A randomised controlled trial of the effectiveness of the Nuffield Early Language Intervention (NELI) Preschool programme Statistical Analysis Plan

Evaluator (institution): National Foundation for Educational Research
Principal investigator(s): Dr Stephen Welbourne

PROJECT TITLE	A randomised controlled trial of the effectiveness of the Nuffield Early Language Intervention (NELI) Preschool programme
DEVELOPER (INSTITUTION)	OxEd and Assessment (OxEd)
EVALUATOR (INSTITUTION)	National Foundation for Educational Research (NFER)
PRINCIPAL INVESTIGATOR(S)	Dr Stephen Welbourne
PROTOCOL AUTHOR(S)	Palak Roy, Lillian Flemons, Gemma Schwendel, Stephen Welbourne, Elena Rosa Speciani and Merrilyn Groom
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at setting level
TRIAL TYPE	Effectiveness
CHILD AGE RANGE AND KEY STAGE	3 to 4-year-old children
NUMBER OF SETTINGS	303
NUMBER OF CHILDREN	6,783 (4,140 in primary analysis)
PRIMARY OUTCOME MEASURE AND SOURCE	Latent oral language variable formed from expressive and receptive subtests in LanguageScreen and the Renfrew Action Picture Test (RAPT)
SECONDARY OUTCOME MEASURE AND SOURCE	Individual expressive and receptive subtests of LanguageScreen and the RAPT 1. LanguageScreen Expressive Vocabulary 2. LanguageScreen Listening Comprehension 3. LanguageScreen Receptive Vocabulary 4. LanguageScreen Sentence Repetition 5. RAPT information 6. RAPT grammar

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	14/07/2025	*N/A*
1.1	29/04/2026	*Due to changes in how children are assessed against each Early Learning Goal, the value of the Early Years Foundation Stage Profile variables in the National Pupil Database have been updated. This has required adjustments to the way the outcome measure is constructed for longitudinal follow-up analysis.*

Table of contents

Introduction

The Nuffield Early Language Intervention (NELI) Preschool is a 20-week oral language enrichment programme for 3 to 4-year-old children in the year before they enter formal education.

The programme aims to support language development in early years settings via enrichment and targeted components. It is designed around the principles of shared book reading and guided play. The programme comprises a blend of language screening for children, online training and delivery support for practitioners, a scripted programme for in-person delivery in preschools, and supportive materials such as storybooks and digital resources. Enrichment sessions focus on whole-class activities, such as reading books, engaging in dialogical questioning and activities to support the learning of related vocabulary, while the targeted component involves small-group and one-to-one sessions tailored to children at the bottom 20-25% of the class[1] in oral language skills, incorporating activities to support and consolidate learning, enhance narrative skills and scaffolded language production. Settings are given access to an online training platform, as well as an online Delivery Support Hub. The online training course includes a range of texts, videos and quizzes, as well as interactive forums for comments and discussion. Both the training platform and the Delivery Support Hub provides participating practitioners with access to the delivery team, experienced practitioners and speech and language specialists who can provide support and answer any questions.

This effectiveness trial, funded by the EEF and the Department for Education's (DfE) Stronger Practice Hubs (SPHs), aims to build on the findings from the previous efficacy trial (West *et al.*, 2023) by testing whether NELI Preschool 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. Since the efficacy trial included only a small number of PVI settings, this evaluation was designed to ensure that at least one-third of the participating settings are PVI.

This evaluation is employing a setting randomised design, with eligible settings being randomised into two groups: intervention and control. Intervention settings will deliver the full version of the NELI Preschool programme over 20 weeks, including both enrichment and targeted components, with the two components being evaluated separately as part of a secondary analysis. The enrichment component comprises daily whole-class sessions lasting 15-20 minutes, whilst the targeted component consists of three small group sessions (each lasting 15 minutes) and one individual session (lasting 5 minutes) per week.

All 3 to 4-year-old children (as of August 31, 2024) in the preschool classroom were eligible to take part in the trial[2]. These children were assessed by settings using LanguageScreen prior to

[1] This is equivalent to selecting six children from a class of 25-30 children. In the evaluations, this has been implemented by selecting six children with the lowest oral language skills in a class or per setting – see impact evaluation design for further details.

[2] While most analyses will include only children who attend the setting 15 hours or more per week, all 3 to 4-year-old children entitled to Early Years Pupil Premium, regardless of attendance, are eligible for the trial. For further details on child eligibility criteria, alongside setting eligibility criteria, see the 'Participant selection' section in the protocol here.

randomisation. LanguageScreen acted as a screening test for intervention settings to select children to target for additional support and will form a baseline for the primary outcome. Children who were found to be unable to access LanguageScreen were not eligible for the trial.

Design overview

Trial design, including number of arms		Two-arm, cluster randomised
Unit of randomisation		Setting
Stratification variables (if applicable)		Setting type (maintained and PVI) and setting size (settings with 30 or fewer 3 to 4 year olds vs those with more than 30 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
	measure(s) (instrument, scale, source)	1. LanguageScreen Expressive Vocabulary subtest (score range 0-24) 2. LanguageScreen Listening Comprehension subtest (score range 65-135[3]) 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)
Baseline for primary outcome	variable	Oral Language Skills
	measure (instrument, scale, source)	Latent oral language variable formed from expressive and receptive subtests in LanguageScreen

[3] While the raw scores will be used for all other LanguageScreen and RAPT subtests, the standard score will be used for the LanguageScreen Listening Comprehension subtest. Further details are given in the Primary Outcome section.

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 6. Oral language skills
	measure (instrument, scale, source)	1. LanguageScreen Expressive Vocabulary subtest (score range 0-24) 2. LanguageScreen Listening Comprehension subtest (score range 65-135[3]) 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

A number of implementation models for the targeted component of NELI Preschool, with different numbers of targeted intervention groups and children per group, were identified by the delivery team to account for variations between settings' class structure and capacity. The delivery team provided guidance on each of these NELI Preschool implementation models, and settings were asked to choose their model prior to randomisation. After randomisation, intervention settings received Targeted Group Selection Guidance that was in line with these models.

Model A	**One targeted intervention group with six children** This is the most common format, where settings are single form entry and have only one preschool class. The setting will run one whole class session and will select six children for targeted intervention.
Model B	**Multiple targeted intervention groups with six children per group** This model works best in larger, multiple form entry settings where NELI Preschool can be run separately for each class. Whole class sessions will be run for each class, and six children per class will be selected for targeted intervention. In cases where there are not enough children attending 15 hours or more per week to make up a complete targeted intervention group in each class, six children will be selected for targeted intervention across the *setting*.
Model C	**One targeted intervention group with six or more children but split across multiple classes** This model works best for settings with multiple classes that run more than one whole group session (one for each class) but find the targeted intervention in Model B too intensive and unrealistic. Instead, settings ideally select three children from each class to be included in targeted

	intervention irrespective of the number of children per class who attend 15 hours or more per week[4].

In addition to selecting their preferred implementation model and sharing it with NFER, settings provided additional information for each child so that NFER can select the children who would ideally be targeted for intervention in each setting, according to the Targeted Group Selection Guidance. While intervention settings were instructed to select children using the same Targeted Group Selection Guidance, it is anticipated that their selection may differ from what is expected as per the guidance (and thus NFER's selection). In response to this, and to facilitate the various aspects being explored in this this study, a series of analysis samples are defined. These are comprised of children who have been tested at baseline and will be tested at endline and then included in various analyses.

S1: Ideally Targeted Children	Children who should have been selected for the targeted component of NELI Preschool if settings correctly followed OxEd's Targeted Group Selection Guidance for their preferred implementation model (implementation model A, B or C, as described above). NFER will define this sample based on the criteria outlined in that guidance, using baseline LanguageScreen scores and additional information provided by the settings. These criteria include the children with the lowest baseline LanguageScreen scores, attending 15 or more hours per week, who are able to access small group work and for whom settings can schedule all sessions. The number of children selected per setting will depend on the implementation model. Those using Models A and B will have six children per class, whilst those using Model C will have three children per class. This sample will be selected after NFER have received final information from settings regarding children who have left or been withdrawn so that they can be excluded from selection.
S2: Practitioner Targeted Children (intervention settings only)	Children in intervention settings who were actually selected for targeted intervention after randomisation. Intervention settings defined this sample using the same information and criteria as for S1 and shared their selection with NFER in January 2025. This sample may differ from S1 if settings have opted to include children not identified as per the guidance. Settings were requested to select six children per class if implementing Models A or B, and three children per class if implementing Model C.
S3: Enrichment-only Sample	Children who attend the setting at least 15 hours per week and are not part of S1 or S2. NFER will randomly select six children per class from each setting.

[4] For further details on child eligibility criteria, see the 'Participant selection' section in the protocol here.

