Statistical Analysis Plan for Using Research Tools to Improve Language in the Early Years

BIT and NIESR

PROJECT TITLE	**Using Research Tools to Improve Language in Early Years**
DEVELOPER (INSTITUTION)	Oxford University, UCL and A+ Education
EVALUATOR (INSTITUTION)	Joint partnership between The Behavioural Insights Team (BIT) and the National Institute of Economic and Social Research (NIESR)
PRINCIPAL INVESTIGATOR(S)	Michael Sanders
TRIAL (CHIEF) STATISTICIAN	Michael Sanders
SAP AUTHOR(S)	Lucy Stokes, Daniel Carr, Jake Anders, Richard Dorsett and Michael Sanders
TRIAL REGISTRATION NUMBER	ISRCTN 18055918
EVALUATION PROTOCOL URL OR HYPERLINK	*https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/using-research-tools-to-improve-language-in-the-early-years/*

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	25/06/2018	Original

Changes to Protocol

A number of changes to the design occurred between the publication of the trial protocol and this statistical analysis plan. These are that randomisation occurred in two waves, and stratification was conducted on fewer variables than originally anticipated, and the Renfrew Bus Story has been removed as an outcome measure.

Table of contents

Introduction

This intervention programme aims to improve children's language, social and emotional outcomes through evidence-based professional development for nursery and reception teachers. The approach involves using research tools (the Environment Rating Scales and others) as a framework for self-evaluation and improvement. The study focuses on pupils aged between 3 and 5 years old, and aims to determine whether the intervention improves children's outcomes over the nursery and reception years of schooling.

The Environment Rating Scales (ERS) are highly regarded research tools used for assessing the quality of early years settings, and thought to predict children's development in the early years and beyond. Three ERS are used as part of this intervention: the Early Childhood Environment Rating Scale, Third Edition (ECERS-3), the Early Childhood Environment Rating Scale Curricular Extension (ECERS-E) and the Sustained Shared Thinking and Emotional Well-being (SSTEW) Scale. These scales provide a framework for observers, in this case, nursery and reception teachers, to assess elements of early years practice including language and reasoning, adult-child interactions, activities and care routines.

The core model that will be tested in this programme involves 5 days of training delivered to early years teachers over 5 months, with a sixth, follow-up, day 3 months later. Practitioners are taught the principles of using audit tools to improve practice in their settings, and how to support children's development through evidence-based practice. Teachers will also receive mentoring to help them apply the training in practice. The intervention was developed by a team from Oxford University, UCL Institute of Education, and A+ Education.

The evaluation was structured to be a two-armed randomised controlled trial involving at least 120 primary schools with nursery classes. Sixty schools were to be allocated to receive the intervention and 60 to a business as usual control group.

Recruitment began in April 2016 with the aim of starting the intervention with the cohort of children starting nursery in September 2016. The evaluation will look at the impact of the programme on language and social-behavioural development, as well as the impact on changes in practice among early years practitioners.

A number of changes to the design occurred between the publication of the trial protocol and this statistical analysis plan. These are that randomisation occurred in two waves, and stratification was conducted on fewer variables than originally anticipated, and the Renfrew Bus Story has been removed as an outcome measure.

Design overview

Trial type and number of arms	**Two-arm, cluster randomised**
Unit of randomisation	School
Stratification variables (if applicable)	Geographic area and proportion of FSM pupils

Primary outcome	variable	Language skills
	measure (instrument, scale)	Composite language skill score, based on: • British Picture Vocabulary Scale, • Renfrew Action Picture Test, • Clinical Evaluation of Language Fundamentals (CELF) Preschool 2 UK – Sentence Structure
Secondary outcome(s)	variable(s)	Each of the language measures used in the construction of the composite score Social-behavioural development Quality of provision for language and social development
	measure(s) (instrument, scale)	Language skill measures: • British Picture Vocabulary Scale • Renfrew Action Picture Test (both information and grammar scores) • CELF Preschool 2 UK – Sentence Structure Social-behavioural development: • Adaptive Social Behaviour Inventory Provision quality: • Composite environment rating scale score based on items from ECERS-3, ECERS-E and SSTEW

Study design

This is a cluster randomised controlled trial, with randomisation taking place at the school level. As the programme involves training teachers to improve practice in their classes, the choice was between randomising at school or class level. Randomisation at class-level would have entailed substantial risk of cross-contamination, especially as part of the programme involves teachers sharing practice with other staff. Furthermore, as the trial involves following children from nursery into their reception year, it would not have been practical to ask all schools to keep class groups the same when moving from nursery to reception, as school's choices around class allocation are necessarily driven by many other factors. The trial aimed to recruit 120 primary schools (ultimately 122 were recruited) with nursery and reception classes, 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 were offered the opportunity to take part in the programme following the completion of the study (August 2018), or a payment of £1,000, whichever they preferred.

