Efficacy Trial of *Talking Time©*, an oral language intervention for early years
Statistical Analysis Plan

Evaluator: NIESR
Principal investigator: Edoardo Masset

PROJECT TITLE	**Efficacy Trial of *Talking Time*©, an oral language intervention for early years**
DEVELOPER (INSTITUTION)	IOE, UCL's Faculty of Education and Society and University of Oxford
EVALUATOR (INSTITUTION)	National Institute for Economic and Social Research (NIESR)
PRINCIPAL INVESTIGATOR(S)	Edoardo Masset
SAP AUTHOR(S)	Edoardo Masset, Anisa Butt
TRIAL DESIGN	Two-arm randomised control trial with random allocation at the setting level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Age 3-4, Key Stage – Early Years
NUMBER OF SETTINGS	123
NUMBER OF CHILDREN	1,613
PRIMARY OUTCOME MEASURE AND SOURCE	Oral language skills (a composite index of Expressive vocabulary measured by the Renfrew Expressive Vocabulary Test (Renfrew,2023) and of Information and grammar measured by the Renfrew Action Picture Test (Renfrew, 2019))
SECONDARY OUTCOME MEASURE AND SOURCE	1. Expressive vocabulary, from the Renfrew Expressive Vocabulary test (Renfrew, 2023). 2. Information, from the Renfrew Action Picture Test) (Renfrew, 2019) 3. Grammar, from the Renfrew Action Picture Test) (Renfrew, 2019) 4. Sentence repetition from the Grammar and Phonology Screening (GAPS) (Gardner et al., 2006)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	12.09.2025	*N/A*

Table of contents

Introduction

Talking Time© is an oral language development intervention aimed at children aged 3 to 5. The intervention is designed to provide high quality professional development to early years practitioners (Dockrell et al., 2010). The training provided will enable practitioners to deliver structured activities in the nursery setting to support oral language development. The programme consists of engaging and structured small-group activities. Over a 20-week period, children participate in two 15-minute sessions each week as part of the provision, focusing on enhancing their communication skills through interactive methods. The intervention activities are delivered to all children in the class or room where the intervention is being implemented. Due to its focus on building staff expertise, *Talking Time*© offers a sustainable approach to improving pupil progress. Through professional development, the aim is to gradually decrease reliance on the provided programme materials, so that practitioners are empowered to adapt the plans and prompts provided, plan activities of their own; and ultimately adapt and embed the programme into regular classroom practice.

Talking Time© consists of two components: a professional development programme for practitioners in early years settings, and structured activities delivered by the practitioners to small (up to 5 children), mixed-ability groups of children. Two practitioners in each setting are identified as *Talking Time*© leads and are responsible for delivering the majority of sessions. After initial welcome visits to settings a team of trainers-mentors delivers a professional development package, which consists of:

- 3 *twilight* training sessions, provided to all staff delivering the programme, which introduces practitioners to each of the activities that form part of *Talking Time*© and the language-supporting strategies which underpin them
- 4 in-class mentoring sessions for the *Talking Time*© leads to support implementation and skill in language-supporting interactions, along with three online mentoring sessions for each team's pedagogical leader to support longer-term leadership, planning, and adaptation.
- a weekly individual professional reflection by practitioners

Settings are also provided with a manual that includes flexible plans, and conversation prompts for each activity, as well as ideas for wider classroom activities to reinforce learning. In addition, settings receive five picture books and a starter pack to support the implementation of activities. In the final weeks of the programme, it is expected that practitioners will plan the activities independently.

The timetable for the professional learning component is closely linked to the timetable for programme implementation in the classroom. This means practitioners are prepared and supported to introduce and embed each of the three *Talking Time*© activities at the relevant point in the programme.

During the programme, the *Talking Time*© Leads implement three types of playful, structured small-group activities:

- *Story Conversations*: shared storytelling and conversation using the illustrations in storybooks as prompts
- *Word Play*: games and guided role play designed to develop vocabulary breadth and depth through meaningful experiences
- *Hexagons*: narrative discussion and retelling based on photos of real situations and routines

The programme is designed to promote children's language skills. Sensitive and responsive adult-child oral language interactions, such as those promoted by *Talking Time©,* can improve children's language skills by increasing the diversity and complexity of language in early years, by talking "with" children rather than "to" children, and through a gradual transition from contextualised to decontextualised conversations (Rowe and & Snow, 2020; Rowe, 2022). Such interactions should occur within small groups in supportive learning environments (Morra Pellegrino and & Scopesi, 1990; Hassinger-Das et al., 2017), and should be guided by trained staff to support children's language growth (Dockrell et al., 2017).