	This sample will be selected alongside S1, after we have received final information from settings.
S4: EYPP	All 3 to 4-year-old children in the setting who are entitled to Early Years Pupil Premium (EYPP). This may include children who attend the settings for fewer than 15 hours a week and are part of the other samples. This sample will be selected by NFER as late as feasible to ensure that information about children who are entitled to EYPP is as up to date as possible.

The research questions that will be answered within this study are set out in the table below, along with the analysis sample that will be used.

Research Question	Analysis Sample
RQ1 (Primary Research Question): How effective is NELI Preschool at improving oral language skills of 3 to 4-year-old children in intervention settings compared to children in control group settings?	S1 and S3 in both intervention and control settings
RQ2: How effective is NELI Preschool at improving different aspects of children's oral language skills as measured by the subtests of LanguageScreen and RAPT?	
RQ3a: How effective is NELI Preschool at improving the language skills of the subgroup of six children selected to receive the targeted component of the intervention?	S2 intervention settings and S1 control settings
RQ3b: How effective is NELI Preschool at improving the language skills of the subgroup of six children who should have been selected to receive the targeted component of the intervention?	S1 intervention and control settings
RQ4: How effective is NELI Preschool at improving the language skills of children who only receive the whole-class (enrichment) component of the intervention?	S3 intervention and control settings
RQ5: How effective is NELI Preschool at improving the language skills of disadvantaged children as identified by EYPP?	S4 intervention and control settings
RQ6: Is NELI Preschool effective at improving language skills of 3 to 4-year-old children in PVI intervention settings compared to children in PVI control settings?	S1 and S3 in both intervention and control settings for PVI settings only
RQ7: How effective is NELI Preschool at improving the language skills of EAL children?	S1 and S3 in both intervention and control settings

Randomisation

This trial is employing a setting-randomised design. 303 early year settings who returned the necessary child data and had completed at least one baseline assessment were randomised on a 1:1 basis into two groups, intervention and control, stratified by setting type (maintained nurseries or PVI settings) and setting size (30 or fewer 3 to 4 year olds vs more than 30 3 to 4 year olds) to ensure equal group allocation in each stratum.

Originally, sample size calculations indicated that 318 settings would need to be recruited to deliver a well-powered result, with an initial target set of 320 settings. A total of 329 settings were recruited who then began the first stages of the trial. However, 26 withdrew prior to randomisation, leaving 303 settings. Of those 303, 177 were Maintained and 126 were PVI settings. The majority (244) opted for Implementation Model A, with 38 and 21 choosing Implementation Models B and C, respectively. In total, there were 6,783 3 to 4-year-old children at randomisation who had completed LanguageScreen at baseline, with 3,454 children in the intervention group, and 3,329 children in the control group.

Two stratification variables were used for the randomisation:

(i) Setting type – maintained settings and PVI settings, and

(ii) Setting size – settings with 30 or fewer children aged 3 to 4 who had completed baseline LanguageScreen versus those with more than 30 children aged 3 to 4 who had completed baseline LanguageScreen. Setting size was defined within the protocol using number of classes (one class vs more than one class). However, this may not reflect the 'true' size of a setting, as some settings updated the number of classes to accommodate the programme delivery, and the definition of class is not consistent across settings. Therefore, using number of children to measure setting size provides a more consistent definition across all settings.

The randomisation allocation sequence was generated using the built-in sample function within R (version 4.2.1; R Core Team 2022) and used a simple stratified sample approach, with settings randomised into two arms (intervention and control) on a 1:1 basis. Randomisation was carried out by the project statistician who was not blinded to randomisation but was provided the minimum data necessary to carry out the randomisation correctly.

Originally, it was planned to randomise settings in two batches to account for settings that may have been delayed in fulfilling randomisation requirements. However, it was subsequently decided to postpone randomisation of the first batch and complete it for all settings together. 151 settings were randomised to the intervention group with the remaining 152 randomised by default to the control group. Table 1 below sets out the results of the randomisation.

Table 1 – Randomisation Results

	PVI, More than 30 3 to 4 year olds	PVI, 30 or fewer 3 to 4 year olds	Maintained, More than 30 4 to 4 year olds	Maintained, 30 or fewer 3 to 4 year olds	Total
Control	10 (6.6%)	53 (34.9%)	28 (18.4%)	61 (40.1%)	152 (100%)
Intervention	10 (6.6%)	53 (35.1%)	27 (17.9%)	61 (40.4%)	151 (100%)

A copy of the R syntax used to carry out the randomisation is given in Appendix A.

Sample size calculations overview

Tables 2 and 3 below sets out the sample size calculations performed both when the protocol was written and post randomisation, to take into account the actual number of settings and children randomised.

Table 2 – Protocol sample size calculations

		All children (RQ1 & RQ2)	Targeted (RQ3)	Enrichment-only (RQ4)	EYPP (RQ5)	PVI settings (RQ6)
Study parameters taken from		OxEd (West et al., 2023)	RAND (Dimova et al., 2020)	RAND (Dimova et al., 2020)	OxEd (West et al., 2023)	OxEd (West et al., 2023)
MDES*		**0.163**	**0.213**	**0.213**	**0.200**	**0.231**
Pre-test/ post-test correlations	Level 1 (child)	0.81	0.75	0.75	0.81	0.81
	Level 2 (setting)					
ICC	Level 2 (setting)	0.21	0.349	0.349	0.21	0.21
Alpha		0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8
One or two-sided?		Two	Two	Two	Two	Two
Average Cluster Size		12	6	6	2.3	12
Number of Settings	Intervention	151	151	151	151	63
	Control	152	152	152	152	63
	Total	**303**	**303**	**303**	**303**	**126**
Number of Children	Intervention	1908	954	954	366	960
	Control	1908	954	954	366	960
	Total	**3816**	**1908**	**1908**	**732**	**1920**

* The MDES figures presented are adjusted for setting-level attrition of 10% and child-level attrition of 23%

Table 3 – Randomisation Sample Size Calculations

		All children (RQ1 & RQ2)	Targeted (RQ3)	Enrichment-only (RQ4)	EYPP (RQ5)	PVI settings (RQ6)
Study parameters taken from		OxEd (West et al., 2023)	RAND (Dimova et al., 2020)	RAND (Dimova et al., 2020)	OxEd (West et al., 2023)	OxEd (West et al., 2023)
MDES*		**0.156 (0.165)**	**0.206 (0.217)**	**0.206 (0.217)**	**0.190 (0.200)**	**0.244 (0.257)**
Pre-test/ post-test correlations	Level 1 (child)	0.81	0.75	0.75	0.81	0.81
	Level 2 (setting)					

ICC	Level 2 (setting)	0.21	0.349	0.349	0.21	0.21
Alpha		0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8
One or two-sided?		Two	Two	Two	Two	Two
Average Cluster Size**		13.7	6.8	6.8	3.1	13.7
Number of Settings	Intervention	151	151	151	151	63
	Control	152	152	152	152	63
	Total	**303**	**303**	**303**	**303**	**126**
Number of Children	Intervention	2063	1032	1032	473	861
	Control	2077	1038	1038	476	861
	Total	**4140**	**2070**	**2070**	**949**	**1722**

* The two sets of MDES figures presented are adjusted for child-level attrition of 23% without setting level attrition, and for 23% child-level attrition and 10% setting-level attrition (in brackets)

** Average cluster size and number of children takes into account the number of 3 to 4-year-old classes within each setting and the number of children from each setting that will be sampled. For EYPP-eligible children, this is based on the total number of children within each setting and assumes 14% of children are eligible for EYPP. The proportion of EYPP-eligible children assumed here differs from the 9.6% assumed when the protocol was written and reflects the most up-to-date published figures.

Sample size calculations were carried out using the 'PowerUpR' package (Bulus *et al.*, 2021) within R (version 4.2.1; R Core Team 2022). A copy of the R syntax used to perform the sample size calculations is given in Appendix B.