The trial is being conducted across schools in the Liverpool, Manchester and West Midlands areas. These areas were chosen as they possess above average proportions of students from disadvantaged backgrounds (as measured by the proportion of neighbourhoods in the top 20 per cent of areas in the Index of Multiple Deprivation) and/or below average results at age 5 for communication and language (30% worst performing authorities for the proportion of children reaching expected level of development for communication and language in their EYFS profile).

The eligibility criteria for schools to participate were:

- Participating schools should be located in one of the study areas, and be a one or two form entry state primary school with a nursery class. Three or Four form entry schools were only accepted where they agreed to channel nursery children who have completed a baseline assessment into a reception class led by a participating teacher (defined as one nominated during the EOI process). This applies to both control and treatment schools. That is, in control group schools nursery children should move to a reception class led by a teacher who would have received the intervention had they been assigned to the treatment condition.
- One nursery and one to two reception teachers (with three reception teachers encouraged where the school is three form or more entry) agree to attend the ERS training and engage with mentoring if allocated to the treatment group;
- Schools should not have previously accessed training by A+ Education Ltd which is substantially similar to that being provided via the current intervention, received substantial support from their local authority using rating scales such as the Environment Rating Scales (ECERS and others) or used such tools themselves on a regular basis;
- If allocated to the control group, that schools continue with 'business as usual' for the duration of the trial (i.e. that they do not procure similar training that they otherwise would not have done);
- A completed Memorandum of Understanding;
- Consent to participate in the study – including the collection of outcome measures in summer 2018 – regardless of which trial arm they are assigned to;
- Agreement to collect opt-in consent from the parents of children involved in the study, and the provision of both school and pupil level data.
- Agreement 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
- Agreement that teachers in both trial arms complete a survey at the end of the trial period, and attend an interview with evaluation staff if requested.

Priority was given to schools with a higher proportion of FSM pupils.

Baseline outcome measure collection occurred over October and November 2016 when pupils were in nursery classes, and post-intervention outcome measure collection is scheduled for May to July 2018 when pupils are in their reception year. The programme runs over a five month period, with a follow-up day three months later. The decision to collect outcomes after 18 months was made due to the need to ensure sufficient time for changes in teachers' practices to become embedded and for the pupils to potentially benefit from these practices.

Given the young age of the pupils involved in the study, an opt-in consent process was used, with participants' parents (or legal guardians) making an informed decision regarding whether they consented to their child's participation in the assessments and data sharing. It is important to acknowledge that the use of opt-in consent may result in a threat to external validity; if consent was less likely to be obtained for certain groups of pupils. It is also possible that consent was more likely to be obtained in schools that are more effective generally. This should not be systematically different across the two arms of the trial. However, it may mean that the analysis is not generalisable to the full population of interest.

Randomisation

Randomisation followed recruitment of schools, including the signing of Memoranda of Understanding (MoUs) and baseline data collection in the majority of schools. Randomisation was stratified on the basis of school-level characteristics (proportion of FSM students and school location) to ensure balance between treatment and control groups (to be of equal or near-equal size). This was conducted using Stata. The randomisation followed a two-stage process:

1. The schools were stratified on the basis of FSM students (split across the median sample proportion) and location (split into West Midlands, Manchester, and Liverpool groups).

2. A random number was generated within each block and the subsamples split into two groups of equal size to ensure that school FSM proportion and location were balanced across trial arms. We used the Department for Education's Performance Tables to determine the blocking characteristics.

Randomisation ultimately occurred in two waves due to baseline measure collection and recruitment time constraints. The two waves were randomised as described above. A total of 122 schools were recruited, with 62 randomised to treatment and 60 to control. Of the 122 schools, 96 were randomised in the first batch (on the 25th November 2016) and 26 in the second batch (on the 31st November 2016).

