Efficacy Trial of Early Talk Boost
Statistical Analysis Plan
Evaluator (institution): NFER
Principal investigator(s): Dr Stephen Welbourne

Education Endowment Foundation

PROJECT TITLE	Efficacy Trial of Early Talk Boost
DEVELOPER (INSTITUTION)	Speech and Language UK
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Dr Stephen Welbourne
SAP AUTHOR(S)	Dr Jose Liht, Chris Morton, Palak Roy and Dr Stephen Welbourne
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at setting level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	3-4-year-old, Early Years
NUMBER OF SCHOOLS	224
NUMBER OF CHILDREN	2,656
PRIMARY OUTCOME MEASURE AND SOURCE	Oral language skills (latent variable formed from expressive and receptive subtests in LanguageScreen and the Renfrew Action Picture Test (RAPT))
SECONDARY OUTCOME MEASURES AND SOURCES	Individual expressive and receptive subtests of LanguageScreen and the RAPT (for 1 to 6) 1.Expressive Vocabulary, LanguageScreen Expressive Vocabulary 2.Listening Comprehension, LanguageScreen Listening Comprehension 3.Receptive Vocabulary, LanguageScreen Receptive Vocabulary 4.Sentence Repetition, LanguageScreen Sentence Repetition 5.Information (Communication skills), RAPT information

	6.Grammar, RAPT grammar 7.Communicative Participation, Focus 34 8.Practitioners' knowledge and confidence, lead practitioner survey

SAP version history

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

Table of contents

Introduction

Language is foundational to all education, acting as the gateway to literacy, science, mathematics, history, and all other subjects which are predominantly conveyed through the medium of language. Neuroscientific research indicates that the neural mechanisms that underpin reading largely overlap with those that support language (Welbourne *et al.*, 2011; Woollams, Halai and Lambon Ralph, 2018). In English (along with other European languages) the phonological representations crucial for spoken language comprehension and production also play a central role in decoding written words. Language underpins children's learning across the curriculum, and strong early phonological and oral-language skills are known to predict later literacy success (Liberman, Shankweiler and Liberman, 1989) (Hulme *et al.*, 2024). Research shows that early interventions can positively influence children's linguistic development (Spinner and Aoife, 2025). A recent meta-analysis focusing on early years language and communication programmes has underscored the potential for substantial positive impact on children (EEF, 2023), yet in the UK the evidence base remains limited, with the Nuffield Early Language Intervention (NELI) targeted intervention being one notable exception (EEF, 2025). Given the substantial time and resource demands of programmes like NELI, there is growing interest in developing lighter-touch interventions that can support children at risk of language delay without significantly disrupting the wider curriculum.

Early Talk Boost (ETB) is one such programme, aimed at 3- to 4-year-olds with delayed speech and language. Delivered over nine weeks in small groups, ETB includes practitioner training, structured story- and song-based sessions, and a parent workshop. A tracker helps identify suitable children, particularly those with "amber" developmental scores. A small scale randomised controlled trial found that ETB improved expressive and receptive language (Reeves *et al.*, 2018), although methodological limitations mean the evidence is graded as Level 2 by the Early Intervention Foundation meaning it is not possible to make causal inferences based on this data alone (Early Intervention Foundation, 2019).

A more recent efficacy trial commissioned by the Education Endowment Foundation (EEF) in 2023 attempted to evaluate ETB at scale but encountered substantial implementation challenges, including high attrition, screening issues, and data-collection delays. These difficulties ultimately led to the trial being cancelled in early 2024. The experience highlighted the need for clearer eligibility criteria, simpler assessment processes, and improved logistical coordination in any future evaluation of ETB. The current trial seeks to improve upon these attempts and provide more robust evidence for the efficacy of ETB.

Design overview

Trial design, including number of arms	Two-arm, cluster randomised
Unit of randomisation	Setting
Stratification variables	Setting type (maintained and Private, Voluntary, and Independent (PVI)) and setting size (settings with fewer than 20 3- to 4-year-olds vs those with 20 or more 3- to 4-year-olds)

Primary outcome	variable	Oral language skills
	measure (instrument, scale, source)	Latent oral language variable formed from expressive and receptive subtests in LanguageScreen and the Renfrew Action Picture Test (RAPT)
Secondary outcome(s)	variable(s)	1. Vocabulary knowledge 2. Literal and inferential language comprehension and expressive language skills 3. Vocabulary understanding 4. Language comprehension and production 5. Information 6. Grammar 7. Communicative participation 8. Practitioners' knowledge and confidence
	measure(s) (instrument, scale, source)	1. LanguageScreen Expressive Vocabulary subtest (score range 0-24) 2. LanguageScreen Listening Comprehension subtest (score range 65-135; age standardised) 3. LanguageScreen Receptive Vocabulary subtest (score range 0-23) 4. LanguageScreen Sentence Repetition subtest (score range 0-14) 5. RAPT information subtest (score range 0-41) 6. RAPT grammar subtest (score range 0-39) 7. FOCUS-34 Communicative participation 8. Leader practitioner survey, Practitioners' Knowledge and confidence
Baseline for primary outcome	**variable**	Oral Language Skills
	measure (instrument, scale, source)	Latent oral language variable formed from expressive and receptive subtests in LanguageScreen
Baseline for secondary outcome	**variable**	1. Vocabulary knowledge 2. Literal and inferential language comprehension and expressive language skills 3. Vocabulary understanding 4. Language comprehension and production 5. Oral language skills (baseline for primary outcome)

		6. Oral language skills (baseline for primary outcome) 7. Oral language skills (baseline for primary outcome) 8. Practitioners' Knowledge and confidence
	measure (instrument, scale, source)	1. LanguageScreen Expressive Vocabulary subtest (score range 0-24) 2. LanguageScreen Listening Comprehension subtest (score range 65-135) 3. LanguageScreen Receptive Vocabulary subtest (score range 0-23) 4. LanguageScreen Sentence Repetition subtest (score range 0-14) 5. Latent oral language variable formed from expressive and receptive subtests in LanguageScreen 6. Latent oral language variable formed from expressive and receptive subtests in LanguageScreen 7. Oral language skills (baseline for primary outcome) 8. Practitioners' Knowledge and confidence obtained from the Leader practitioner survey,

This efficacy trial, funded by EEF, aims to build on the findings from the earlier trial by testing whether ETB is effective at improving children's oral language skills when delivered on a larger scale across both maintained nursery settings and Private, Voluntary, and Independent (PVI) settings.

We employed a setting-randomised design. We randomised 224 Early Years settings on a 1:1 basis into two groups, intervention and control, stratified by setting type (maintained nurseries or PVI settings) and setting size (less than 20 eligible children vs 20 or more eligible children) to ensure equal group allocation in each stratum. Intervention settings will deliver ETB to up to four groups of children per class/room, with a maximum of 32 children per setting (i.e., two groups of eight children in each of two classes or four groups of six to eight children across the entire setting).

All 3–4-year-old children (as of August 31, 2025), who did not have Special Educational Needs and Disabilities (SEND) diagnosis and were not awaiting SEND diagnoses were eligible to take part in the trial providing they attended the setting for at least three different days (or half-days) every week during 2025-26 academic year. These children were assessed by settings using LanguageScreen prior to selection of children to receive ETB. LanguageScreen is a baseline for the primary outcome[1], and also act as a screening test for the trial. NFER selected children using the LanguageScreen traffic light system, with priority given to those in the amber band, as they

[1] Although we are not basing our analysis on change scores from baseline to endpoint, and are instead using the observations made before the intervention takes place as covariates, we refer to measurements made before randomisation to the treatment conditions as baseline measurements.

are most likely to benefit from the ETB intervention. If more children were needed to achieve the sample size requirements, selection extended to those near the amber threshold from the red and green bands, forming the final trial cohort for analysis (see child selection section below for further details).

The primary outcome for this trial is an oral language skills latent variable based on the endline subtests of the LanguageScreen and Renfrew Action Picture Test (RAPT). LanguageScreen measures core receptive and expressive language skills through structured tasks assessing vocabulary, listening comprehension, and sentence repetition. RAPT, in contrast, captures children's ability to convey information and use grammatical structures within narrative descriptions. Combining LanguageScreen and RAPT into a latent outcome therefore provides a more sensitive, comprehensive and robust measure of oral language by integrating complementary domains and maximising the variance captured across skills.

