1 if the student 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*
- $Year_i$ is a binary variable for the year group (1 for Year 2 and 0 for Year 1)
- $School_i$ is a vector of school fixed effects
- ϵ_i is the individual error term

Given the assumptions about the baseline measure, we will conduct exploratory analysis using a more flexible specification of 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.

Oral Communication

The outcome measure for oral communication will be the Renfrew Bus Story[17] test. The Renfrew Bus Story is short standardised test that assesses narrative aspects of oral language. Pupils' ability to recall the story is measured based on information content, sentence length, grammatical usage and independence. The assessment of narrative skills is a growing area of research. However, the Renfrew Bus Story remains the most commonly used measure.[18] This assessment has some evidence of moderate test-retest reliability and high inter-rater reliability on two of the three constructs measured.[19] We initially had some concerns about whether children with speech delays or challenges would be adequately able to engage with the assessment, so we conducted a pilot in the year prior to the evaluation in three schools with 88 children – most of whom were currently in the Speech Bubbles programme. The majority displayed full to partial engagement, with only two children not engaging at all. The results of this pilot provided confidence that the test is suitable to deliver to the vast majority of pupils taking part in the evaluation.

The assessment will be conducted on a one-to-one basis by RAs trained in language assessment by an experienced child psychologist. These RAs will be blind to trial arm assignment.

Outcome variables will be regressed using an OLS model on treatment arm indicators, strata indicators (year indicators and school fixed effects), and pre-test raw EYFSP score. For reading attainment, our baseline covariate will be the EYFSP composite score for four learning goals:

1) understanding (FSP_COM_G02);
2) speaking (FSP_COM_G03);
3) listening and attention (FSP_COM_G01)

These goals were selected as they are most closely linked to oral communication, one of our co-primary outcome measures. For each goal, teachers will assign point scores as follows:

- Not yet meeting expectation (emerging) - 1 point

[17] http://www.talkingpoint.org.uk/slts/assessment-children-slcn/expressive-language-assessments

[18] Dockrell, J. E. (2001). Assessing language skills in preschool children. *Child Psychology and Psychiatry Review*, *6*(2), 74-85.

[19] Education Endowment Foundation. Early Years Measures Database. Retrieved from: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/early-years-measure-database/early-years-measures-database/bus-story/

- Meeting expectation (expected) - 2 points
- Exceeding expectation (exceeding) - 3 points
- Not assessed (A) – coded as "missing"

The aggregate score will range from 3 to 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.

The analysis will use the Renfrew Bus Story test and will be carried out using an OLS linear model:

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

where

- Y_i is the Renfrew Bus Story score for student *i*;
- $Treat_i$ is a binary indicator for the treatment assignment for student *i* (1 if the student 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*
- $Year_i$ is a binary variable for the year group (1 for Year 2 and 0 for Year 1)
- $School_i$ is a vector of school fixed effects
- ϵ_i is the individual error term

As described for the reading attainment analysis, we will conduct the exploratory analysis using a more flexible specification.

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 SSiS.[20] The SSiS Social Skills scale assesses pupils' skills across the following sub-scales: communication, cooperation, assertion, responsibility, empathy, engagement and self-control.

SSiS is a commonly used social skills assessment for young children, is standardised and has been used in prior EEF evaluations.[21] 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 the 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 across the following learning goals:

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

[21] 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.

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 3-9. As previously stated, we believe that aggregating the measures is the best way to utilise this data as a baseline measurement to indicate whether the pupil is at, above, or below expectations on the learning goals most closely associated with the outcome measure.

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 OLS linear model:

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

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 for student *i* (1 if the student 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*
- $Year_i$ is a binary variable for the year group (1 for Year 2 and 0 for Year 1)
- $School_i$ is a vector of school fixed effects
- ϵ_i is the individual error term

Creative self-efficacy analysis

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 three 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.