Randomisation strata by batch

Batch 1			
	West Midlands	Manchester	Liverpool
Below median FSM	29	11	7
Above median FSM	25	8	16
Batch 2			
Below median FSM	7	3	5
Above median FSM	2	6	3

Calculation of sample size

Sample size calculations were based on the assumptions below.

- **Randomisation will be performed at the school-level.** This means that 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 and mentoring, which will impact on whole class attainment.

- **Number of children per cluster is 24.** This is an estimate of the average number of children in each class.

- **An intracluster correlation coefficient (ICC) of 0.20.** This defines how alike individual children are within each school (the cluster unit of randomisation). The ICC increases the more individuals within the clusters resemble one another. An ICC of 0.20 is commonly used in clustered randomised control trials in school settings. We note that this is higher than the ICC generally used in EEF trials; but should mean our estimate of the MDES is more conservative.

- **Power: 80%; Significance level: 5%.** These are standard assumptions.

- **The required minimum detectable effect size (MDES) is 0.22.** This specifies the minimum effect size our trial is powered to detect, in terms of a given standardised difference between two means (of a continuous outcome measure). If the effect of the intervention is below this amount, our trial may not be able to detect it.

At the protocol stage, we considered a pre- and post-test of our outcome measures, which was factored in to our original sample size calculations.[1] For the specific oral communication assessments used in this trial we do not have information on test-retest correlation over the length of time this study run for. However, based on test-retest correlation coefficients for language development studies of similarly aged groups over similar lengths of time, we assumed a test-retest correlation coefficient of 0.50 (Sibieta et al., 2016). On this basis, maintaining a MDES of 0.22 required 114 schools to be enrolled in this trial (no allowance for cluster-level attrition). With 15% attrition at the student-level, (effectively 20 students per cluster, rather than 24), 117 schools would be required. This was rounded to 120 to adjust for the possibility of attrition at the school level. However, due to issues with pre-testing, it was decided that the primary analysis would use post-test outcomes only (although controlling for school-level average pre-test scores).[2]

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.22	0.29	0.21	0.25
Pre-test/ post-test correlations	level 1 (pupil)				
	level 2 (class)				
	level 3 (school)	0.3	0.3	0.3	0.3
	level 2 (class)				

[2] These issues are documented later in the SAP.

Intracluster correlations (ICCs)	level 3 (school)	0.20	0.20	0.20	0.20
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		20	4	17	4
Number of schools	intervention	57	57	62	62
	control	57	57	60	60
	total	114	114	122	122
Number of pupils	intervention	1140	228	1,069	248
	control	1140	228	1,042	240
	total	2280	456	2,111	488

Recruitment update

The original recruitment target was ultimately overshot, as fear of schools withdrawing at the last minute or failing to provide data needed prior to randomisation meant that the trial partner over recruited as planned with the EEF.

However, the number of observations per cluster was ultimately lower than anticipated.

The lower number of children assessed per school partly reflects the number of children for whom it was possible to obtain parental consent, due to the following reasons:

- Some schools recruited into the trial only had a small number of children (eleven schools had fewer than ten children)

- Some schools in the trial were unable to gain consent from all (or nearly all) parents, and some showed little willingness to assist in engaging parents with a view to increasing the number of parents within participating schools giving consent.

It also reflects the fact that it was not possible to assess all children for whom parental consent was granted, for the following reasons:

- Several schools in the trial had a high number of children from migrant backgrounds with low levels of proficiency in English. This meant these pupils were unable to engage with the test – hence this left a smaller number of children in these schools to be assessed.

- Some children were absent across multiple days, and therefore not present at the time of the assessments.

- Some children did not consent to participate in the assessments at the time of testing.

- Some children had learning and/or physical impairments that meant they were unable to participate in the assessments.

Furthermore, for 24 schools, full sets of opt-in consent forms were not received by the evaluation team, although all schools confirmed that they had distributed and received the forms. Six of these schools subsequently returned full sets of forms and one additional school has withdrawn from the trial. At the time of writing, forms were still being sought from the remaining schools.