The secondary outcomes for children in the trial are subtests of LanguageScreen and RAPT: Expressive Vocabulary, Listening Comprehension, Receptive Vocabulary, Sentence Repetition, Information (Communication skills) and Grammar, and Communicative Participation from Focus-34. It assesses the effectiveness with which children communicate in life situations where knowledge, information, ideas or feelings are externalised and includes, for example, behaviours relating to intelligibility, communicating independently with unfamiliar adults, turn-taking and willingness to talk (Thomas-Stonell et al., 2013; Thomas-Stonell et al., 2013). We are also exploring the intervention's effect on practitioner self-reported knowledge and confidence scale, as measured by self-reported questions in the lead practitioner survey at baseline and endline.

The Implementation and Process Evaluation (IPE) will investigate the extent to which the programme is delivered with fidelity under ideal conditions. It will particularly focus on differences between PVI and maintained settings, programme benefits for disadvantaged children, securing and maintaining child engagement in ETB sessions, whether sessions were delivered as intended, and impact on practitioner outcomes.

Research questions

Research Question
RQ1 (Primary Research Question): How effective is ETB at improving oral language skills of 3 to 4-year-old children with low to medium[2] prior oral language skills?
RQ2: How effective is ETB at improving different aspects of oral language skills of 3–4-year-old children with low to medium prior oral language skills as measured by the subtests of LanguageScreen and RAPT?
RQ3: How effective is ETB at improving Communicative Participation of 3–4-year-old children (measured by FOCUS-34) with low to medium prior oral language skills?
RQ4: How does the effectiveness of ETB vary for PVI settings compared to maintained settings?
RQ5: How effective is ETB at improving the oral language skills of 3–4-year-old disadvantaged children (children eligible for Early Years Pupil Premium, EYPP) with low to medium prior oral language skills?
RQ6: What is the differential impact of ETB at improving the oral language skills of 3–4-year-old children based on a traffic light classification of children's prior oral language skills as measured by LanguageScreen at baseline?
RQ7: How effective is ETB at improving practitioners' knowledge and confidence in teaching language skills?
RQ8: Is the effectiveness of ETB mediated by practitioners' knowledge and confidence?
RQ9: Is the effectiveness of ETB mediated by children's communicative participation[3] (as measured by FOCUS-34)?

Randomisation

This trial is employing a setting-randomised design. 224 settings who returned requested child data and completed baseline LanguageScreen assessments were randomised into two arms, intervention and control, on a 1:1 basis. Randomisation was stratified by setting type (maintained nurseries or PVI settings) and setting size (settings with fewer than 20 3- to 4-year-olds vs those with 20 or more 3- to 4-year-olds) to ensure equal group allocation in each stratum. Randomisation was conducted by the trial statistician on 1st November 2025. . Settings were notified shortly thereafter. This resulted in 112 settings assigned to each arm with 139 maintained vs 85 PVI settings and 114 settings with less than 20 3- to 4-year-olds vs 110 settings with 20 or more 3- to 4-year-olds. See Table 1 below for more information.

[2] The wording of the Research Questions was revised during the peer review of the Analysis Plan. Further details can be found in the *Child Selection for ETB* section.

[3] Please note that the outcome measure in this document refers to children's attention and social skills. The label has been adjusted here to maintain consistency with other research questions, including those related to IPE. This relabelling does not reflect any change in how the outcome is measured.

Table 1: Randomisation Results

	PVI, <20 children	PVI, 20+ children	Maintained, <20 children	Maintained, 20+ children	Total
Control	30	12	27	43	112
Intervention	30	13	27	42	112

Please see the randomisation code in Appendix A.

Child selection for ETB

At baseline the LanguageScreen traffic light system was used as a tool to select children for ETB. The traffic light system categorises children into three bands—red, amber, and green—based on their standard scores. Children with standard scores of 90 or above are classed as green (no cause for concern), children with scores in the range 82 - 89 are classed as amber (may benefit from additional language support), and children with scores of 81 or below are classed as red (definitely need additional language support). The ETB selection prioritised all children identified within the amber band, as they represent the target group thought most likely to benefit from the ETB intervention. This is akin to selection for ETB outside the trial where children with scores falling in the amber category of ETB tracker are considered to be most suitable for the ETB intervention. RQ6 will test the validity of this assumption.

As additional children were required to be selected beyond the amber group, further selection was made from the children whose scores were closest to the amber threshold. This included those at the upper end of the red band and the lower end of the green band, ensuring that the extended sample remained as closely aligned as possible with the intended ETB target population. Ties were broken at random: for example, if two children had scores of 91 and 80, the selection priority order amongst these two children was decided at random. The resulting sample formed the trial cohort and final analysis sample for both intervention and control settings.

Of the 4,497 children who were assessed at baseline, 457 were categorised as red, 606 amber and 3434 green on the LanguageScreen traffic light system. After the selection procedure described above, 2,656 children were selected to participate in the trial: 312 were categorised as red, 606 as amber and 1,738 as green. The selection procedure therefore resulted in 100% of amber category children, 68% of red category children and 51% of green category children being selected. This is largely in line with what we expected when writing the study protocol, although we had anticipated a lower proportion of green category children to be selected (40% rather than 51%). As expected, the majority of red category children were selected for ETB and therefore we can be reasonably confident that we can detect effects for children with lower baseline language skills. However, as the sample now includes a higher proportion of children of green category, describing the sample exclusively as “low prior oral language skills” is no longer fully accurate. To reflect this selection of children spanning the lower to mid-range prior language skills, the wordings in the research questions are revised to refer to them as “low to medium prior oral language skills”.

During the set-up stage, it was agreed between NFER, EEF and Speech and Language UK (delivery team) that each intervention setting would be asked to run up to four groups of ETB sessions, comprised of between six to eight children. The exact number and size of groups

depended on the number of eligible children who were nominated to complete baseline LanguageScreen (see Table 2 in the study protocol;[4] Roy *et al.*, 2022).

In cases where settings had multiple classes or rooms, two different approaches could be taken in consultation with the setting. In the first approach, settings nominated which classes would take part in the trial (with a maximum of 70 eligible children across all nominated classes). Each class was then treated separately, with either one or two ETB groups selected per class. Alternatively, the setting could elect to pool all their eligible children in which case the number of children and groups selected was determined by the total number of eligible children in the setting. In practice almost all settings opted for this second approach. The majority of settings selected either one group of six, one group of seven, and two groups of six children. The resulting number of children selected at each setting is shown below in Table 2. For one very large setting with more than 70 eligible children who wished to pool their children, we first randomly selected 70 of the eligible children. Only these children completed baseline LanguageScreen in this setting and therefore only these children were taken forward to the selection process described above.

Due to the timeline on which the baseline LanguageScreen data became available, it was necessary to select the final trial children shortly after randomisation, rather than pre-randomisation. To ensure against any bias that might result from selecting children with a known group allocation, the selection was performed by a statistician who was blinded to each settings' randomised allocation. The one exception is the control setting described in the next paragraph, where it was not possible to preserve blinding for the selection.

In one control setting with 22 completed LanguageScreen baselines, 12 children were accidentally given the Australian version of LanguageScreen, rather than the British version. Only a minority of items were different between the two versions. We reviewed the affected items and decided five items (three from Expressive vocabulary, two from Receptive vocabulary) were sufficiently different that it would be more accurate to impute scores for these items. The 12 children had their scores imputed for the five items using predictive mean matching (one imputation only), where age in months and every individual LanguageScreen item score were used in the imputations. Total raw and standard scores were then recalculated for these pupils and the recalculated scores were used for trial selection and as a baseline in modelling.