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 OLS linear model:

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

where:

- Y_i is the raw writing self-efficacy measure score for student *i*

- $Treat_i$ is a binary indicator for the treatment assignment for student *i* (1 if the student 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*
- $Year_i$ is a binary variable for the year group (1 for Year 2 and 0 for Year 1)
- $School_i$ is a vector of school fixed effects
- ϵ_i is the individual error term

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 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:

$$Y_i = \alpha + \beta_1 Treat_i + \beta_2 FSM_i + \beta_3 FSM_i \times Treat_i + \gamma X_i + \delta Year_i + \theta School_i + \epsilon_i$$

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 FSM; 0 if not)
- X_i is a vector of baseline attainment specified in the corresponding model above
- $Year_i$ is a binary variable for the year group (1 for Year 2 and 0 for Year 1)
- $School_i$ is a vector of school-level 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 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 mean baseline EYFSP scores utilised in the primary analysis, 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)[22] between the treatment and control groups and these will be presented in the report.
- Count and % EVERFSM

[22] 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 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 logistic regression to test whether the 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 English as an Additional Language (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.

Compliance

We will estimate treatment effects across all four outcome measures for compilers using a Complier Average Causal Effect (CACE) analysis, using a pupil-level measure of compliance with the intervention. Compliance in this trial will be defined as having attended at least 16 of the 24 Speech Bubbles sessions. Attendance will be recorded by the drama practitioner and held centrally by the project team.

The CACE estimation will use a two-stage least squares (2SLS) approach[23]:

$$Comply_i = \gamma_0 + \gamma_1 Treat_i + \delta School_i + \kappa Year_i + \zeta X_i + \mu_i$$

$$Y_i = \beta_0 + \beta_1 \hat{Comply}_i + \theta School_i + \xi Year_i + \phi X_i + \epsilon_i$$

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

where:

- $Treat_i$ is a binary indicator for the treatment assignment (1 if the student is assigned to treatment and 0 if the student 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 school-level fixed effect
- $Year_i$ is a binary variable for the year group (1 for Year 2 and 0 for Year 1)
- X_i is a vector of baseline attainment measured through aggregated EYFSP learning scores for student *i,* as specified in the primary and secondary analyses above
- 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 the first equation
- Y_i is the raw PIRA score for student *i*

Intra-cluster correlations (ICCs)

We will estimate the ICC of the baseline and primary 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 from the primary analysis for pre-test ICC and PIRA scores for post-ICC;
- γ_i is the school-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 β_1 in the primary ITT analysis model;

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 using the above method, we will instead use the following approximation:

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

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 SAP 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 ITT analysis:

regress pira i.treat eyfsp_pira i.block, robust

regress busstory i.treat eyfsp_busstory i.block, robust

is a linear regression model estimated on individual-level full randomised sample data where *pira* is the Progress in Reading Assessment (PIRA) raw score and *busstory* the Renfrew Bus Story 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 for PIRA and *eyfsp_busstory* for Renfrew Bus Story (corresponding to Xin the regression equation), and *block* is a categorical stratification variable (corresponding to Year and School in the regression equation).

CACE analysis:

ivregress 2sls pira eyfsp_pira i.block (comply = treat), robust

ivregress 2sls busstory eyfsp_busstory i.block (comply = treaT), robust

is an instrumental variable (two stage least squares) regression model estimated on individual-level full randomised sample data where *pira* is the Progress in Reading Assessment (PIRA) raw score and *busstory* the Renfrew Bus Story 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 for PIRA and *eyfsp_busstory* for Renfrew Bus Story (corresponding to Xin the regression equation), and *block* is a categorical stratification variable (corresponding to Year and School in the regression equation).

Sub-group analysis:

regress pira i.treat i.EVERFSM_6_P treat#EVERFSM_6_P eyfsp_pira i.block, robust

regress busstory i.treat i.EVERFSM_6_P treat#EVERFSM_6_P eyfsp_pira i.block, robust

is a linear regression model estimated on individual-level full randomised sample data where 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).