Children from more disadvantaged backgrounds on average perform more poorly in standardised assessments of language skills (Nelson et al. 2011, Law et al., 2018). Attending school in a socially and economically deprived neighbourhood may affect language proficiency, and there is evidence that children learning English as a second language are at some risk of literacy difficulties (August & Shanahan, 2006; Kieffer, 2008). This can have important consequences for the remainder of their education throughout the school system. While *Talking Time*© is a universal intervention (delivered to all children in the target age range within a setting), it is expected to particularly benefit disadvantaged children, given the typically higher prevalence of oral language needs among this group. Children eligible for the Early Years Pupil Premium (EYPP), and children with English as an Additional Language (EAL), are relevant sub-groups of analysis of our study.

The study is an efficacy trial of the Talking Time© programme to provide evidence for what works to support oral language development of young children, particularly disadvantaged children.

The primary research question that this impact evaluation is designed to address is:

RQ1. What is the impact of *Talking Time*© on children's oral language skills as measured by a composite index of expressive vocabulary (Renfrew Expressive Vocabulary test) and of Information and grammar abilities (Renfrew Action Picture test)?

The following sub-questions of this (RQ1) primary research question will also be explored:

> RQ1a. What is the impact of *Talking Time*© on oral language skills of disadvantaged children that are eligible for the Early Years Pupil Premium (EYPP)?
>
> RQ1b. What is the impact of *Talking Time*© on oral language skills of children with EAL?

The secondary research questions of the study are:

RQ2. What is the impact of *Talking Time*© on different aspects of children's oral language skills as measured by the subtests of Renfrew Expressive Vocabulary Test, Renfrew Action Picture Test and the sentence repetition assessment from the GAPS test?

The following sub-questions of this (RQ2) secondary research question will also be explored:

> RQ2a. For EYPP children, what is the impact of *Talking Time*© on different aspects of oral language skills, as measured by the subtests of Renfrew Expressive Vocabulary Test, Renfrew Action Picture Test and the sentence repetition assessment from the GAPS test?

RQ2b. For children with EAL, what is the impact of *Talking Time* on different aspects of oral language skills, as measured by the subtests of Renfrew Expressive Vocabulary Test, Renfrew Action Picture Test and the sentence repetition assessment from the GAPS test

RQ4. How does the impact of the intervention vary with compliance?

The project is delivered within a larger initiative by the Department for Education's Stronger Practice Hubs (SPH), which focuses on evidence-based development in early years education. The goal of the study is to support education recovery following the pandemic, while also generating evidence on effective professorial development in the early years.

Participant selection

Setting level

The trial was open to state-maintained and private, voluntary, and independent early years settings (PVIs). Settings could only take part in one SPH programme for the 2024–2025 academic year and could not be involved in another trial that included the same children and outcomes of interest. Settings involved in other SPH-funded programmes, including the control groups for trials such as Early Talk Boost, The One Programme, Early Years Conversation Project, Concept Cat, or Communication Friendly Settings, were not eligible to participate in Talking Time.

The delivery team recruited settings from five geographic regions: North-West, Yorkshire & Humber, West Midlands, East of England, and London. The delivery team paid efforts to recruit settings from disadvantaged areas in terms of a high percentage of children with EAL or eligible for EYPP (since information on current EYPP status of children was not yet available, the team relied on records from previous years).

A total of 130 settings were recruited from selected local authorities across the four regions. After some withdrawals, 123 settings ultimately joined the study. Table 1 presents the number of settings recruited in each region alongside the number that would have been recruited proportionally to the regional population size. This comparison is provided for illustrative purposes as the sample was not designed to be representative of the regional population distribution.

Table 1 Number of settings by region

Region	Proportion of study population	Population-based number of settings	Actual recruited settings
London	0.26	32	22
North-West (inc. Y&H)	0.38	46	31
East of England	0.19	23	32
West Midland	0.18	22	38
TOTAL	1.00	123	123

Child level

Parents or carers were provided with a parent information sheet and with a consent form. Only children whose parents/carers signed the consent form were included in the baseline assessments. Settings distributed and collected consent forms from parents and confirmed

the consented status of each child through a data collection form shared with the evaluation team. Settings handed out consent forms to all parents of attending children. Settings handed out an average of 23 forms and received an average of 18 signed copies. The average return rate of consent forms per setting was 83% (for the 117 settings that kept a record of the forms sent and received).

There was no eligibility criteria related to SEND status, although it should be noted that both settings and assessors found it challenging to include children with SEND in the study. The data collection provider reported that children with SEND often completed the assessment with great difficulty, and setting staff often recommended not assessing them. Children with SEND are therefore likely to be under-represented in our study.

Finally, we set the minimum setting size to 5 consented children, and we did not conduct assessments in settings with fewer than 5 consented children. Because of the consent process and because some of the recruited nurseries were very small, some settings had very few children eligible for the intervention. In principle, settings with only 2 or 3 children could be included in a study. However, there are practical reasons for not doing so, as there are fixed costs in conducting assessments in a setting and it is not efficient carrying out very few assessments in a cluster. Statistical theory suggests that fewer than 5 observations per cluster lead to poor variance estimation when using mixed models (Eldridge et al., 2006). We therefore decided to use a minimum threshold of 5 children per cluster for inclusion in the study, although it must be noted that the occurrence never materialised.