[4] Efficacy Trial of Early Talk Boost

Table 2: Children groups allocated for the trial

	Number of classes[5]		
Selected for Trial/class	**Intervention**	**Control**	Total
4 (one group of 4)	0	1	1
5 (one group of 5)	1	0	1
6 (one group of 6)	28	25	53
7 (one group of 7)	31	33	64
12 (two groups of 6)	15	8	23
13 (two groups of 6 or 7)	4	5	9
14 (two groups of between 6 and 8 each)	4	5	9
15 (two groups of 7 or 8)	5	6	11
16 (two groups of 8)	6	9	15
18 (three groups of 6)	3	2	5
20 (three groups of between 6 and 8)	4	5	9
21 (three groups of between 6 and 8)	1	3	4
22 (three groups of between 6 and 8)	0	1	1
23 (three groups of 7 or 8)	2	1	3
24 (three groups of 8)	4	4	8
25 (four groups of 6 or 7)	1	2	3
26 (four groups of between 6 and 8)	1	2	3
27 (four groups of between 6 and 8)	2	1	3
28 (four groups of between 6 and 8)	1	1	2
32 (four groups of 8)	1	0	1
Total	114	114	228

[5] Most settings chose to pool all their children for the purpose of selection, so are counted as one 'class' in this table. Four settings instead chose to perform selection separately in two classes, so these settings are counted twice in the table, once for each class.

Outcome measures

Primary outcome

The primary outcome for this trial will be oral language skills (including both expressive and receptive language skills). It will be measured using LanguageScreen and RAPT.

LanguageScreen

LanguageScreen is an online language assessment tool which can be accessed in a web browser via any desktop, laptop computer, Chromebooks with an internet connection. It can also be used on iPads or Android tablets, if preferred. It assesses language-related abilities through structured tasks or assessments that require children to answer questions about pictures or words/stories. It is designed to evaluate oral language skills in children aged 3-11. It is administered one-to-one with children and takes approximately five to ten minutes to complete. It is comprised of the following four subtests:

1. Expressive Vocabulary (24 items; raw score range 0-24) assesses the ability to name pictures.
2. Listening Comprehension (4-12 items; standard score range 65-135) assesses children's ability to understand spoken language by asking questions about short stories being played to them. The number of stories and questions varies by age: children younger than 4.5 are played one story and are asked four questions (raw score range: 0-4), while children aged 4.5 and above are played two stories and are asked 12 questions (raw score range: 0-12). There is no overlap between the stories offered to the two different age groups. As scores are on different scales depending on age and there will be children on either side of the 4.5 age threshold, the standard score will be used in all analyses instead of the raw score.
3. Receptive vocabulary (23 items; raw score range 0-23) assesses the ability to match spoken words to pictures by asking them to match a word they hear to one of four pictures on the screen.
4. Sentence Repetition (14 items; raw score range 0-14) assesses the ability to repeat sentences verbatim by asking children to repeat the sentences they hear.

Children are scored correct or incorrect for each item with automated discontinuation rules. Responses are scored by the App and raw and standard scores for each subtest are provided along with LanguageScreen total standard score.[6]

Renfrew Action Picture Tests (RAPT)

Like LanguageScreen, RAPT is a one-to-one administered test. It assesses the speech and language development of children who are between 3 and 8 ½ years of age by using 10 picture cards, depicting a range of everyday scenarios that stimulate children to give samples of spoken language. These speech samples are than scored on two criteria measuring information content and grammatical structure:

1. Information (10 items, score range 0-41), where children are asked to describe the information shown in a set of pictures.

[6] Standard scores are age-standardised and express a child's performance relative to their age.

2. Grammar (10 items, score range 0-39), which checks the grammar used by children, such as the use of verb tenses, while describing the information shown in a set of pictures.

The primary outcome measure will be a latent variable constructed from a Principal Components Analysis (PCA) of the LanguageScreen subtest scores and RAPT Information and Grammar scores collected at endline. These measures capture complementary domains of oral language, with LanguageScreen assessing structured receptive and expressive skills and RAPT capturing spontaneous, narrative-based expressive language. Combining the measures via the first component of a PCA provides a more sensitive and comprehensive measure of language skills for the primary outcome then picking an individual component. For both LanguageScreen and RAPT, the raw scores will be used in the PCA, apart from the Listening Comprehension score, for which the standard score will be used. Each subtest's raw score (and the age standardised score in the case of the listening comprehension score of the LanguageScreen test) will be centred and scaled to have a mean of 0 and standard deviation of 1 before subjecting to PCA. This is necessary so that differences in the scale magnitude for each subtest's raw score do not influence the PCA. The importance that each subtest will have in the makeup of the score will be determined by the PCA method which weights based on how much variance it accounts for.

PCA is a dimensionality reduction method that works by finding successive uncorrelated linear combinations of the variables in a dataset while trying to minimise information loss. The first principal component is the linear combination of variables that has maximum variance of all possible linear combinations and thus it accounts for as much variation in the data as possible. PCA is one of the oldest and most widely used data dimensionality reduction techniques. A detailed description of PCA can be found in Groth *et al.*, 2013. The PCA will be run using the built-in prcomp function within R version 4.2.1 (R Core Team, 2022).

Secondary outcomes

The six subscales (four LanguageScreen subtests and two RAPT subtests) outlined above will be treated as secondary outcome measures, with analyses run separately for each of the six subscales. In four of the five analyses, the raw subscore will be used. The analysis for the LanguageScreen Listening Comprehension subtest will use the standard score for the reasons described above.

FOCUS-34

The FOCUS-34 (CanChild, 2020) is a validated outcome measure used to assess communication skills in children between the ages of 18 months to five years 11 months including children who have a variety of communication disorders. The instrument assesses the effectiveness with which children communicate in life situations where knowledge, information, ideas or feelings are externalised and includes, for example, behaviours relating to intelligibility, communicating independently with unfamiliar adults, turn-taking and willingness to talk (Thomas-Stonell et al., 2013; Thomas-Stonell et al., 2013). For this evaluation, we will use the parent version which will be administered by the practitioner. The instrument allows for people other than the parents to administer the measure and provides instructions on how to achieve this. There are two parts of the assessment: part 1 has 23 items with response score range of 1-7 focusing on how much the statements align with the child, while part 2 has 11 items with a score range of 1-7 focussing on what the child can do linguistically. The scores for each of the items is added and this constitutes

the total score (range of 34 – 238 points).

Practitioners' knowledge and confidence

The Practitioners' knowledge and confidence, lead practitioner survey is an adapted version of the second scale of the Early Math Beliefs and Confidence Survey (EM-BCS),14 which has been validated for assessing teachers 'confidence in helping preschool children (aged 3-4 years) to learn maths (Chen and McCray, 2013). The adapted version was also used in the effectiveness trial of NELI Preschool (EEF, 2025). The measure consists of 11 statements with a Likert scale to indicate level of agreement with each. The statements have been adapted to capture the shift in domain (from maths to language development). Each question receives a score between one and five (depending on the selected option on the Likert scale). Scores for all questions are then combined to produce a summary score between 11 and 55, with higher scores demonstrating greater confidence in helping nursery aged children develop their language skills.

The psychometric properties of the assessments can be found in the study protocol (Roy et al., 2022).

Baseline measure

Baseline for Primary outcome

LanguageScreen will form the baseline measure for the primary outcome. As with the primary outcome measure, a PCA will be performed on the four baseline LanguageScreen subscores (using raw scores apart from Listening Comprehension, which will use the standard score). The first component score will form the baseline measure. The PCA will follow the same procedure as that used for the primary outcome measure.

Baseline for Secondary outcomes

We will also use LanguageScreen as a baseline measure for secondary outcomes. For the four LanguageScreen outcomes (Expressive Vocabulary, Listening Comprehension, Receptive Vocabulary and Sentence Repetition), we will use the corresponding score from each baseline LanguageScreen subtest. For RAPT Information and Grammar outcomes, as well as for the FOCUS-34 measure, the first component score obtained from the PCA performed on the four baseline LanguageScreen subscores will be included as a baseline measurement. All PCAs will be run using the built-in prcomp function within R version 4.2.1 (R Core Team, 2022). Moreover, the Practitioners' knowledge and confidence, (lead practitioner survey) will be used to establish both a baseline and an outcome for when that measure is used.

Sample size calculations overview

All sample size calculations below were performed using the PowerUpR package (Bulus *et al.*, 2021) in the R statistical software, using the function 'mdes.cra2'. This function computes the Minimum Detectable Effect Size (MDES) for a two-level cluster-randomised trial.

The target MDES for this study is 0.16, a figure informed by the effect size seen in the ETB pilot trial (EEF, 2024) and similar Early Years language interventions, particularly the 0.16 effect size

observed in the targeted subgroup of children for the NELI Preschool efficacy trial (Dimova *et al.*, 2020). A similar effect size is anticipated for this trial.