The numbers of settings and children set out in the tables present the recruited and randomised samples before any attrition. Two settings have subsequently withdrawn from the intervention (but not the evaluation). Under ITT principles, these two settings are still considered part of the trial (i.e. are not considered as part of setting-level attrition), will be subject to endline testing and are included in the calculations set out in Table 3. However, MDES has been calculated including an allowance of 23% for child-level attrition, in line with EEF EY lessons learned (EEF, 2019). MDES is additionally presented (in brackets) that assumes 10% setting-level attrition alongside the child-level attrition. The ICC and pre-post correlation parameters used are taken from two prior NELI trials and, for each research question, have been selected from the paper whose population is expected to reflect that used in the analyses for each of the research questions. The main difference between the sets of parameters arises from the much higher ICC that is observed when the analysis is restricted to only six children selected for their low LanguageScreen scores. Broadly speaking, the population in the NELI Preschool efficacy trial (West *et al.*, 2023) reflect the analysis samples being used to answer RQ1 & RQ2 (all children), RQ5 (EYPP children) and RQ6 (children in PVI settings), whilst the population in the NELI effectiveness trial (Dimova *et al.*, 2020) reflects the analysis samples proposed to answer RQ3 (targeted) and RQ4 (enrichment). There are differences in the number of settings and children between protocol (318 recruited settings; assumed 12 children per setting, 3816 in total) and randomisation (303 settings and 4140 children). Whilst there are fewer settings at randomisation, this is offset by a larger number of children per setting than anticipated, and we are still expecting a well-powered trial. Even when taking child-level and setting-level attrition into account the trial is adequately powered

to detect an effect for the EYPP subgroup, and therefore very well powered for the primary research question.

Outcome measures

Primary outcome

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

LanguageScreen is a language assessment tool accessed via an App on a tablet. 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[5].

Similar to 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 that can be evaluated in terms of grammatical structures, sentence length and identifying information.

RAPT has two score schemes (assessing either information or grammar), both of which use the same 10 items and scored independently using scoring guidelines:

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

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

ii 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. The first component from this analysis will be used as the primary outcome measure. 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.

One potential concern with using the standard score instead of the raw score is the possible loss of information. To assess the suitability of this approach, a correlation analysis was conducted on baseline LanguageScreen data. In this analysis, two PCAs were performed. The first used the *raw* Listening Comprehension score along with the raw scores for the other LanguageScreen subtests (Expressive Vocabulary, Receptive Vocabulary and Sentence Repetition), and the first principal component was extracted. The second used the *standard* Listening Comprehension score along with the raw scores for the three other LanguageScreen subtests, and again the first principal component was extracted. These two principal components were then correlated, yielding a Pearson correlation coefficient greater than 0.99 which indicated that this is an appropriate approach which will not result in any appreciable loss of information. The scores used will all be centred and scaled as part of the analysis, with the range of values of the first component likely to be similar to the range of the scaled raw and standard scores. 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.

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, we will use the same baseline measure as the primary outcome, the first component score obtained from the PCA performed on the four baseline LanguageScreen subscores as outlined above.

All PCAs will be run using the built-in prcomp function within R (version 4.2.1; R Core Team 2022).

Analysis

All analyses will be conducted on an intention-to-treat (ITT) basis and follow the EEF's statistical analysis guidance (EEF, 2022). Analysts will not be blinded to group allocation for any of the analyses.

Primary outcome analysis

RQ1: How effective is NELI Preschool at improving oral language skills of 3 to 4-year-old children in intervention settings 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 Although incorporating random slopes at the setting level is theoretically feasible, this approach is not considered appropriate for this trial. This decision is based on two key considerations. First, the available data may be insufficient to reliably estimate all model parameters, particularly for settings with a small number of children. Second, increasing the model's complexity by adding random slopes can lead to convergence issues, especially in data structures characterised by many clusters with relatively few individuals per cluster. These challenges could compromise the stability and interpretability of the model.

The two-level random intercept model is given by:

$$OralLanguage_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (1)$$

Where $OralLanguage_{ij}$ 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), $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 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:

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

$$\varepsilon_{ij} \sim N(0, \sigma_e^2) \quad (3)$$

The analysis samples used will be S1 (ideally targeted children) and S3 (enrichment-only sample) in both intervention and control settings. The two-level random effects model will be run in R (version 4.2.1; R Core Team 2022) using the package 'lme4' (Bates et. al, 2015).

Secondary outcome analysis

RQ2: How effective is NELI Preschool at improving different aspects of children's oral language skills as measured by the subtests of LanguageScreen and RAPT?

RQ2 will be answered by running six multilevel 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. Six 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:

$$ExpressiveVocabulary_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ ExpressiveVocabulary_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (4)$$

$$ListeningComprehension_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ ListeningComprehension_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (5)$$

$$ReceptiveVocabulary_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ ReceptiveVocabulary_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (6)$$

$$SentenceRepetition_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ SentenceRepetition_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (7)$$

$$Information_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (8)$$

$$Grammar_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (9)$$

Where $ExpressiveVocabulary_{ij}$ is the endline Expressive Vocabulary raw score for child i at setting j, $BLExpressiveVocabulary_{ij}$ is the baseline Expressive Vocabulary raw score for child i at setting j, $ListeningComprehension_{ij}$ is the endline Listening Comprehension standard score for child i at setting j, $BLListeningComprehension_{ij}$ is the baseline Listening

Comprehension standard score for child i at setting j, $ReceptiveVocabulary_{ij}$ is the endline Receptive Vocabulary raw score for child i at setting j, $BLReceptiveVocabulary_{ij}$ is the baseline Receptive Vocabulary raw score for child i at setting j, $SentenceRepetition_{ij}$ is the endline Sentence Repetition raw score for child i at setting j, $BLSentenceRepetition_{ij}$ is the baseline Sentence Repetition raw score for child i at setting j, $Information_{ij}$ is the endline Information raw score for child i at setting j, and $Grammar_{ij}$ is the endline Grammar raw score for child i at setting j. As in the primary outcome analysis, u_{0j} is the random intercept for setting j, $intervention_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention), $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 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.

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.

The analysis samples used will be S1 (ideally targeted children) and S3 (enrichment-only sample) in both intervention and control settings. As per the primary analysis, the two-level random effects models will also be run in R (version 4.2.1; R Core Team 2022) using the package 'lme4' (Bates et. al, 2015).

Subgroup analyses

Several subgroup analyses are planned to ascertain the intervention's effect on specific subgroups as defined in RQ3 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.

RQ3a: How effective is NELI Preschool at improving the language skills of the subgroup of six children selected to receive the targeted component of the intervention?

RQ3a will be answered using the subgroup of children selected by settings to receive the targeted intervention. The analysis sample used will be S2 (practitioner targeted children) for intervention settings and S1 (ideally targeted children) for control settings.

RQ3b: How effective is NELI Preschool at improving the language skills of the subgroup of six children who should have been selected to receive the targeted component of the intervention?

RQ3b will be answered using the subgroup of children who should have been selected to receive the targeted component of the intervention, using the S1 analysis sample (ideally targeted children) for both intervention and control settings.

RQ4: How effective is NELI Preschool at improving the language skills of children who only receive the whole-class (enrichment) component of the intervention?

RQ4 will be answered using the subgroup of six randomly selected children who are eligible for NELI Preschool and were not selected for targeted intervention, using the S3 analysis sample (enrichment-only sample) for both intervention and control settings.

RQ5: How effective is NELI Preschool at improving the language skills of disadvantaged children?

This analysis will seek to ascertain the intervention's effect on the subgroup of EYPP-eligible children (S4), where EYPP eligibility will be identified by settings towards the end of the trial[6] and will include all EYPP eligible children, regardless of whether they attend the nursery for 15 hours per week or not or were selected to receive the targeted intervention or not.

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 (S4 for both intervention and control settings). The second will be similar to the primary analysis model but will include all EYPP-eligible children (S1+S3+S4 in both intervention and control settings), as well as an EYPP eligibility covariate and interaction term between that covariate and the intervention/control dummy variable. This interaction model is defined below:

$$OralLanguage_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \beta_5 EYPP_{ij} + \beta_6 intervention_j * EYPP_{ij} + \varepsilon_{ij} \quad (10)$$

Where $OralLanguage_{ij}$ 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), $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 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, $\beta_6 intervention_j * EYPP_{ij}$ is the interaction between group allocation and EYPP eligibility for child i at setting j, and ε_{ij} is the residual error term for child i at setting j.

RQ6: Is NELI Preschool effective at improving language skills of 3 to 4-year-old children in PVI intervention settings compared to children in PVI control settings?

This analysis will seek to ascertain the intervention's effect on children who attend PVI settings, using analysis samples S1 and S3 for PVI settings in both the intervention and control groups. The model is defined below:

[6] We will collect EYPP eligibility at the start of the trial along with all children's data. Due to the rolling nature of applications for EYPP, we will also ask settings to update this information for each child towards the end of the trial to ensure we assess all EYPP eligible children.

$$OralLanguage_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingSize_j + \varepsilon_{ij} \quad (11)$$

Where $OralLanguage_{ij}$ 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), $BLOralLanguage_{ij}$ 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, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 3 to 4 year olds) used as a stratifier at randomisation, and ε_{ij} is the residual error term for child i at setting j.

RQ7: How effective is NELI Preschool at improving the language skills of EAL children?

This analysis will seek to ascertain the intervention's effect on the subgroup of children who speak English as an Additional Language (EAL), using the EAL flag provided by settings to select this subgroup. The analysis samples used will be S1 (ideally targeted children) and S3 (enrichment-only sample) in both intervention and control settings.