Based on the 114 schools who have currently returned full sets of consent forms, and assuming an average of 15 children per school (which factors in attrition of 10% at endline data collection), and an ICC of 0.20, the MDES stands at 0.25 using only endline data and controls for the school level average at baseline data collection with an assumed school-average-pre-test to individual post test correlation of 0.35. Using a lower ICC equal to 0.13, the MDES would be 0.22.

Outcome measures

Primary outcome

The primary outcome measure will be a composite language skill score. This measure will draw on the results reported in three language assessments:

- **British Picture Vocabulary Scale (BPVS):** A one-to-one test that assesses a child's receptive vocabulary. For each question, the test administrator says a word and the child responds by selecting a picture from four options that best illustrates the word's meaning.
- **Renfrew Action Picture Test (APT):** In this test, the child is asked to describe the actions shown in a set of pictures. Two scores are recorded, one for the level of information they provide (for example nouns and verbs) and one for the grammar they use (such as use of tenses). Both information and grammar scores will be incorporated into the composite measure.
- **Clinical Evaluation of Language Fundamentals (CELF) Preschool 2 UK - Sentence Structure:** This subtest provides information about how a child understands spoken language. This is achieved by asking the child to interpret spoken sentences of increasing length and complexity by pointing to the picture that illustrates a given sentence.

These three tests were chosen following extensive discussion with the project team, taking into account the most appropriate measures for the aspects of language development targeted by the intervention, but also bearing in mind practical considerations, such as length of assessments. At the pre-test, the language assessments took a total of between 15 and 20 minutes to administer. Each test captures a different dimension of language development, covering vocabulary, comprehension and expressive language. It was decided that the primary outcome should be a composite score based on these different measures, as a priori the intervention could potentially affect each of these dimensions. Although some of the component measures are different, this approach is broadly consistent with that adopted in the evaluation of the Nuffield Early Language Intervention (Sibieta et al., 2016).

To arrive at a composite language skill score we standardised each of the components (including both APT scores) to have a mean of zero and standard deviation of one. These were added together to create composite measures and re-standardised. As such, the four[3] language measure scores are equally weighted in the composite language skill score. In addition, we will also explore using factor analysis as a sensitivity analysis.

The factor analysis posits that there is a latent factor describing language skills, with our four (counting the two APT scores separately) observed tests scores representing manifest measures of this underlying construct. Unlike the composite score proposed above, this allows the four measures to load on the common latent factor to differing degrees. We will estimate the loadings of the factor on the four measures using an exploratory factor analysis principal factor approach, constraining there to be a single retained factor.[4] From this model, we will predict values of the language skills latent factor using the regression scoring measure.[5]

All three tests are used at both pre- and post-test. All tests (both pre and post) were administered on a 1:1 basis and scored by research assistants with an academic background in speech therapy or psychology. Recruited by BIT, the research assistants (RAs) were trained by an experienced language development psychologist in how to use the language assessments prior to visiting schools. RAs were blind to the trial arm allocation of schools they visit. Tests are conducted at two intervals during the course of the trial:

- **Pre-test:** this was conducted between 5 October – and 12 December 2016. The great majority occurred prior to schools being informed of their trial arm assignment, but some follow-up visits occurred one to ten days after randomisation was communicated to include children who were absent when RAs first visited.

- **Post-test:** this is being undertaken during May - July 2018.

Secondary outcomes

Individual scores for each of the above language measures will also be reported as secondary outcome measures.

An additional secondary outcome is social-behavioural development as measured by the Adaptive Social Behavior Inventory (ASBI). The ASBI is a questionnaire that is being completed by class teachers for each student, at the same time point as primary outcome measures are collected (at both pre and post-test). We will report scores for the three subscales, Express (13 items), Disrupt (7 items) and Comply (10 items), as well as the total score (min. score=30, max score=90).

As trial arm allocation was only revealed after pre-test outcome measures were collected, the pre-test ASBI scores are blind to trial arm assignment, despite being collected by classroom teachers. Post-test ASBI scores will not be blind to trial arm assignment. This is unavoidable given the need for a teacher familiar with the student to complete the questionnaire. The ASBI

[3] As this includes the two separate APT measures, along with the scores from the BPVS and the CELF Sentence Structure sub-test.

[4] This method can not be considered a reliable approach if the first factor is not sufficiently strong.