We estimated the ICC using the NELI Preschool effectiveness trial baseline data. Because as the evaluators of NELI Preschool we had direct access to the data, we were able to select how many children were included per class when calculating the ICC. The ICC for the NELI Preschool whole-class component was 0.21, but we found this increased as we restricted the sample to the children with the lowest scores in each class. We decided on an ICC of 0.335, which was obtained by calculating the ICC for the 16 children with the lowest baseline LanguageScreen scores in each setting (reflecting this trial's design). Due to this high expected ICC we considered using a design that randomises children within settings (see study protocol for details), which would reduce the impact of a high ICC on the sample size calculation. However, due to concerns around contamination, we eventually decided to randomise at the setting level, which is reflected in the sample size calculations described here.

We also estimated a pre/post correlation of 0.75, based on the value reported in the NELI effectiveness evaluation (EEF, 2020). The formula used by PowerUp requires the proportion of variance explained at level 1 and level 2 (R_1^2 and R_2^2) are both specified. As we do not have prior evidence for exact values of R_1^2 and R_2^2 we squared 0.75 and used this for both parameters (i.e. assuming they would both be of similar size to the overall R^2). Given that R_1^2 and R_2^2 cannot both be higher or both be lower than the overall R^2 (i.e. if R_1^2 and R_2^2 are not both equal to R^2 one will be higher and one will be lower) we believe $R_1^2 = R_2^2 = R^2$ to be a reasonable default estimate, in the absence of other evidence.

Depending on setting size, we allowed between six and 32 children to be randomised per setting and assumed an average of 16 randomised. Amongst the 16 children randomised we expected approximately 12 to be analysed, anticipating 23% child-level attrition (EEF, 2019)[7]. Using these estimates, 211 settings were required to detect an effect size of 0.16 with a power of 0.8. This is the number of settings analysed, allowing for 10% setting-level attrition between recruitment and analysis, so the recruitment target for the trial was to achieve 235 settings completing baseline. Given this number of settings, the MDES in the subgroup of EYPP children would be 0.21 at the analysis stage, assuming a 15.1% eligibility rate[8]. The proportion of EYPP children in the trial may be higher than this if EYPP children have lower LanguageScreen scores than their peers on average, so the MDES of 0.21 was a somewhat conservative estimate.

The number of settings randomised was 224, somewhat lower than the target of 235. There were also fewer children randomised per setting: about 12 on average rather than the expected 16. The lower number of children and settings slightly increased the MDES anticipated at the analysis stage, from 0.16 to 0.17. In the EYPP subgroup the anticipated MDES increased from 0.21 to 0.24. As EYPP data has not yet been provided by settings at the time of writing, we again assume a 15.1% EYPP eligibility rate for this calculation. As in the protocol calculations, we assume 10% setting-level attrition and 23% child-level attrition. The assumption of 10% setting-

[7] After the protocol power calculation an updated version of this document became available indicating a mean of 16% child-level attrition amongst EEF EY trials. For consistency with the protocol calculation we continue to use 23% here and note that the difference in the MDES resulting from the two attrition estimates is only around 0.001.

[8] Based on DfE statistics for 2025.

level attrition may be conservative: in practice settings often withdraw between recruitment and randomisation, rather than after randomisation (i.e. part of the 10% deduction has already occurred when the 224 settings are randomised).

Table 3: Sample size calculations

		Protocol				Randomisation			
		OVERALL		EYPP		OVERALL		EYPP	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.15	0.16	0.19	0.21	0.16	0.17	0.21	0.24
Pre-test/ post-test correlations	level 1 (child)	0.75		0.75		0.75		0.75	
	level 2 (setting)	0.75		0.75		0.75		0.75	
	level 2 (setting)	0.3352		0.3352		0.3352		0.3352	
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		16	12.32	2.42	1.86	11.86	9.13	1.79	1.38
Number of settings	intervention	117.5	105.5	117.5	105.5	112	101	112	101
	control	117.5	105.5	117.5	105.5	112	101	112	101
	total	235	211	235	211	224	202	224	202
Number of children	intervention	1880	1300	284	196	1311	909	198	137
	control	1880	1300	284	196	1345	932	203	141
	total	3760	2600	568	392	2656	1841	401	278

	Setting level	Child level
Expected attrition (%)	10%	23%

Analysis

All analysis will be conducted in line with EEF's Statistical Analysis Guidance. Apart from the missing data analysis, all analysis will include 'complete cases' only: children with any missing analysis variables will not be included. Analysis will be conducted on an 'intention-to-treat' basis (except the compliance analysis), with children analysed according to their setting's randomised allocation, regardless of how much of the ETB intervention they received. The analyst will not be blinded to treatment allocation for any of the analysis described.

Primary outcome analysis

RQ1: How effective is ETB at improving the oral language skills of children with low to medium prior oral language skills when compared to children in control group settings?

The primary research question (RQ1) will be answered using a linear multilevel model, specifically a two-level random intercept model. The two-level random intercept model is given by:

:

$$ORAL\_LANGUAGE_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \mu_{0j} + \varepsilon_{ij} \quad (1)$$

Where:

- $ORAL\_LANGUAGE_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above
- μ_{0j} is the random intercept for setting j,
- $INTERVENTION_j$ is the intervention/control dummy variable for setting j (0= control; 1 = intervention)
- $BL\_ORAL\_LANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.
- $\mathrm{SETTING\_TYPE}_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) also used as a stratifier at randomisation
- and ε_{ij} is the residual error term for child i at setting j.

Both the random intercept and residual term are assumed to be independently normally distributed:

$$\mu_{0j} \sim N\left(0, \sigma^2_{\mu_0}\right) \quad (2)$$

$$\epsilon_{ij} \sim N(0, \sigma^2_e) \quad (3)$$

The two-level random effects model will be run in R version 4.5.2 (R Core Team, 2025) using the package 'lme4' (Bates *et al.*, 2015).

Secondary outcome analysis

RQ2: How effective is ETB at improving different aspects of children's oral language skills who have low to medium prior oral language skills as measured by the subtests of the LanguageScreen and RAPT instruments when compared to children in control group settings?

RQ2 will be answered by running six models. For each, the secondary outcome measure will be the total raw score (apart from the standard score used for Listening Comprehension) for each subtest of LanguageScreen and RAPT at endline as outlined above. Multilevel models will be run, with each secondary outcome forming the dependent variable. As per the primary analysis model, each model will account for the clustering of children within settings as a random effect and will include the same covariates. The two-level models are given by:

$$\begin{aligned}\text{EXPRESSIVE\_VOCABULARY}_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_\text{EXPRESSIVE\_VOCABULARY}_{ij} \\ +\beta_3 \text{SETTING\_TYPE}_\text{j} + \beta_4 \text{SETTING\_SIZE}_\text{j} + \mu_{0j} + \varepsilon_{ij} \end{aligned} \quad (4)$$

$$\begin{aligned}LISTENING\_COMPREHENSION_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_LISTENING\_COMPREHENSION_{ij} \\ +\beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \mu_{0j} + \varepsilon_{ij} \end{aligned} \quad (5)$$

$$\begin{aligned}RECEPTIVE\_VOCABULARY_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_RECEPTIVE\_VOCABULARY_{ij} \\ +\beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_J + \mu_{0j} + \varepsilon_{ij} \end{aligned} \quad (6)$$

$$\begin{aligned}SENTENCE\_REPETITION_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_SENTENCE\_REPETITION_{ij} \\ +\beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \mu_{0j} + \varepsilon_{ij} \end{aligned} \quad (7)$$

$$\begin{aligned}INFORMATION_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} \\ +\beta_{434} SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_J + \mu_{0j} + \varepsilon_{ij} \end{aligned} \quad (8)$$

$$\begin{aligned}GRAMMAR_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} \\ +\beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_J + \mu_{0j} + \varepsilon_{ij} \end{aligned} \quad (9)$$

Where:

- $EXPRESSIVE\_VOCABULARY_{ij}$ is the endline Expressive Vocabulary raw score for child i at setting j in equation (4) above.
- $LISTENING\_COMPREHENSION_{ij}$ is the endline Listening Comprehension standard score for child i at setting j in equation (5) above.
- $BL\_LISTENING\_COMPREHENSION_{ij}$ is the baseline Listening Comprehension standard score for child i at setting j in equation (5) above.
- $RECEPTIVE\_VOCABULARY_{ij}$ is the endline Receptive Vocabulary raw score for child i at setting $j$in equation (6) above.
- $BL\_RECEPTIVE\_VOCABULARY_{ij}$ is the baseline Receptive Vocabulary raw score for child i at setting j in equation (6) above.
- $SENTENCE\_REPETITION_{ij}$ is the endline Sentence Repetition raw score for child i at setting j in equation (7) above.
- $BL\_SENTENCE\_REPETITION_{ij}$ is the baseline Sentence Repetition raw score for child i at

setting j in equation (7) above.