To answer question RQ7, two models will be run. The first will replicate the primary analysis model (equation (1)) on the subgroup of EAL children. The second will be similar to the primary analysis model but will include all children (S1+S3) in both intervention and control settings), as well as an EAL covariate and interaction term between that covariate and the intervention/control dummy variable. This interaction model is defined below:

$$OralLanguage_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \beta_5 EAL_{ij} + \beta_6 intervention_j * EAL_{ij} + \varepsilon_{ij} \quad (12)$$

Where $OralLanguage_{ij}$ 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), $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 3 to 4 year olds) also used as a stratifier at randomisation, EAL_{ij} is the binary variable that shows whether a child speaks EAL (0 = not EAL; 1 = EAL) for child i at setting j, $\beta_6 intervention_j * EAL_{ij}$ is the interaction between group allocation and EAL for child i at setting j, and ε_{ij} is the residual error term for child i at setting j.

For all models run for the subgroup analyses, both the random intercept and residual term are assumed to be independently normally distributed and are defined, respectively, by equations (2) and (3) above.

All subgroup analyses will be run using R (version 4.2.1; R Core Team 2022) and will use the package 'lme4' (Bates et. al, 2015).

Additional analyses

MEDIATION ANALYSIS

Practitioner confidence has been conceptualised as a short-term outcome in the ToC (see protocol here). We will conduct a mediation analysis to examine the extent to which improvements in children's oral language skills are mediated by increased practitioner confidence as implied by the ToC.

Practitioner Confidence Measure

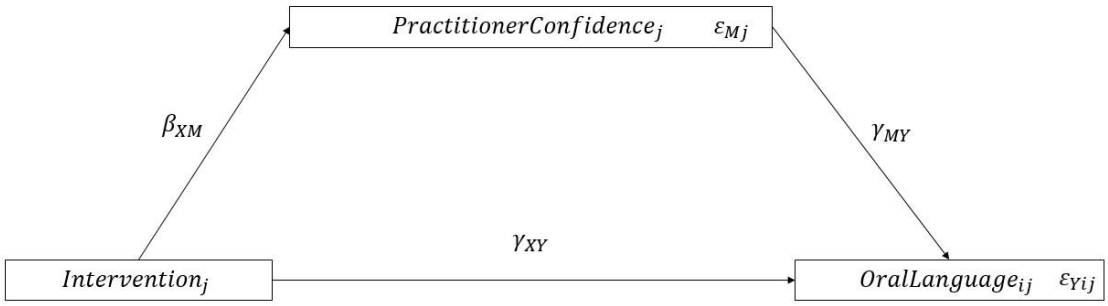
Practitioner confidence will be measured using a version of the second scale of the Early Math Beliefs and Confidence Survey (EM-BCS) that was adapted for assessing teachers' confidence in helping preschool children (aged 3 to 4 years) to learn maths (Chen and McCray, 2013) The measure was further adapted to capture the shift in domain (from maths to language development) and consists of eleven questions about practitioners' confidence in their knowledge of, and ability to help improve, 3 to 4 year olds' language skills. Each question is answered using a 5-point Likert scale (from strongly disagree to strongly agree), meaning it can receive a score between 1 and 5. The full set of questions can be found in Appendix C. The questions were included in the practitioner outcome survey that was administered at baseline and will be administered again at endline as part of the Implementation and Process Evaluation data collection activities.

It was originally anticipated that practitioner confidence would be defined at the setting level, and the score would be generated by summing the responses (to give a total score per practitioner in the range of 11-55) and then calculating the mean across practitioners within each setting. However, there were concerns that the eleven questions are correlated with each other which would need to be accounted for. An Exploratory Factor Analysis (EFA) was therefore performed using the practitioner confidence data collected from the baseline practitioner outcomes survey to determine the most appropriate way to measure practitioner confidence. It established that, whilst the survey questions are highly correlated with each other, the score for the latent variable constructed from a one-factor Confirmatory Factor Analysis (CFA) is highly correlated to the score outlined above ($r^2 > 0.99$). Therefore, the practitioner confidence score will be calculated as originally anticipated, by summing the scores for each practitioner and then calculating the mean score per setting. The results from the EFA of the baseline practitioner confidence measure is presented in Appendix D.

Model for Mediation Analysis

A basic mediation analysis is assumed and will be conditional on there being sufficient evidence that the intervention has an impact on the primary outcome (i.e. that the effect size is greater than 0.1). It will use a Structural Equation Model (SEM) approach. The outcome measure will be children's oral language skills, 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.



Equation (13) 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 NELI Preschool on children's oral language skills. It is defined as:

$$OralLanguage_{ij} = \gamma_0 + u_{0j} + \gamma_{XY} intervention_j + \gamma_{MY} PractitionerConfidence_j + \beta_2 BLOralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \beta_5 BLPractitionerConfidence_j + \varepsilon_{Yij} \quad (13)$$

Where $OralLanguage_{ij}$ 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, $BLOralLanguage_{ij}$ 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, $PractitionerConfidence_j$ is the mean endline practitioner confidence score for setting j, $\beta_5 BLPractitionerConfidence_j$ is the equivalent setting-level baseline practitioner score, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 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.

Equation (14) models how the intervention affects practitioner confidence. β_{XM} represents how much practitioner confidence changes for intervention practitioners compared to control practitioners. It is defined as:

$$PractitionerConfidence_j = \beta_0 + \beta_{XM} intervention_j + \varepsilon_{Mj} \quad (14)$$

Where, $PractitionerConfidence_j$ is the mean endline practitioner confidence score for 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_{Mj} and ε_{Yij} are independently normally distributed:

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad (15)$$

$$\varepsilon_{Mj} \sim N(0, \sigma_{eM}^2) \quad (16)$$

$$\varepsilon_{Yij} \sim N(0, \sigma_{eY}^2) \quad (17)$$

For this mediation analysis it is assumed that the total effect of NELI Preschool (i.e. the effect measured by the primary analysis model) is the sum of the direct effect of NELI Preschool on children's oral language skills and the indirect effect of NELI Preschool mediated through practitioner confidence. The direct effect is measured by γ_{XY} whilst the indirect effect is measured by the product $\beta_{XM} * \gamma_{MY}$ (and hence the total effect is given by $\gamma_{XY} + (\beta_{XM} * \gamma_{MY})$). The direct effect will represent the effect that NELI Preschool has directly on children's language skills, controlling for practitioner confidence, whilst the indirect effect will measure how much of the effect of NELI Preschool is transmitted through practitioner confidence. In both instances, a positive effect indicates that NELI Preschool improves a child's language skills. The extent of the mediation (i.e. the proportion of the effect of NELI Preschool that is mediated through practitioner confidence) will be calculated using the formula:

$$\frac{\beta_{XM} * \gamma_{MY}}{\gamma_{XY} + (\beta_{XM} * \gamma_{MY})} \quad (18)$$

We will report the coefficients of the direct, indirect and total effects, their confidence intervals, and associate p-values. Furthermore, we will report the proportion mediated.

The mediation analysis will be run in R (version 4.2.1; R Core Team 2022) using the 'lavaan' package (Rosseel, 2012).

DOSAGE ANALYSIS

The Theory of Change hypotheses that through exposure to a variety of outputs (attendance at both whole-class and targeted sessions, experience of a wide variety of words, opportunities to engage in activities intended to improve their language skills), children's oral language skills are improved. We will complete a dosage analysis to determine whether higher exposure to these outputs leads to better outcomes. As part of this, we will also explore whether attendance at both the enrichment and targeted component of NELI Preschool leads to higher oral language skills than attendance at the enrichment component only. This analysis will focus on what was actually delivered rather than what was planned. We will use data that is received from the intervention settings about children's attendance at each type of NELI Preschool session. These are completed by settings for each week and returned to NFER in five-week batches in the Session Delivery Logs (SDL). In the SDLs, practitioners record the

number of whole group sessions and targeted sessions each child attended each week during the programme delivery. This information will be used to derive the dosage measures in the analysis models below.

There are concerns that settings may struggle to return SDLs for weeks 16-20 in a timely manner given other trial-related expectations over the same period. Therefore, only settings that have completed all SDLs for weeks 1-15 will be included in the dosage analysis. As a sense check for this, a correlation analysis between data from SDLs in weeks 1-15 and data from SDLs in weeks 16-20 will be completed to determine whether this is reasonable. A Pearson correlation coefficient will be calculated to assess the relationship between the mean total number of weekly sessions (including whole group, small group, and one-to-one sessions) delivered by each setting during weeks 1-15 and weeks 16-20, respectively. If the correlation is statistically significant (i.e., $p < 0.05$), data from weeks 1-15 will be used for the dosage analysis, and linear extrapolation will be applied to estimate values for the missing data between week 16 and the endline testing date. However, if the correlation is not statistically significant, the dosage analysis will be limited to settings that have completed SDLs for all weeks of the programme (weeks 1-20).