[5] This analysis will be carried out using Stata's 'factor' command as follows:

- factor BVPS APT1 APT2 CELF
- predict factor_languageskills

was chosen as it has been widely used as a measure of social-behavioural development and was relatively straightforward and practical to administer.

The final secondary outcome to be considered as part of this study is the quality of the provision for language and social development, as measured by a composite Environment Rating Scale (ERS) score based on items from the ECERS-3, ECERS-E and SSTEW. The items from each scale to be included in this composite measure are detailed in Appendix 1 of the Trial Protocol. For transparency, total scores for the ECERS-3 and SSTEW, alongside the literacy subscale of the ECERS-E, will also be reported in our evaluation. All scores will use the standard scoring approach of summing question scores and dividing by the number of questions (min. score =1, max score =7). ERS scores for reception classes were collected by A+ Education staff prior to randomisation of schools to trial arms, and will be collected again in the autumn term 2017. The observers will be blind to trial arm assignment.

The trial protocol indicates that the Renfrew Bus Story Test would be used as an additional secondary analysis. Due to the funding implications of running this test, it was decided not to administer this test. Instead, the mean utterance length from the Renfrew Action Picture Test (APT) will be used in its place.

Analysis

Primary outcome analysis

Our primary analysis will focus on the composite language skill score, and will be performed using Stata (version 14). Using the composite score of four language development measures (the APT provides two separate measures) will allow for a more holistic measure of language development, covering comprehension, vocabulary and expressive language.

Outcome variables will be regressed using a least squares linear model with treatment arm indicators, strata indicators (i.e. whether the school was above or below the median FSM proportion, school location, plus whether the school was randomised as part of the first or second batch). Due to the issues with the collection of pre-test data discussed above, the primary analysis will not include the pre-test composite language skill score. To account for the experimental design, standard errors will be clustered at the school level to allow for correlation of pupil outcomes within schools.

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.

$$Y_{ijt} = \alpha + \beta_1 Treat_j + +\beta_2 Y_{jt-1} + \beta_3\, \gamma_j + \varepsilon_{ijt}$$

where i are individuals and j are schools, Y_{ij} is our composite language skill score, $Y_{jt-1}$is the school average pre-test score, $Treat$ is our school-level treatment indicator, γ_j being a vector of stratification variables, and ε being 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 (γ_j).

Secondary outcome analyses

We will repeat the primary analysis but replace Y_{ijt} being the composite language measure with, separately, each of the four separate measures. As for some pupils not all four language measures may be available, we will check the sensitivity of the results to basing the analyses on a consistent sample for which all four language measures are available, as well as allowing the sample to vary by measure to incorporate all pupils for whom each outcome is available. Results based on the consistent sample will be considered as the main source of evidence.

The same approach will also be adopted for the analysis of ASBI scores, which form an additional secondary outcome. Effectively the same approach will be used for analysis of impact on the quality of provision as measured by the composite measure described above, here the models will control for the quality of provision as measured at baseline.

The exploration of a number of outcomes can raise concerns around multiple comparisons. Our primary outcome is clearly defined in both the trial protocol and this analysis plan as our composite language score, with all other outcomes considered secondary analysis. However, given the number of secondary outcomes we will adjust for the fact that we are undertaking multiple comparisons by applying the Benjamini-Hochberg procedure.

Interim analyses

No interim analyses are planned.

Subgroup analyses

We will also conduct the analysis for the following subgroups of pupils, using the same model as our primary analysis:

1. Those who are registered for free school meals (FSM) in the National Pupil Database (using the variable EVERFSM_6_P, following EEF guidance);
2. Those who are marked as English as an Additional Language (EAL) by their schools;
3. Those with language difficulties, as defined as those who score in the bottom 15 percent of BPVS age-standardised scores (in the “extremely” or “moderately” low score range) during pre-test. This is equivalent to a score one standard deviation below the mean of the normed population.
4. The analysis will also be conducted separately for boys and girls.

These subgroups were identified in the trial protocol. FSM pupils are clearly a key subgroup to be analysed in all EEF trials. As the primary outcome to be measured is improvement in children’s language, there is particular interest in whether the programme has differential effects for those children for whom English is an additional language, and also those identified as having language difficulties at the point of baseline data collection. It is also relevant to consider differences by gender, given considerable interest in differences in attainment by gender, especially in terms of language outcomes.