- $INFORMATION_{ij}$ is the endline Information raw score for child i at setting j in equation (8) above.
- $GRAMMAR_{ij}$ is the endline Grammar raw score for child i at setting j in equation (9) above.
- As in the primary outcome analysis, u_{0j} is the random intercept for setting j in all equations above.
- $INTERVENTION_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention) in all equation above.
- $BL\_\mathrm{ORAL\_LANGUAGE}_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined in equations (8) and (9) above.
- $SETTING\_TYPE_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation in all equations above.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) also used as a stratifier at randomisation in all equation above.
- εij is the residual error term for child i at setting j in all equations above.

In each model, both the random intercept and the residual term are assumed to be independently normally distributed as specified, respectively, in equations (2) and (3) above. As per the primary analysis, the two-level random effects models will also be run in R (version 4.5.2; R Core Team 2026) using the package 'lme4' (Bates *et al.*, 2015).

RQ3: How effective is ETB at improving Communicative Participation of 3–4-year-old children (measured by FOCUS-34) with low to medium prior oral language skills?

$$COMMUNICATIVE\_PARTICIPATION_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING_j + \beta_4 SETTING\_TYPE_j + \beta_5 SETTING\_SIZE_j + \mu_{0j} + \varepsilon_{ij} \quad (10)$$

Where:

- $COMMUNICATIVE\_PARTICIPATION_{ij}$ is the endline raw score Communicative Participation obtained through the FOCUS-34 instrument for child i at setting j.
- μ_{0j} is the random intercept for setting j.
- $INTERVENTION_j$ is the intervention/control dummy variable for setting j (0= control; 1 = intervention).
- $BL\_ORAL\_LANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.
- $SETTING\_TYPE_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) also used as a stratifier at randomisation.
- ε_{ij} is the residual error term for child i at setting j.

Subgroup analyses

Several subgroup analyses are planned to ascertain the intervention's effect on specific subgroups as defined in RQ4 to RQ6. In each instance, they will be analysed with a two-level random intercepts model as defined for the primary analysis (equation (1)), using the same dependent variable and covariates.

RQ4: How does the effectiveness of ETB vary for PVI settings compared to maintained settings?

To answer question RQ4, three models will be run. The first will replicate the primary analysis model (equation (1)) on the subgroup of maintained settings. The second will replicate the primary analysis model (equation (1)) on the subgroup of PVI settings. The third will be similar to the primary analysis model but will include an interaction term between the setting type covariate and the intervention/control dummy variable. This interaction model is defined below:

$$ORAL\_LANGUAGE_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \beta_5 INTERVENTION_j * SETTING\_TYPE_j + \mu_{0j} + \varepsilon_{ij} \quad (11)$$

Where:

- $ORAL\_LANGUAGE_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above.
- u_{0j} is the random intercept for setting j.
- $INTERVENTION_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention).
- $BL\_ORAL\_LANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) used as a stratifier at randomisation.
- $INTERVENTION_j * SETTING\_TYPE_j$ is the interaction between group allocation and setting type eligibility for child i at setting j.
- ε_ij is the residual error term for child i at setting j.

RQ5: How effective is ETB at improving the oral language skills of 3–4-year-old disadvantaged children (children eligible for EYPP) with low to medium prior oral language skills?

This analysis will seek to ascertain the intervention's effect on the subgroup of EYPP-eligible children, where EYPP eligibility will be identified by settings during the Spring term 2026 and will include EYPP eligible children who were selected for ETB prior to randomisation.

To answer question RQ5, two models will be run. The first will replicate the primary analysis model (equation (1)) on the subgroup of EYPP-eligible children. The second will be similar to the primary analysis model but will include interaction term between the EYPP eligibility covariate and the intervention/control dummy variable. This interaction model is defined below:

$$ORAL\_LANGUAGE_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \beta_5 EYPP_{ij} + \beta_6 INTERVENTION_j * EYPP_{ij} + \mu_{0j} + \varepsilon_{ij} \quad (12)$$

Where:

- $ORAL\_LANGUAGE_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above.
- $\mu_{0j}$is the random intercept for setting j.
- $INTERVENTION_j$is the intervention/control dummy variable for setting j (0 = control; 1 = intervention).
- $BL\_ORAL\_LANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.
- $SETTING\_TYPE_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) also used as a stratifier at randomisation.
- $EYPP_{ij}$is the binary variable that shows EYPP eligibility (0 = not eligible; 1 = eligible) for child i at setting j.
- $INTERVENTION_j * EYPP_{ij}$ is the interaction between group allocation and EYPP eligibility for child i at setting j.
- ε_{ij} is the residual error term for child i at setting j.

RQ6: What is the differential impact of ETB at improving the oral language skills of 3–4-year-old children based on a traffic light classification of children's prior oral language skills as measured by LanguageScreen at baseline?

To answer question RQ6, four models will be run. The first three will replicate the primary analysis model (equation (1)) on the subgroup of (i) children classified as RED by the baseline traffic light tracking method described above, (ii) children classified as AMBER and (iii) children classified as GREEN. The fourth model will include all children regardless of traffic light colour. The model will include a traffic light covariate and an interaction term between the covariate and the intervention/control dummy variable. The interaction model is defined by the equation below:

$$ORAL\_LANGUAGE_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \beta_5 BL\_TRAFFIC\_LIGHT\_GREEN_{ij} + \beta_6 INTERVENTION_j * BL\_TRAFFIC\_LIGHT\_GREEN_{ij} + \beta_7 BL\_TRAFFIC\_LIGHT\_RED_{ij} + \beta_8 INTERVENTION_j * BL\_TRAFFIC\_LIGHT\_RED_{ij} \mu_{0j} + \varepsilon_{ij} \quad (13)$$

Where:

- $ORAL_LANGUAGE_{ij}$is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above.
- $\mu_{0j}$is the random intercept for setting j

- $INTERVENTION_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention).
- $BL\_ORAL\_LANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.
- $SETTING\_TYPE_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) also used as a stratifier at randomisation.
- $BL\_TRAFFIC\_LIGHT\_GREEN_{ij}$ is the performance colour status dummy variable obtained from the baseline screening (0 = Not green; 1 = Green)) for child i at setting j.
- $INTERVENTION_j * BL\_TRAFFIC\_LIGHT\_GREEN_{ij}$ is the interaction between group allocation and the traffic light green colour status dummy variable for child i at setting j.
- $BL\_TRAFFIC\_LIGHT\_RED_{ij}$ is the performance colour status dummy variable obtained from the baseline screening (0 = Not red; 1 = red) for child i at setting j.
- $INTERVENTION_j * BL\_TRAFFIC\_LIGHT\_RED_{ij}$ is the interaction between group allocation and the traffic light red colour status dummy variable for child i at setting j,
- and ε_{ij} is the residual error term for child i at setting j.

Additional analyses

RQ7: How effective is ETB at improving practitioners' knowledge and confidence in teaching language skills?