The dosage analysis will include all children who received the intervention and for whom we have endline assessment data and information about session attendance from SDLs. This mostly corresponds to analysis samples S2 and S3 in intervention settings but may also include children in S1 who weren't selected by practitioners for targeted intervention but still participated in whole class enrichment sessions. This will also include EYPP-eligible children (S4) who attended fewer than 15 hours per week.

Two separate analyses will be conducted, with each taking the form of a linear model that will set the primary outcome measure as the dependent variable and include the appropriate dosage measure(s) as (a) covariate(s). The model will also include all other covariates used in the primary analysis model.

The first will analyse the impact of the enrichment component only and will include all children who attended enrichment sessions only. The dosage measure will be the number of whole class sessions each child attended.

The two-level random intercepts model for this analysis is given by:

$$OralLanguage_{ij} = \beta_0 + u_{0j} + \beta_1 whole_{ij} + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (19)$$

Where $OralLanguage_{ij}$ 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, $whole_{ij}$ is the number of whole class sessions attended by child i at setting j, $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 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.

The second analysis will analyse the impact of both the enrichment and targeted components and will include all children who attended both enrichment and targeted sessions. Three dosage measures will be derived: the number of whole class sessions attended, the number of small group sessions attended, and the number of one-to-one sessions attended by each child. By keeping the type of session separate within the model, we will be able to assess how much they influence language skills separately.

The two-level random intercepts model for this analysis is given by:

$$OralLanguage_{ij} = \beta_0 + u_{0j} + \beta_1 whole_{ij} + \beta_2 small_{ij} + \beta_3 onetoone_{ij} + \beta_4 BL\ OralLanguage_{ij} + \beta_5 SettingType_j + \beta_6 SettingSize_j + \varepsilon_{ij} \quad (20)$$

Where $OralLanguage_{ij}$ 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 intercept for setting j, $whole_{ij}$ is the number of whole class sessions attended by child i at setting j, $small_{ij}$ is the number of small group sessions attended by child i at setting j, $onetoone_{ij}$ is the number of one-to-one sessions attended by child i at setting j, $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 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.

For both analyses both the random intercept and residual term are assumed to be independently normally distributed and are defined, respectively, by equations (2) and (3) above.

The coefficients of interest in these analyses are β_1, β_2 and β_3, and they will represent how much of an improvement (or decline) in oral language skills is associated with attendance at one of the sessions in question. A positive coefficient will indicate an improvement whilst a negative coefficient will indicate a deterioration. These coefficients will be presented as effect sizes, and we will also report confidence intervals and p-values (see below for how they will be derived).

As before, the two-level random effects model will be run in R (version 4.2.1; R Core Team 2022) using the package 'lme4' (Bates et. al, 2015).

Longitudinal follow-up analyses

The long-term outcome for this trial, as set out in the Theory of Change, is that more children have a better foundation for learning. To assess the impact of NELI Preschool in the longer term, we will carry out a longitudinal analysis, following up children at the end of their Early Years Foundation Stage (i.e. at the end of their Reception Year, one year after the end of the programme delivery). The analysis will use analysis samples S1 (ideally targeted children) and S3 (enrichment sample) from both intervention and control settings (i.e. as per the primary analysis) for whom we have the longitudinal outcomes data (see below). The trial data for these children will be matched to their record within the National Pupil Database (NPD) in order to

access their Early Years Foundation Stage Profile (EYFSP) to determine the effect of the intervention in the longer term. Within the latest version of the EYFSP, there are two Early Learning Goals (ELGs) that relate to language skills, EL1 (listening, attention and understanding) and EL2 (speaking), and three that relate to literacy, EL8 (Comprehension), ELG9 (Word Reading) and ELG10 (Writing). Children are scored against these using the scale 1 = Emerging, 2 = Expected. The scores attained by each child across these five ELGs will be summed to give a score between 5-10. We will produce summary statistics for each of the five ELGs, making comparisons between the intervention and control groups. We will then repeat the primary analysis of the impact evaluation using the derived EYFSP measure outlined above as the outcome measure instead of the latent oral language variable. All the other model parameters will remain the same as the primary analysis.

The two-level random intercepts model is given by:

$$EYFSP_{ij} = \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij} \quad (21)$$

Where $EYFSP_{ij}$ is the EYFSP score constructed from the five EYFSP ELG scores as outlined above for child i at setting j, u_{0j} is the random intercept for setting j, $intervention_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention), $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 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.

In addition, we will run two models that consider the language skills ELGs separately from those that relate to literacy. The first will use ELG1 (listening, attention and understanding) and ELG2 (speaking), with a total score ranging from 2 to 4. The second will use ELG8 (Comprehension), ELG9 (Word Reading) and ELG10 (Writing), with a total score ranging from 3 to 6. For each of these analyses, the model definition is as per equation (21), with $EYFSP_{ij}$ representing the language skills ELG score and literacy ELG score, respectively. The rest of the model remains the same. As it is unlikely that the results from these two models will be sufficiently powered, these results will be presented as exploratory in nature.

In each of these three models, both the random intercept and residual term are assumed to be independently normally distributed and are defined, respectively, by equations (2) and (3).

As before, the two-level random effects models will be run in R (version 4.2.1; R Core Team 2022) using the package 'lme4' (Bates et. al, 2015).

Children with a code of A (not assessed) for any of the ELGs listed above will be counted as missing and excluded from the analysis. As the EYFSP is a statutory assessment and results are being obtained from administrative data, we expect this missing data to be lower than the primary analysis, but the removal of these children may introduce bias depending on the scale and pattern of non-assessment. In line with the recommendations in the EEF's *Statistical*

Analysis Guidance (EEF, 2022), we will undertake a series of checks to assess the potential impact of this missing data. Specifically, we will compare the characteristics of excluded children with those included in the longitudinal analysis to identify whether non-assessment is systematically associated with particular child or setting factors. A missing data analysis will be conducted to examine patterns of missingness in EYFSP scores across ELGs if the proportion of missing EYFSP outcome data exceeds 5%. The analysis will use a logistic regression model analogous to that specified for the missing data analysis of the primary outcome (see below). The dependent variable will be an indicator for whether the EYFSP score is missing (1 = missing, 0 = observed). The model will include baseline oral language skills, a treatment indicator (intervention vs control), and indicators for setting type and setting size as core covariates. Additional covariates hypothesised a priori to be associated with missingness will also be included. Child-level covariates will comprise EYPP eligibility, EAL status, gender, and age in months. Setting-level covariates will include setting type, setting size, the proportion of 3 to 4-year-old children eligible for EYPP, IDACI decile, implementation model, most recent Ofsted rating, and SPH.

The results will be used to assess whether missingness in EYFSP scores is consistent with Missing Completely at Random (MCAR), or whether there is evidence of association between missingness and observed covariates, consistent with a Missing at Random (MAR) or Missing Not at Random (MNAR) mechanism. Specifically, statistically significant associations between the missingness indicator and observed covariates will be interpreted as evidence against MCAR. Covariates found to be predictive of EYFSP missingness will be incorporated into the primary longitudinal analysis model to support the MAR assumption. Estimates from models with and without these additional covariates will be compared to assess the robustness of the primary findings to alternative missing data specifications. If the results differ, this will be indicative that data is either MAR or MNAR and sensitivity analysis may be required. If necessary, sensitivity analysis built on a multi-level multiple imputation will be implemented. The missing EYFSP scores will be imputed using predictive mean matching, with five plausible values derived for each case. The longitudinal analysis model will then be re-run on the five sets of imputed plausible values and the estimates for each model will be pooled into a single set of estimates and standard errors that will be compared to the results of the original analysis. This comparison will indicate the robustness of our findings, and to what degree missing values bias the results of the longitudinal analysis model.

Finally, we anticipate that there may be some children present in this analysis that were missing for the primary analysis. Therefore, we will rerun both the primary and longitudinal analyses using samples where we have all three data points (baseline, primary outcome and longitudinal outcome) as a sensitivity check. In each case, the results of these models will be compared to the primary and longitudinal analysis results to determine their robustness and assess the degree to which missing values may be biasing the results.

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 fair, and whether adjustments need to be made in analyses (e.g., by weighting) to account for any differences.