The subgroup analyses will be conducted for both the primary and secondary outcomes. To test whether there are differences for all the above subgroups other than FSM pupils, interaction terms will be incorporated into the models. For FSM pupils, analysis will be run separately for this subgroup, in line with EEF analysis guidance.

Additional analyses

Given the proportion of pupils for whom it was not possible to obtain a pre-test, we will also conduct additional robustness analyses as follows:

- Including pre-test scores in the analysis, but restricting the analysis sample to only those pupils for whom both pre-test and post-test scores are available
- For the full sample for whom post-test scores are available, additionally including pre-test scores where available, and for those where the pre-test is missing, imputing a score for the pre-test using multiple imputation (see Missing Data Section).
- In addition, we will check the sensitivity of the results to both the inclusion and exclusion of those pupils for whom some but not all four language measures are available.

In order to explore the question of whether any effect of the programme is working through changes in the learning environment (as proxied by composite ERS scores), as is hypothesised by the project's logic model, we will carry out additional analysis to explore the extent to which change in ERS scores mediate the treatment effect. We stress that this analysis is exploratory in nature.

This will be carried out using the following regression model:

$$Y_{ijt} = \alpha + \beta Treat_j + \beta_2 Y_{ijt-1} + \beta'_3 \gamma_j + \beta_4 \Delta ERS_j + \ \beta_5(\Delta ERS_j * Treat_j) + \varepsilon_{ijt}$$

where ΔERS is the change in composite ERS scores between those collected by the project team before randomisation ERS_{jt-1} and then collected contemporaneously with outcome measures ERS_{jt}, as appropriate. Our primary parameters of interest in this model will be as follows: $\beta_4$will report the estimated change in outcome measure unexplained by the change in ERS score in treatment schools; $\beta_5$will report the estimated change in outcome measure associated with the change in the ERS score in treatment schools. To explore whether it is change in particular aspects of the ERS that are associated with the change in outcome measure, we will also conduct analyses that replace the composite ERS score with its component subscales.

Imbalance at baseline

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

- pre-test composite language score (including its subscales),
- proportion female,
- proportion ever eligible for Free School Meals,
- institutions Ofsted ratings
- proportion for whom English is an Additional Language, and
- Age in months

We will do this by calculating absolute standardised differences (Imbens & Rubin, 2015) between the treatment and control groups and these will be presented in the report. In line with EEF reporting guidelines, differences in the pre-test measures will be reported as effect sizes. The interpretation of these will need to bear in mind the issues discussed earlier regarding the collection of the pre-tests, however, there is no reason to suspect that these should differ by trial arm.

Missing data

We will report the distribution of missing observations by treatment arm. In the event of greater than 5% missing data at either cluster or individual level or a significant difference in missingness between treatment and control arms we will conduct further investigation into the mechanisms of missingness. We will include an assessment of missing data at both the school and pupil level, and will investigate the extent to which baseline characteristics (at school and pupil level) are correlated with non-response, using linear regression and the same set of variables as detailed above in our balance checks. As a sensitivity analysis, we intend to undertake multiple imputation of baseline data stratified on treatment condition as an additional robustness check. The model for imputation will control for participant school, gender and free school meals status, and we will use chaining to use baseline scores that were collected where we have partial cases.

The extent of missingness in terms of baseline outcomes is already known. As mentioned earlier in this document, some pupils for whom parental consent was granted did not complete the pre-test (around 21 per cent of all pupils for whom consent was obtained). For this reason, the primary analysis will use post-test outcomes only. We will explore the sensitivity of our results to alternative approaches, discussed below in the section on secondary analysis.

As a number of schools (23 of 122) have not returned ASBI score sheets for children (at the pre-test) we will explore the potential to impute these using other characteristics of children and schools. As ASBI forms were completed prior to randomisation we would not expect there to be a relationship between trial arm and likelihood of return. Indeed, the number of schools who returned ASBI forms is similar across trial arms (49 returned in Control group, 50 for Treatment). We propose to undertake multiple imputation stratified on treatment condition as an additional robustness check. The model for imputation will control for participant school, gender and free school meals status, and we will use chaining to use baseline scores that were collected where we have partial cases.