RQ7 will be answered using a linear model at the school level given by:

$$PRACTITIONER\_CONFIDENCE_i = \beta_0 + \beta_1 INTERVENTION_i + \beta_2 BL\_PRACTITIONER\_CONFIDENCE_i + \beta_3 SETTING\_TYPE_i + \beta_4 SETTING\_SIZE_i + \varepsilon_i \quad (14)$$

Where:

- $PRACTITIONER\_CONFIDENCE_i$ is the practitioners' knowledge and confidence score obtained from the leader practitioner survey at setting i.
- $INTERVENTION_i$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention) at setting i.
- $BL\_PRACTITIONER\_CONFIDENCE_i$ is the baseline knowledge and confidence score at setting i.(range 11-55 points)
- $SETTING\_TYPE_i$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation at setting i,
- $SETTING\_SIZE_i$ is the setting size dummy variable (0 = fewer than 20 3- to 4-year-olds; 1 = 20 or more 3- to 4-year-olds) also used as a stratifier at randomisation at setting i

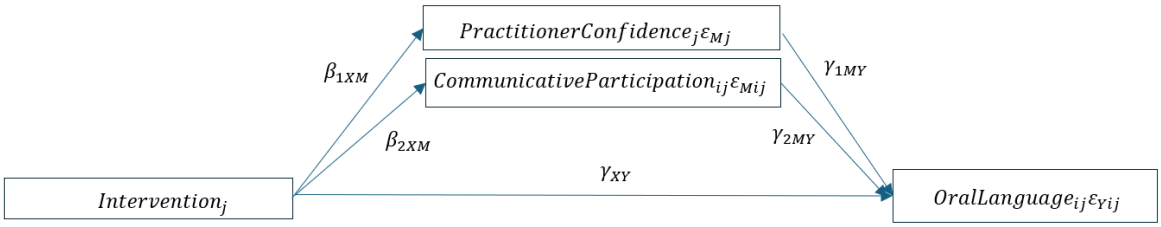
- and ε_i is the residual error term at setting i.

RQ8: Is the effectiveness of ETB mediated by practitioners' knowledge and confidence?

RQ9: Is the effectiveness of ETB mediated by children's Communicative Participation (as measured by FOCUS-34)?

These two questions will be addressed by three mediation models, the main model will look at the simultaneous mediation effect of practitioners' knowledge and confidence and children's Communicative Participation while the other two models will look at each individual mediation effect in isolation. This analysis will be conditional on there being sufficient evidence that the intervention has an impact on the primary outcome. It will use a Structural Equation Model (SEM) approach. The outcome measure will be children's oral language skills, the primary outcome measure as defined in the primary outcome analysis. Data will be considered at two levels (child and setting), with clustering at the setting level.

The SEM is set out in the diagram below, followed by the formulae for the mediation model.



The Equation below models oral language skills as a function of the mediator, with γ_{MY} measuring the direct effect of practitioner confidence on children's oral language skills and γ_{XY} measuring the direct effect of ETB on children's oral language skills. It is defined as:

$$ORAL\_lANGUAGE_{ij} = \gamma_0 + u_{0j} + \gamma_{XY} INTERVENTION_j + \gamma_{1MY} PRACTITIONER\_CONFIDENCE_j + \gamma_{2MY} COMMUNICATIVE\_PARTICIPATION_{ij} + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \beta_5 BL\_PRACTITIONER\_CONFIDENCE_j + \varepsilon_{Yij} \quad (15)$$

Where:

- $ORAL\_LANGUAGE_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above.
- u_{0j} is the random intercept for setting j.
- $INTERVENTION_j$ is the intervention/control dummy variable (0 = control; 1 = intervention) for setting j.
- $BL\_ORALLANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.
- $PRACTITIONER\_CONFIDENCE_j$ is the endline practitioner confidence score for setting j.
- $COMMUNICATIVE\_PARTICIPATION_{ij}$ is the endline communicative participation for child i at

setting j

- $BL\_PRACTITIONER\_CONFIDENCE_j$ is the equivalent setting-level baseline practitioner score.
- $SETTING\_TYPE_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = fewer than 20 3- to 4-year-olds; 1 = 20 or more 3- to 4-year-olds) also used as a stratifier at randomisation
- and ε_{Yij} is the residual error term for child i at setting j.

The Equation below models how the intervention affects practitioner confidence. β_{Xj} represents how much practitioner confidence changes for intervention practitioners compared to control practitioners. It is defined as:

$$PRACTITIONER\_CONFIDENCE_j = \beta_0 + \beta_{1X} INTERVENTION\_j + \varepsilon_J \quad (16)$$

Where:

- $PRACTITIOJNER\_CONFIDENCE_j$ is the endline practitioner confidence score for setting j.
- $INTERVENTION_j$ is the intervention/control dummy variable (0 = control; 1 = intervention) for setting j
- and ε_j is the residual error term for setting j.

It is assumed that the random effect u_{0j} and two residuals ε_j and ε_{Yj} are independently normally distributed:

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad (17)$$

$$\varepsilon_j \sim N\left(0, \sigma_{ej}^2\right) \quad (18)$$

$$\varepsilon_{Yj} \sim N(0, \sigma_{eY}^2) \quad (19)$$

Similarly the Equation below models how the intervention affects Communicative Participation. β_{XM} represents how much practitioner confidence changes for intervention children compared to control children. It is defined as:

$$COMMUNICATIVE\_PARTICIPATION_{ij} = \beta_0 + \beta_{2X} INTERVENTION_j + \varepsilon_{ij} \quad (20)$$

Where:

- $COMMUNICATIVE\_PARTICIPATION_{ij}$ is the communicative participation score of child i in setting j

- $INTERVENTION_j$ is the intervention/control dummy variable (0 = control; 1 = intervention) for setting j
- and ε_{Mj} is the residual error term for setting j.

It is assumed that the random effect u_{0j} and two residuals ε_j and ε_{Yij} are independently normally distributed:

$$u_{0j} \sim N(0, \sigma_{u0}^2) \tag{21}$$

$$\varepsilon_j \sim N(0, \sigma_{eM}^2) \tag{22}$$

$$\varepsilon_{Yij} \sim N(0, \sigma_{eY}^2) \tag{23}$$

For this mediation analysis it is assumed that the total effect of ETB (i.e. the effect measured by the primary analysis model) is the sum of the direct effect of ETB on children's oral language skills and the indirect effect of ETB mediated through practitioner confidence and communicative participation. This assumption is consistent with standard causal mediation frameworks and additionally assumes that there are no unmeasured confounders affecting both practitioner confidence and the children's language skills. Randomisation at the setting level reduces the risk of confounding biasing the estimates of the relative strengths of the treatment–mediator and treatment–outcome pathways. Under these conditions, treating the total effect as the sum of direct and indirect effects is appropriate and aligns with the programme's Theory of Change.

The direct effect is measured by γ_{XY}whilst the indirect effect is measured by the sum of the products $\beta_{1X} * \gamma_{1Y}\ and\ \beta_{2X} * \gamma_{2Y}$ (and hence the total effect is given by $\gamma_{XY} + (\beta_1 X * \gamma_{1Y}) + (\beta_2 X * \gamma_{2Y})$. The direct effect will represent the effect that ETB has directly on children's language skills, controlling for practitioner confidence and communicative participation, whilst the indirect effect will measure how much of the effect of ETB is transmitted through practitioner confidence and communicative participation. In both instances, a positive effect indicates that ETB improves a child's language skills. The extent of the mediation (i.e. the proportion of the effect of ETB that is mediated through practitioner confidence and communicative participation) will be calculated using the formula:

$$\frac{\beta_{1X} * \gamma_{1Y} + \beta_{2X} * \gamma_{2Y}}{\gamma_{XY} + (\beta_{1X}\gamma_Y) + \gamma_{XY} + (\beta_{2X}\gamma_{2Y})} \tag{24}$$

We will report the coefficients of the direct, indirect and total effects, their confidence intervals, and associate p-values. We will run two additional models each of which will contain only one of the mediation variables to verify that the results for the simultaneous mediation hold. Furthermore, we will report the proportion mediated for each of the models run. The mediation analysis will be run in R version 4.2.1 (R Core Team, 2022) using the 'lavaan' package (Rosseel, 2012).

Imbalance at baseline

To assess imbalance between intervention and control groups at baseline we will produce cross-tabulations of background characteristics of the settings in the sample, reporting both counts and percentages. Running this analysis will ascertain whether the comparisons between intervention and control groups are biased, and whether adjustments need to be made in the analyses.