We will examine the following background characteristics:

1. Setting type, as defined in the primary outcome analysis
2. Size of setting, as defined in the primary outcome analysis
3. Implementation model
4. Latest Ofsted rating
5. Setting's IDACI decile
6. Setting's SPH
7. Proportion of 3 to 4-year-old children eligible for EYPP within the setting

To run this analysis, we will link the settings taking part in the trial to the relevant information contained on the most up to date childcare providers and inspections statistics published by Ofsted[7].This dataset is being used to enable us to access demographic information for PVI settings that isn't held on DfE's Get Information about Schools (GIAS) register. However, this means that we are unable to get hold of published data about the overall proportion of EYPP pupils at each setting, and whether they are in urban or rural areas.

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. EAL status
4. Age in Months

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 them as Hedge's g effect sizes. For the purpose of computing the effect sizes we will fit models without covariates at the two levels defined by the main primary outcome analysis.

It's important that this balance is maintained within the analysis sample throughout the trial. We will therefore repeat this imbalance check for the longitudinal analysis sample in order to ensure that comparisons made at this point between intervention and control groups are fair, and whether any adjustment (e.g., by weighting) is required.

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)

[7] Childcare providers and inspections statistics are published here: Childcare providers and inspections as at 31 August 2024 - GOV.UK

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:

- EYPP eligibility
- EAL status
- Gender
- Age in months

Setting Level:

- Setting type, as defined in the primary outcome analysis
- Size of setting, as defined in the primary outcome analysis
- Proportion of 3 to 4-year-old children eligible for EYPP within the setting
- Setting's IDACI decile
- Implementation model
- Latest Ofsted rating
- Setting's SPH

The outcome variable will be the logit probability of the endline latent oral language variable being missing. After this stage the analysis will follow the roadmap from EEF statistical analysis guidance[8]. We will determine whether the primary outcome data is MCAR or whether there is a degree of correlation between a missing indicator and other covariates (MAR or MNAR). Covariates that are found predictive of missingness for the primary outcome will be added to the primary analysis model and the results of the two models compared. If the results of these two models differ, this will be indicative that data is either MAR or MNAR and sensitivity analysis may be required.

If necessary, sensitivity analysis built on a multi-level multiple imputation will be implemented. The missing primary outcome values will be imputed using predictive mean matching, with five plausible values derived for each case. The primary analysis model will

[8] We are working under the expectation that there will be will no missing values among the models' covariates under MAR (missing at random), and that it will be possible to obtain valid estimates by including covariates predictive of non-response in the substantive models. The models' interpretation is conditional on these covariates being included.

then be re-run on the five sets of imputed plausible values and the estimates for each model will be pooled into a single set of estimates and standard errors. The results of this model will be compared to the results of the primary analysis model. This comparison will indicate the robustness of our findings, and to what degree missing values bias the results of the primary analysis model.

The missing data analysis will be run in R (version 4.2.1; R Core Team 2022) using the packages 'mice' (van Buuren and Groothuis-Oudshoorn, 2011) and 'smcfcs' (Barlett *et al.*, 2024) (pooling of the results of the plausible values models).

Compliance Analysis

Sample: S1 (ideally targeted children) and S3 (enrichment sample)

Compliance of intervention settings will be measured at the setting level and will be defined as completion of NELI Preschool training and delivery of a minimum proportion of NELI Preschool whole class, small group and one-to-one sessions.

A setting-level compliance measure will help us understand whether NELI Preschool would, on average, benefit *all* children if the setting were to implement it as expected. In this scenario, there will be individual children who may not experience NELI Preschool as expected. By taking a setting-level approach, these children will not be excluded from the analysis, thus allowing us to assess the impact of compliance for *all* children within the setting. This setting-level compliance measure is likely to be of greater relevance and use to settings in ascertaining the impact of NELI Preschool on *all* children when implementation is carried out as intended.

It will be measured using a binary compliance variable, with settings being considered compliant (compliance = 1) if they fulfil **all** the following criteria:

1. At least 2 staff members have completed full online training.
2. At least 75% of total expected whole class sessions delivered.
3. At least 70% of total expected small group sessions delivered.
4. At least 65% of total expected individual sessions delivered.

Data on the number of sessions delivered will be recorded on SDLs that are completed by settings for each week and returned to NFER in five-week batches. Due to concerns about settings' ability to return SDLs in later weeks as outlined in the dosage analysis section, we will use SDLs that cover weeks 1-15 of NELI delivery to derive the proportions for criteria 2-4, and will only include settings that return the complete set of week 1-15 SDLs. As per the dosage analysis, this is subject to the results of a correlation analysis between data taken from week 1-15 SDLs and data taken from week 16-20 SDLs indicating that this is an appropriate strategy. If the results of this correlation analysis indicate this is not appropriate, we will use SDLs covering the whole twenty weeks of the programme and only include settings that have returned all SDLs for weeks 1-20. Regardless of which set of SDLs are used, it is likely that the compliance analysis will only include a subset of the settings used in the primary outcome analysis. Therefore, we will also repeat the primary outcome analysis on these settings in order to check the robustness of the primary outcome analysis results. If fewer than 50% of settings submit completed SDLs for weeks 1-15 (or weeks 1-20), we will use only the first

criteria (i.e. at least 2 staff members have completed the full online training) to define compliance.

These criteria will apply across all three implementation models. Whole class and small group sessions will be deemed to have been delivered if at least one child was in attendance in that session for each setting. For individual sessions, the total expected number of sessions for each setting is calculated by multiplying the number of children in the S1 sample within a setting by the number of expected weekly individual sessions (i.e. 3) and the number of weeks that the intervention was delivered for. The denominator for the percentage calculations will be the actual number of sessions that settings delivered within weeks 1-15 and will be derived from the SDLs.

We will explore the effect of this binary compliance indicator on the primary outcome measure. A complier average causal effect (CACE) estimate will be obtained using instrumental variable modelling (IV) as prescribed by EEF statistical analysis guidance. The IV regression will be run via a two-stage least squares model with random group allocation as the instrumental variable for compliance and is given by:

Stage One:

$$Compliance_{ij} = \beta_0 + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_3 SettingType_j + \varepsilon_{ij} \quad (22)$$

Stage Two:

$$OralLanguage_{ij} = \beta_0 + \beta_1 \widehat{Compliance}_{ij} + \beta_2 BLOralLanguage_{ij} + \beta_3 SettingType_j + \beta_3 SettingType_j + \varepsilon_{ij} \quad (23)$$

Where $\widehat{Compliance}_{ijkl}$ is the estimated compliance measure for child i in class j obtained from the first stage of the two-stage regression, $OralLanguage_{ij}$ 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, $intervention_j$ is the intervention/control dummy variable for setting j, $BLOralLanguage_{ij}$ 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, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, and $SettingSize_j$ is the setting size dummy variable (0 = 30 or fewer 3 to 4 year olds; 1 = more than 30 3 to 4 year olds) also used as a stratifier at randomisation. The residuals in both stages are assumed to be independently normally distributed:

$$\varepsilon_{ij} \sim N(0, \sigma^2) \quad (24)$$

The model will be fit using the function ivreg from the R package 'ivreg' (Fox *et al*., 2024*)* and the estimation of causal effects will be done using the functions contained in the 'ivpack' package (Jiang and Small, 2014).

The analyses will, as before, be run in R (version 4.2.1; R Core Team 2022).

Intra-cluster correlations (ICCs)

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

$$ICC_{setting} = \frac{\sigma_j^2}{\sigma_j^2 + \alpha} \quad (25)$$

Where σ_j^2 is the between-setting variance and α is the between-child variance. Pre-test ICCs will be computed using two-level random intercept models with no covariates, 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:

$$g = \frac{M_1 - M_2}{SD_{pooled}} \quad (26)$$

Where: M_1 is the mean of the first group, M_2 is the mean of the second group, and SD_{pooled} is the pooled standard deviation of the two groups. The pooled standard deviation is calculated as follows:

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}} \quad (27)$$

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.

For analyses that have a binary dependent variable, effect sizes will be reported as an odds ratio.

REFERENCES

Bartlett, J., Keogh, R., Bonneville, E. and Thorn Ekstrøm, C. (2024). *smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification*. Available at: https://CRAN.R-project.org/package=smcfcs (Accessed: 28 April 2025)

Bates, D., Maechler, M., Bolker, B., and Walker, S (2015). *Fitting Linear Mixed-Effects Models Using lme4.* Journal of Statistical Software, 67(1), 1-48. Available at: https://doi.org/10.18637/jss.v067.i01 (Accessed: 28 April 2025)

Bulus, M., Dong, N., Kelcey, B. and Spybrook, J. (2021) *Power analysis tools for multilevel randomized experiments*. Available at: https://cran.r-project.org/web/packages/PowerUpR/PowerUpR.pdf (Accessed: 31 March 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 2025)

Dimova, S., Ilie, S., Brown, E.R., Broeks, M., Culora, A. and Sutherland, A. (2020) *Nuffield Early Language Intervention: evaluation report*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Nuffield_Early_Language_Intervention.pdf (Accessed: 28 April 2025)

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: 31 March 2025)

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=1709857308 (Accessed: 31 March 2025)

Fox, J., Kleiber, C., Zeileis, A. and Kuschnig, N. (2024). *ivreg: Instrumental-Variables Regression by '2SLS', '2SM', or '2SMM', with Diagnostics*. R package version 0.6.3 Available at: https://CRAN.R-project.org/package=ivreg (Accessed: 28 April 2025)

Jiang, Y. and Small, D. (2014). *ivpack: Instrumental Variable Estimation*. R package version 1.2. Available at: https://CRAN.R-project.org/package=ivpack (Accessed: 28 April 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: 30 May 2025).