Compliance

The trial protocol specified a set of criteria identifying the minimum level of engagement required in order for the intervention to be considered to be taking place. This draws principally on school and teacher engagement scores and attendance data. As the intervention is focused on the Early Years Foundation Stage (EYFS) phase, it considers data from both the nursery teacher, reception teacher and the phase as a whole.

The following approach, agreed between the evaluation and delivery team, is to be used to assess compliance for analysis of the primary outcome:

1) Did the pupil have a nursery teacher who attended over 3 training sessions?
2) Did the pupil have a reception teacher who attended over 3 training sessions?
3) The school’s engagement score, with a score of 1 or 2 considered to be complying with the intervention, and a score of 3 or 4 considered not to be complying. This will be assessed by the mentors (part of the intervention team) who work with the schools and teachers.

Engaging with the intervention means that schools have:
(a) used at least some elements of the URLEY tools and/or materials (e.g. the ERS, TROLL, the Language Learning Principles, the interaction audit or other tools, the action planning process);
(b) attempted to implement changes within their classroom/s, even though they may have faced challenges in doing this;
(c) made some attempt to introduce new staff to the approach, where staffing has changed.

Where there are differences between teachers/classes within a school (e.g. one has engaged while others have not) then mentors are asked to make an overall judgement regarding participating schools as a whole, according to the following scale:
1 = good, consistent engagement
2 = reasonable engagement (or, where this has been mixed, half or more of participating teachers have engaged)
3 = some engagement (or, where this has been mixed, fewer than half of participating teachers have engaged)
4 = little or no engagement

The three indicators set out above (the two measures of training attendance and the engagement score) will be combined into a single indicator of compliance (with compliance defined as occurring when all three of the criteria are met). This is the binary measure of compliance that we will use in the Complier Average Causal Effect analysis described below.

This approach to assessing compliance is aligned with the activities and inputs set out in the original logic model for the intervention, which are attendance at training, mentoring and access to online resources. Of these, the first two are considered to be the key activities, although accessing online resources is partly captured through the engagement score. The engagement score is also considered to be preferable to including a measure based on hours of mentoring received (as a greater number of mentoring hours may reflect schools that required more support, rather than greater engagement).

We will use Complier Average Causal Effect (CACE)[6] 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, using the binary measure of compliance described above. The predicted values from the first stage are then used in the estimation of a model of our outcome measure Y_{ijt}. In other respects, the specification remains the same as the primary outcome ITT model. 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.

In addition, we will also conduct a similar analysis for the ERS measures. Here the compliance criteria will be defined by fulfilling the next three conditions as follows:

1) Whether the teacher (of the class that is observed at post-test) had attended over 3 training sessions (i.e. 4 or more)

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

2) The school's[7] engagement score, (receiving a score of 1 or 2 considered to be complying with the intervention)
3) Whether this teacher was also the teacher observed at the pre-test

Intra-cluster correlations (ICCs)

We will estimate the school-level ICCs for pre-tests and post-tests using empty hierarchical linear models including school-level random effects as follows:

$$\boldsymbol{Y_{ij} = \beta_0 + \eta_j + \varepsilon_{ij}}$$

where $\boldsymbol{Y_{it}}$ is the pre- or post-test of individual $\boldsymbol{i}$ in school $\boldsymbol{j}$, $\boldsymbol{\beta_0}$ is a constant term, $\boldsymbol{\eta_j}$ is a school-level random effect and $\boldsymbol{\varepsilon_{ij}}$ is an individual-level idiosyncratic error term. The ICC estimate is recovered as follows:

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

Effect size calculation

Effect sizes will be calculated in line with the EEF's analysis policy for cluster randomised trials i.e. estimating Hedges' g using total variance (rather than within cluster variance) to maximise comparability with other trials. This will require estimates of:

- the unstandardised conditional treatment effect (β_1) from the primary ITT analysis regression model reported above;
- the unconditional total standard deviation of the outcome variable for the analysis sample (s_t). (Note that $s_t$is a combination of variance within schools –s_w–and between schools – s_b– although there is no need to decompose given that estimation is not carried out using a hierarchical model.)

Hedges' g is calculated as follows:

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

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);

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

[7] Note that as only one class per school is observed this is effectively a class level variable

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

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