We will examine the following background characteristics:

1. Setting type, as defined in the primary outcome analysis
2. Size of setting at baseline, as defined in the primary outcome analysis
3. Setting's IDACI quintile
4. Proportion of 3 to 4-year-old children selected for ETB who were eligible for EYPP

We will also consider imbalance between intervention and control groups at the child level by producing cross-tabulations of these background characteristics:

1. EYPP eligibility
2. Gender
3. Age in months
4. Proportions of red, amber and green LanguageScreen classification assigned.

We will further assess imbalance at baseline in terms of children's baseline latent oral language skills score (taken from the first component of the PCA) by comparing the difference in means of the two scores between the intervention and control groups and reporting it as a Hedge's g effect size. For the purpose of computing the effect sizes we will fit two level models without covariates.

Missing data

There is likely to be some degree of missing data for the primary outcome at endline. Where children are unavailable for testing, the reason for this will be established where possible and described in the final report. As per the EEF's statistical analysis guidance (EEF, 2022), if more than 5% of primary outcome data is missing, further missing data analysis will be undertaken. After evaluating to what extent data are missing (i.e. number and proportion of missing cases) and counting the number of complete cases, we will proceed to identify patterns of missingness in terms of the primary outcome variable.

By design, only children who were assessed at baseline will be included in the trial, and so we are not expecting to find missing cases in the data corresponding to any of the covariates of the primary analysis model (baseline oral language skills, setting type and setting size, whose values are already known). As such, we will not investigate missingness in terms of any variables other than the primary outcome measure.

We will investigate patterns of missing data by means of a substantive model for the primary outcome variable, at the two levels assumed by the primary outcome analysis. It will be a logistic regression model, with baseline oral language skills, intervention/control dummy variable, setting type and setting size indicators included as covariates, along with other covariates that might be indicative of missingness. This list includes:

Child Level:

1. EYPP eligibility
2. Gender
3. Age in months

Setting Level:

1. Proportion of 3 to 4-year-old children selected for ETB who were eligible for EYPP
2. Setting's IDACI decile

The outcome variable will be the logit probability of the latent oral language variable being missing. After this stage the analysis will follow the roadmap from EEF statistical analysis guidance. Depending on whether only the outcome or both the outcome and other covariates, conditionally missing on other covariates are statistically significant, we will assess the need for multiple imputation and further sensitivity analysis.

Compliance Analysis

Compliance will be defined at the child level and will be based on attendance at ETB sessions. Attendance records for each child selected for the ETB intervention will be recorded in the Session delivery logs, which will be completed by all intervention settings. There are 3 sessions delivered per week over 9 weeks, so it is possible for participating children to attend up to 27 sessions in total. Child attendance data will be collected via a session delivery log, which will be completed by practitioners for each session. If a session is either not delivered or not attended by a child, the session will not be counted towards their total for the compliance measure. There will be a single compliance measure for this evaluation: the impact of complete attendance (all 27 ETB sessions[9]), compared to no attendance. This will be obtained by estimating the benefit of each additional ETB session (as described below) then multiplying this number by 27.

The benefit of each additional ETB session (i.e. sessions treated as a continuous number) will be estimated using instrumental variable analysis, estimated using two stage least-squares. For the first stage, the compliance variable (number of sessions) will be regressed on the intervention indicator and covariates from the primary analysis. This first stage linear regression will be:

$$COMPLIANCE_{ij} = \beta_0 + \beta_1 INTERVENTION_j + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \varepsilon_{ij} \quad (25)$$

Where:

- $COMPLIANCE_{ij}$ is the number of sessions attended by child *i* at setting *j*.
- $INTERVENTION_j$ is the intervention/control dummy variable for setting *j* (0 = control; 1 = intervention).
- $BL\_ORAL\_LANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.

[9] We note that the number of children attending all 27 sessions may be relatively low, but don't regard this as problematic. We will present the estimate for 27 sessions as the theoretical maximum benefit of the intervention, while also discussing the observed distribution of sessions attended to provide context. The measure will not be adapted or removed based on the number of children attending 27 sessions.

- $SETTING\_TYPE_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) also used as a stratifier at randomisation.
- ε_{ij} is the residual error term for child i at setting j.

For the second stage the primary outcome will be regressed on each pupil's predicted compliance value $\widehat{COMPLIANCE}_{ij}$ obtained from the first stage, in the following linear regression:

$$ORAL\_LANGUAGE_{ij} = \beta_0 + \beta_1 \widehat{COMPLIANCE}_{ij} + \beta_2 BL\_ORAL\_LANGUAGE_{ij} + \beta_3 SETTING\_TYPE_j + \beta_4 SETTING\_SIZE_j + \varepsilon_{ij} \quad (26)$$

Where:

- $\widehat{COMPLIANCE}_{ij}$ is the estimated number of sessions attended by child *i* at setting *j*.
- $ORAL\_LANGUAGE_{ij}$ is the endpoint latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the data as outlined above.
- $BL\_ORAL\_LANGUAGE_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above.
- $SETTING\_TYPE_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation.
- $SETTING\_SIZE_j$ is the setting size dummy variable (0 = 20 or fewer 3- to 4-year-olds; 1 = more than 20 3- to 4-year-olds) also used as a stratifier at randomisation.
- ε_{ij} is the residual error term for child i at setting j.

The coefficient for predicted compliance β_1 in this second stage can be interpreted as the average change in oral language score per additional ETB session received. This coefficient and its 95% confidence interval bounds will then be multiplied by 27 to obtain the compliance measure.

Results from both regression stages will be reported. Instrumental variable analyses will be performed using the R package 'ivreg' (Fox *et al.*, 2021). These models do not include school-level random effects, so instead cluster-robust standard errors will be calculated using the R package 'sandwich' (Zeileis, 2006; Zeileis, Köll and Graham, 2020). The correlation between the intervention indicator and the compliance variables, as well as the F-test for the intervention indicator from the first stage of the two-stage regressions, will be reported.

Intra-cluster correlations (ICCs)

For the primary outcome analysis, we will calculate ICC as:

$$\text{ICC} = \frac{\sigma^2}{\sigma^2 + \alpha} \quad (27)$$

Where σ^2 is the between-setting variance and α is the between-child variance. Pre-test ICCs will be computed using a null model, and post-test ICCs will be derived from the primary outcome analysis ITT model described above (1).

Effect size calculation

We will follow the EEF's statistical analysis guidance (EEF, 2022) to calculate appropriate effect sizes for each analysis. For analyses that have a continuous dependent (outcome) variable, the following equation will be used to calculate effect size, expressed as Hedge's g:

$$g = \frac{\beta}{SD_{pooled}} \quad (28)$$

Where: $\beta$and SD_{pooled} is the pooled standard deviation of the two groups. The pooled standard deviation is calculated as follows:

$$SD_{pooled} = \sqrt{\frac{(SD_1 + SD_2)}{n_1 + n_2}} \quad (29)$$

Where n_1 and n_2 are the sample sizes of the first and second groups, respectively, and SD_1 and SD_2 are the standard deviations of the first and second groups, respectively.

The numerator for the effect size calculation will be the coefficient of the intervention/control dummy variable from the multilevel model for each analysis. The effect sizes will be calculated using the total variance without covariates as the denominator (i.e. equivalent to Hedges' g). Confidence intervals for each effect size will be derived by multiplying the standard error of the intervention/control dummy variable coefficient by 1.96. These will be converted to effect size confidence intervals using the same formula as the effect size itself.FThe effect size of the interactions will be calculated following the EEF analysis guidelines (EEF, 2022):

$$\frac{\beta\, TREATMENT_j + \beta\left(TREATMENT_j * COVARIATE_{ij}\right)}{SD} \quad (30)$$

References

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Bulus, M., Dong, N., Kelcey, B. and Spybrook, J. (2021) 'PowerUpR: power analysis tools for multilevel randomized experiments'. Available at: https://cran.r-project.org/web/packages/PowerUpR/index.html (Accessed: 31 March 2026).

CanChild (2020) *Focus-34*, *CanChild*. Available at: https://canchild.ca/shop/30-focus-34/ (Accessed: 24 July 2025).

Chen, J.-Q. and McCray, J. (2013) *A survey study of early childhood teachers' beliefs and confidence about teaching early math*. Available at: https://earlymath.erikson.edu/wp-content/uploads/2013/12/2013-1-Chen-McCray-Survey-Study-of-Early-Childhood-Teachers-Beliefs-and-Confidence-about-Teaching-Early-Math.pdf (Accessed: 31 March 2026).