Rosseel, Y. (2012). *Lavaan: An R Package for Structural Equation Modelling.* Journal of Statistical Software, 48(2), 1-36. Available at: https://doi.org/10.18637/jss.v048.i02 (Accessed: 28 April 2025)

van Buuren, S., Groothuis-Oudshoorn, K. (2011). *mice: Multivariate Imputation by Chained Equations in R.* Journal of Statistical Software, 45(3), 1-67. Available at: https://doi.org/10.18637/jss.v045.i03 (Accessed: 28 April 2025)

West, G., Birchenough, J., Korell, C., Diaz, M.R., Lervåg, A., Duta, M., Cripps, D., Gardner, R., Fairhurst, C. and Hulme, C. (2023) *An efficacy trial of the Nuffield Early Language Intervention in Preschool (NELI Preschool)*. ISRCTN. Available at: https://doi.org/10.1186/ISRCTN29838552 (Accessed: 28 April 2025).

APPENDICES

APPENDIX A – RANDOMISATION SYNTAX

```
#
================================================================================

# Script Name: NELI_Randomisation_C.R                              =

# Script Purpose: Carry out NELI Randomisation                     =

#                                                                  =

# Author: Gemma Schwendel                                          =

# Date Created: 15th November 2024                                 =

# Notes:                                                           =

#
================================================================================

#                                                                  =

# Revision Log                                                     =

#
================================================================================

# Date          | Reason for revision                              =

#                                                                  =

#
================================================================================


# Load required packages: uncomment as required. Add others as required

library(openxlsx)

library(dplyr)


### Clear working environment

rm(list=ls())


#1. Set work directory

setwd("K:\\NELIN\\RPO\\Main Trial\\Randomisation")
```

```
#2.identify project

project<-"NELI"


#3.identify classification: c, r or p

classification<-"C"


#4. Number of the randomisation: 1st, 2nd, 3rd ...

randomisation<-1

randomisation<-as.character(as.roman(randomisation))


#5. Load data

Experiment<-read.xlsx("NELIN_Randomisation File_C.xlsx")


# Derive Size of Setting based on number of 3 or 4 year olds in the trial: 30 or less is one class

Experiment$Setting_size <- ifelse(Experiment$Children_34<=30,"One Class","More than One Class")


# Create two-level setting type

Experiment$Setting_type_twolevel <- ifelse(Experiment$Setting_type=="State-
maintained",Experiment$Setting_type,"PVI")


#6.  Stratified sample?

#6a. List the stratification variables if Yes

stratify <- "Yes" # Yes or No

if (stratify =="Yes"){

 stratification<-list("Setting_type_twolevel","Setting_size") # list the stratification variables here

} else {

 Experiment$stratify_dummy <- 1

 stratification<-list("stratify_dummy")  # leave this as is

}

n_strats<-length(stratification)
```

```
#7.identify the cluster variable
cluster<-"NFER_No"

###8. What time is now? (hh.mm)
time_now<-10.15  #### Use this to set the seed

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
#removed (but there aren't any in this case)
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=round(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]


### Create output

excel=createWorkbook()

addWorksheet(excel,"all settings")

addWorksheet(excel,"control settings")

addWorksheet(excel,"intervention settings")


writeData(excel, "all settings", Experiment)

writeData(excel, "control settings", Experiment[Experiment$group=="control",])

writeData(excel, "intervention settings", Experiment[Experiment$group=="intervention",])
```

```
xlsxname<-
paste(paste(paste(paste(project,"Randomisation",sep="_"),randomisation,sep=""),classification,sep
="_"),
            "xlsx",sep=".")


saveWorkbook(excel, xlsxname,overwrite=T)


###Crosstabs

Sys.setenv(R_ZIPCMD="C:/Rtools/bin/zip")


wb<-createWorkbook()

addWorksheet(wb,"cross_tabs")


start<-1


if (stratify=="Yes"){

 for(i in 1:n_strats){

  xtab<-as.matrix(table(Experiment[,stratification[[i]]],Experiment$group))

  xtab<-cbind(xtab,margin.table(xtab,1))

  xtab<-rbind(xtab,margin.table(xtab,2))

  colnames(xtab)[3]<-"Total"

  rownames(xtab)[nrow(xtab)]<-"Total"

  writeData(wb,"cross_tabs",xtab,rowNames=T,startCol=start,startRow = 2)

  writeData(wb,"cross_tabs",stratification[[i]],startCol=start,startRow = 1)

  start<-start+5

 }} else{

  xtab<-as.matrix(addmargins(table(Experiment$group)))

  rownames(xtab)[nrow(xtab)]<-"Total"

  writeData(wb,"cross_tabs",xtab,rowNames=T,startCol=start,startRow = 2)

 }
```

```
addWorksheet(wb,"data")

writeData(wb,"data",Experiment,rowNames=F)

workbookname<-
paste(paste(paste(project,"InfoRandomisation",sep="_"),randomisation,sep=""),"xlsx",sep=".")


addWorksheet(wb,"info")

info<-data.frame(project,classification,randomisation)


for (i in 1:n_strats){

 aux<-paste("stratification",as.character(i),sep="  ")

 info[aux]<-stratification[i]

}

info["clusters"]<-cluster

info["time of run"]<-time_now

info["data xlsx"]<-xlsxname

info["directory"]<-getwd()


writeData(wb,"info",info,rowNames=F)


saveWorkbook(wb,workbookname,overwrite = T)


### Now check the balance

wb<-createWorkbook()

addWorksheet(wb,"Balance Checks")


group <- Experiment %>% group_by(group) %>% summarise(Children=sum(Children_34),Classes =
sum(Classes_34)) %>% mutate(prop_children = Children / sum(Children),prop_class =
Classes/sum(Classes))

writeData(wb,"Balance Checks",group,startRow = 1)
```

```
group <- Experiment %>% group_by(Implementation_model,group) %>%
summarise(Children=sum(Children_34),Classes = sum(Classes_34))%>% mutate(prop_children =
Children / sum(Children),prop_class = Classes/sum(Classes))

writeData(wb,"Balance Checks",group,startRow = 5)

group <- Experiment %>% group_by(Setting_size,group) %>%
summarise(Children=sum(Children_34),Classes = sum(Classes_34))%>% mutate(prop_children =
Children / sum(Children),prop_class = Classes/sum(Classes))

writeData(wb,"Balance Checks",group,startRow = 13)

group <- Experiment %>% group_by(Setting_type_twolevel,group) %>%
summarise(Children=sum(Children_34),Classes = sum(Classes_34))%>% mutate(prop_children =
Children / sum(Children),prop_class = Classes/sum(Classes))

writeData(wb,"Balance Checks",group,startRow = 19)



saveWorkbook(wb,"NELI_Balance_Checks_C.xlsx",overwrite = T)
```

APPENDIX B – MDES CALCULATION SYNTAX

```
#
================================================================================

# Script Name: NELIN_Power_calculations_Post_Randomisation_C.R              =

# Script Purpose: To run the power calculations for NELI after randomisation    =

#             with updated numbers of settings and children             =
=

# Author: Gemma Schwendel                                  =

# Date Created: 15/11/2024                                 =

# Notes:                                          =

#
================================================================================

#                                                =

# Revision Log                                        =

#
================================================================================

# Date          | Reason for revision                        =

#                                                =

#
================================================================================


# Load required packages: uncomment as required. Add others as required

library(openxlsx)

library(dplyr)

library(PowerUpR)


# Load any required functions


# Clear working environment

rm(list=ls())


# Set working directory
```

```
setwd("K:\\NELIN\\CFS\\Sample size calculations")


### Parameters

icc=0.21

prePostCor=0.81


icc2=0.349

prePostCor2=0.75


EYPP_ppn = 0.14 ## Source: https://explore-education-statistics.service.gov.uk/find-statistics/education-provision-children-under-5/2024?form=MG0AV3


### Figure out, on average, how many children for each sample per setting

random <-read.xlsx("K:\\NELIN\\RPO\\Main Trial\\Randomisation\\NELI_RandomisationI_C.xlsx")

random$S1 <- ifelse(random$Implementation_model%in%c("A","B"),random$Classes_34*6,random$Classes_34*3) ## Also the case for S2 & S3

random$S4 <- random$Completed_LS_BL*EYPP_ppn


### Taken from Randomisation Results

nSetting_control = 152

nSetting_intervention = 151

nSetting = nSetting_control+nSetting_intervention

ppn = nSetting_intervention/nSetting

nPupilPerSetting=2*mean(random$S1)

nPupil2=mean(random$S1)

nFSMPupilPerSetting=mean(random$S4)

nPVI = 135

nPVI_cont = 68

nPVI_int = nPVI-nPVI_cont

ppn_PVI = nPVI_int/nPVI
```

For All Children (RQ1 & RQ2) - S1 & S3 for both intervention and control

```
mdes.cra2(power = 0.8,alpha = 0.05,two.tailed = T,rho2=icc,r21=prePostCor^2,p =
ppn,n=round(nPupilPerSetting*0.77),J=round(nSetting))
```

For targeted Children (RQ3) - S2 intervention; S1 control (RQ3a) or S1 for both (RQ3b)

```
mdes.cra2(power = 0.8,alpha = 0.05,two.tailed = T,rho2=icc2,r21=prePostCor2^2,p =
ppn,n=round(nPupil2*0.77),J=round(nSetting))
```

For Enrichment (RQ4) - S3 for both intervention and control

```
mdes.cra2(power = 0.8,alpha = 0.05,two.tailed = T,rho2=icc2,r21=prePostCor2^2,p =
ppn,n=round(nPupil2*0.77),J=round(nSetting))
```

For EYPP (RQ5) - S4 for both intervention and control

```
mdes.cra2(power = 0.8,alpha = 0.05,two.tailed = T,rho2=icc,r21=prePostCor^2,p =
ppn,n=round(nFSMPupilPerSetting*0.77),J=round(nSetting))
```

For PVI Settings (RQ6) - S1 & 3 for PVI settings only

```
mdes.cra2(power = 0.8,alpha = 0.05,two.tailed = T,rho2=icc,r21=prePostCor^2,p =
ppn_PVI,n=round(nPupilPerSetting*0.77),J=round(nPVI))
```

APPENDIX C – PRACTITIONER CONFIDENCE QUESTIONS

I am confident in my knowledge of:	The level of language skills of each child that enters my class
	Reasonable goals for 3–4-year-olds in relation to the development of their language skills
	The best practices and strategies for helping 3-4-year-olds to develop their language skills
	National standards for language skills for 3-4-year-olds
	The best ways to assess the language skills (expressive language and understanding of language) of 3-4-year-olds throughout the year
I am confident in my ability to:	Gauge the level of language skills of 3-4-year-olds in my class
	Incorporate regular opportunities to learn and/or practise language skills into common preschool situations (such as art or dramatic play)
	Plan activities to help 3-4-year-olds to develop their language skills
	Support the language development of 3-4-year-olds when they make spontaneous comments/discoveries
	Help 3-4-year-olds to navigate their confusion when they are developing their language skills
	Translate assessment or screening results into curriculum plans on both a group and individual basis

APPENDIX D – RESULTS FROM THE EXPLORATORY FACTOR ANALYSIS OF THE BASELINE PRACTITIONER CONFIDENCE MEASURE

As outlined in the Mediation Analysis section, it was important to determine the most appropriate way to measure practitioner confidence. It was originally proposed that this would be measured at the setting level by summing the responses for each practitioner and then calculating the setting mean for both baseline and endline. However, there were concerns that the items within the practitioner's confidence survey were correlated and that deriving a latent variable would be more appropriate. To explore this, data from the practitioner confidence measure collected in the baseline practitioner outcomes survey was analysed, using built-in functions and the 'lavaan' package (Rosseel, 2012) within R (version 4.2.1; R Core Team 2022).

1. Assess whether a latent variable is appropriate

The first step was to assess whether the questions used to construct the measure are correlated and therefore whether a factor analysis is required to derive a latent variable. If they are not correlated, a simple sum of the scores would be a sufficient measure. Correlation coefficients between the eleven items range between 0.33-0.72, and were all statistically significant, suggesting that a latent variable is appropriate. This was confirmed by additional data checks (KMO = 0.92; Barlett's check for sphericity $p < 0.000$).

2. Exploratory Factor Analysis to determine factor structure

The next step carried out an EFA to determine the factor structure, focusing on the most appropriate number of latent variables to construct. The eigenvalues and the accompanying scree plot (Figure 1) suggest that a one-factor latent practitioner confidence variable would be appropriate for the mediation analysis, which aligns with the underlying theory given the original measure is also a single construct. The first factor captures most of the variance in the data (52%) and all the items correlate well with it, as demonstrated by the loadings and their p-values set out in Table 4 below. All loadings are greater than or equal to 0.30 and have a p-value < 0.05.

Figure 1 – Scree Plot

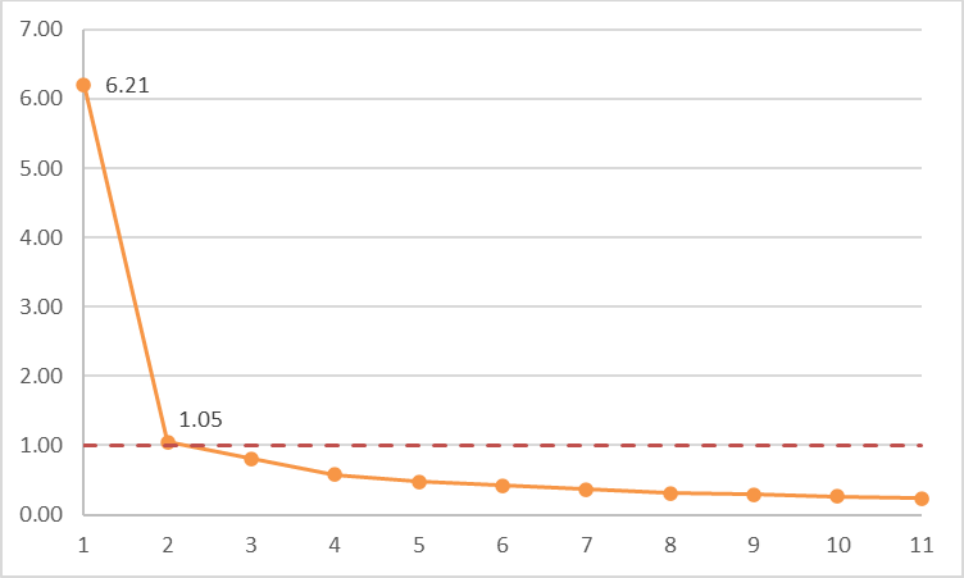


Table 4 – Factor Loadings and p-values

Item		Loading	p-value
I am confident in my knowledge of:	The level of language skills of each child that enters my class	0.39	0.000
	Reasonable goals for 3–4-year-olds in relation to the development of their language skills	0.34	0.000
	The best practices and strategies for helping 3-4-year-olds to develop their language skills	0.41	0.000
	National standards for language skills for 3-4-year-olds	0.36	0.000
	The best ways to assess the language skills (expressive language and understanding of language) of 3-4-year-olds throughout the year	0.45	0.000
I am confident in my ability to:	Gauge the level of language skills of 3-4-year-olds in my class	0.33	0.000
	Incorporate regular opportunities to learn and/or practise language skills into common preschool situations (such as art or dramatic play)	0.33	0.000
	Plan activities to help 3-4-year-olds to develop their language skills	0.37	0.000
	Support the language development of 3-4-year-olds when they make spontaneous comments/discoveries	0.30	0.000
	Help 3-4-year-olds to navigate their confusion when they are developing their language skills	0.34	0.000
	Translate assessment or screening results into curriculum plans on both a group and individual basis	0.42	0.000

3. Confirmatory Factor Analysis to assess validity/reliability

Finally, a CFA was run to assess the validity/reliability and fit of the proposed model, and to compare the latent variable to the originally proposed measure of setting-level practitioner confidence (mean total score). The internal consistency checks (Cronbach's alpha = 0.92; CR = 0.92) and model fit checks (CFI = 0.854; TLI = 0.817) all indicated a model with high validity/reliability and that it is a reasonable fit to the data. The correlation between the latent factor score and the mean total score was determined. A strong correlation between the two would indicate that the mean score captured the majority of the variance in the underlying latent variable and could therefore be used as a proxy for the latent variable. The correlation between the two measures was found to be very strong ($r^2 > 0.99$), as seen in Figure 2 below. Therefore, the mean score will be used to measure practitioner confidence in the mediation analysis.

Figure 2 – Scatter Plot of Mean Score against Factor Score