Dimova, S., Ilie, S., Elena Rosa, B., Broeks, M., Culora, A. and Sutherland, A. (2020) *The Nuffield early language intervention: evaluation report*. EEF. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Nuffield_Early_Language_Intervention.pdf?v=1761235908 (Accessed: 30 April 2026).

Early Intervention Foundation (2019) 'Early Intervention Foundation guidebook. Early Talk Boost'. Available at: https://guidebook.eif.org.uk/files/pdfs/programmes-early-talk-boost.pdf (Accessed: 30 April 2026).

EEF (2019) *Lessons learnt from EEF Early Years trials: recommendations for evaluators*. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-syntheses/EY_lessons_learnt.pdf?v=1630582495 (Accessed: 24 April 2026).

EEF (2020) *Nuffield early language intervention (re-grant)*, *Education Endowment Foundation*. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/nuffield-early-language-intervention1 (Accessed: 21 April 2026).

EEF (2022) 'Statistical analysis guidance for EEF evaluations'. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1699621596 (Accessed: 30 April 2026).

EEF (2024) *Early Talk Boost: lessons learnt report*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/early_talk_boost_lessons_learnt_report.pdf?v=1752846870 (Accessed: 22 April 2026).

EEF (2025) *NELI preschool - trial*, *Education Endowment Foundation*. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/neli-preschool-2023-24-trial (Accessed: 24 February 2026).

Fox, J., Kleiber, C., Zeileis, A. and Kuschnig, N. (2021) 'ivreg: instrumental-variables regression by "2SLS", "2SM", or "2SMM", with diagnostics'. Available at: https://cran.r-project.org/web/packages/ivreg/index.html (Accessed: 27 January 2025).

Groth, D., Hartmann, S., Klie, S. and Selbig, J. (2013) 'Principal components analysis', in B. Reisfeld and A.N. Mayeno (eds) *Computational Toxicology: Volume II*. Totowa, NJ: Humana Press, pp. 527–547. Available at: https://doi.org/10.1007/978-1-62703-059-5_22.

Hulme, C., McGrane, J., Duta, M., West, G., Cripps, D., Dasgupta, A., Hearne, S., Gardner, R. and Snowling, M. (2024) 'LanguageScreen: the development, validation, and standardization of an automated language assessment app', *Language, Speech, and Hearing Services in Schools*, 55(3), pp. 904–917. Available at: https://doi.org/10.1044/2024_LSHSS-24-00004.

Liberman, I.Y., Shankweiler, D. and Liberman, A.M. (1989) 'The alphabetic principle and learning to read'. ERIC. Available at: https://eric.ed.gov/?id=ED427291 (Accessed: 19 June 2025).

R Core Team (2022) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.r-project.org/ (Accessed: 25 March 2026).

R Core Team (2025) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.r-project.org/ (Accessed: 25 March 2026).

Reeves, L., Hartshorne, M., Black, R., Atkinson, J., Baxter, A. and Pring, T. (2018) 'Early talk boost: a targeted intervention for three year old children with delayed language development', *Child Language Teaching and Therapy*, 34(1), pp. 53–62. Available at: https://doi.org/10.1177/0265659018755526.

Rosseel, Y. (2012) 'lavaan: An R Package for Structural Equation Modeling', *Journal of Statistical Software*, 48(2), pp. 1–36. Available at: https://doi.org/10.18637/jss.v048.i02.

Roy, P., Flemons, L., Morton, C. and Welbourne, S. (2022) *Efficacy Trial of Early Talk Boost*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/early_talk_boost_-_evaluation_protocol_-_v.1.0.0_2025-11-25-101645_mwiy.pdf?v=1769078620 (Accessed: 18 March 2026).

Spinner, L. and Aoife, D. (2025) 'Lessons learnt from EEF early years evaluations: Recommendations for evaluators and delivery teams'. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/lessons_learned_from_EEF_EY_trials_-_v2.1.pdf?v=1774252121 (Accessed: 23 March 2026).

Welbourne, S.R., M. Woollams, A., Crisp, J. and A. Lambon Ralph, M.A. (2011) 'The role of plasticity-related functional reorganization in the explanation of central dyslexias', *Cognitive Neuropsychology*, 28(2), pp. 65–108. Available at: https://doi.org/10.1080/02643294.2011.621937.

Woollams, A.M., Halai, A. and Lambon Ralph, M.A. (2018) 'Mapping the intersection of language and reading: the neural bases of the primary systems hypothesis', *Brain Structure and Function*, 223(8), pp. 3769–3786. Available at: https://doi.org/10.1007/s00429-018-1716-z.

Zeileis, A. (2006) 'Object-oriented computation of sandwich estimators', *Journal of Statistical Software*, 16(9), pp. 1–16. Available at: https://doi.org/10.18637/jss.v016.i09.

Zeileis, A., Köll, S. and Graham, N. (2020) 'Various versatile variances: an object-oriented implementation of clustered covariances in R', *Journal of Statistical Software*, 95(1), pp. 1–36. Available at: https://doi.org/10.18637/jss.v095.i01.

Appendix A: R randomisation code

```
#Load data
Experiment<-openxlsx::read.xlsx("...")

#Stratified sample?
#List the stratification variables if Yes
stratify <- "Yes" # Yes or No
if (stratify =="Yes"){
 stratification<-list("Setting Type", "Setting size") # list the stratification variables here
} else {
 Experiment$stratify_dummy <- 1
 stratification<-list("stratify_dummy")  # leave this as is
}
n_strats<-length(stratification)

#identify the cluster variable
cluster<-"NFERNO"

###8. What time is now? (hh.mm)
time_now<-13.02

aux<-100*trunc(time_now)+100*(time_now-trunc(time_now))
set.seed(aux)
seeds<-sample(1:9999,size=(n_strats+2))

#Duplicated cluster information and lines with no cluster identification
#must removed
Experiment<-Experiment[!duplicated(Experiment[cluster]),]
Experiment<-Experiment[!is.na(Experiment[cluster]),]

#Keep the original order of the columns
if (stratify=="Yes"){
 originalColOrder<-colnames(Experiment)
} else{
 originalColOrder<-colnames(Experiment[,c(1:ncol(Experiment)-1)])
}

###Adding a variable that will allow for the recovery
##of the original order of the data frame rows later on
Experiment$originalRowOrd<-1:nrow(Experiment)

### Ordering Experiment by cluster
Experiment<-Experiment[order(Experiment[,cluster]),]

### Assigning a random order to the stratification
rands<-paste("rand",as.character(1:n_strats),sep="_")

for (i in 1:n_strats){

 aux<-as.data.frame(sort(unique(Experiment[,stratification[[i]]])))
 set.seed(seeds[1])
 seeds<-seeds[-1]
```

```
  aux[rands[i]]<-sample(1:nrow(aux))

  Experiment<-merge(Experiment,aux,by.x=stratification[[i]],by.y=colnames(aux)[1])
}

###Randomise by cluster
set.seed(seeds[1])
seeds<-seeds[-1]
Experiment["rand_cluster"]<-sample(nrow(Experiment))

###Reorder the rows of Experiment by rands and rancluster
rands<-c(rands,"rand_cluster")
aux<-do.call(order,Experiment[rands])
Experiment<-Experiment[aux,]

###Assigning Control or Intervention Group
aux<-rep(1:2,times=ceiling(nrow(Experiment)/2))
Experiment$grp<-aux[1:nrow(Experiment)]

rands<-c(rands,"grp")

aux<-data.frame(group=c("control","intervention"))
set.seed(seeds[1])
aux$randgroup<-sample(1:2)

Experiment<-merge(Experiment,aux,by.x="grp",by.y="randgroup")

##Returning the data frame to its original order
Experiment<-Experiment[order(Experiment$originalRowOrd),]

###Removing the variables that are no longer necessary
if (stratify == "Yes"){
  rands<-c("originalRowOrd",rands)} else{
  rands<-c("originalRowOrd",rands,"stratify_dummy")
  }

rands<-which(colnames(Experiment)%in%rands)
Experiment<-Experiment[,-rands]
originalColOrder<-c(originalColOrder,"group")
Experiment<-Experiment[,originalColOrder]
_C.xlsx")
```