In many EEF studies a 15-hour per week attendance threshold is used as an eligibility criterion. Children attending fewer than 15 hours per week are typically excluded from both the intervention and the evaluation. The criterion reflects the minimum free childcare entitlement provided by the government, and is used to focus the evaluation on children that do not receive other types of childcare. In addition, the MOU signed with the settings for the study requested that they facilitate obtaining parental consent from at least five children (aged 3-4 years) attending 15 hours or more of provision per week.

However, since the programme is delivered to the entire classes, excluding individual children based on attendance would have been impractical and it would have been difficult to exclude children based on attendance. We decided not to apply the eligibility criterion to the evaluation for the following reasons:

- Many settings were unable to provide data on expected children attendance at the time assessments were conducted
- Attendance data are often inaccurate due to fluctuating patterns and inaccurate reporting. Applying the criterion could have led to significant inclusion and exclusion errors.
- One of the study's core research questions is to understand how the impact of the intervention varies by attendance level, a question that could not be explored effectively if an attendance-based eligibility cut-off were imposed.

Although the delivery team made efforts to recruit settings in disadvantage areas, it became apparent at the time of conducting the baseline assessments that only few of the consented children were eligible for EYPP. We therefore decided, in settings with more than 15 children, to randomly select 15 children for the assessment after giving higher probability of selection to EYPP children. The probability of selection was set in such a way that the proportion of EYPP in the setting should correspond to the proportion of EYPP recorded in previous years.

Design overview

Trial design, including number of arms		**Two-arm randomised controlled trial with random allocation at the setting level**
Unit of randomisation		Setting
Stratification variables (if applicable)		Six geographic areas (East of England (North), East of England (East), London, North West, West Midlands, Yorkshire and the Humber)) and setting type (maintained or Private, Voluntary, Independent (PVI))
	variable	Oral language skills
Primary outcome	**measure (instrument, scale, source)**	A composite standardised index of Expressive vocabulary – Renfrew Expressive Vocabulary (REV) scored from 0 to 100 (Renfrew, 2023) and of Information and Grammar – Renfrew Action Picture Test (RAPT) Information scored from 0 to 41, Grammar scored from 0 to 39, (Renfrew, 2019)
	variable(s)	1. Sentence repetition 2. Expressive vocabulary 3. Information 4. Grammar
Secondary outcome(s)	**measure(s) (instrument, scale, source)**	1. Sentence repetition score component of the Grammar and Phonology Screening test (GAPS) scored from 0 to 11 (Gardner et al., 2006). 2. Renfrew Expressive Vocabulary (REV) scored from 0 to 100 (Renfrew, 2023). 3. Information – Renfrew Action Picture Test (RAPT) scored from 0 to 41, (Renfrew, 2019). 4. Grammar – Renfrew Action Picture Test (RAPT) scored from 0 to 39, (Renfrew, 2019).
	variable	Oral language skills
Baseline for primary outcome	**measure (instrument, scale, source)**	A composite standardised index of Expressive vocabulary and information and grammar abilities (Renfrew Expressive Vocabulary (REV) scored from 0 to 100 (Renfrew, 2023) and of Information and Grammar – Renfrew Action Picture Test (RAPT) Information scored from 0 to 41, Grammar scored from 0 to 39, (Renfrew, 2019))
Baseline for secondary outcome	**variable**	1. Sentence repetition 2. Expressive vocabulary 3. Information 4. Grammar
	measure (instrument, scale, source)	1. Sentence repetition score component of the Grammar and Phonology Screening test (GAPS) scored from 0 to 11 (Gardner et al., 2006). 2. Renfrew Expressive Vocabulary (REV) scored from 0 to 100 (Renfrew, 2023).

		3. Information – Renfrew Action Picture Test (RAPT) scored from 0 to 41, (Renfrew, 2019). 4. Grammar – Renfrew Action Picture Test (RAPT) scored from 0 to 39, (Renfrew, 2019)

The trial was designed as a two-arm, setting-level randomised trial, where the settings are the units of randomisation. We randomised settings to either the intervention or to the control condition with equal probability of 50%.

We randomised the settings within 11 strata consisting of 6 geographic regions (London, North West, Yorkshire and Humber, East of England (North), East of England (East), and West Midlands) split by PVI status (in one of the strata there were no PVI and therefore we end up with 11 strata rather than 12). This deviates from the protocol in which there were 5 regions rather than 6. The delivery team requested that the East of England should be further divided into North and East subregions because this would help project delivery from a logistical perspective. This led to some very small strata, and in order to maximise balance between arms we decided to adopt a restricted randomisation procedure (Hayes & Moulton, 2017), which is described in Appendix 2.

We performed randomisation in the week of the 4th November 2024. This is the week before the beginning of the intervention, which started officially on the 11th of November 2024. A total of 123 settings were randomised, of which 62 were assigned to the intervention group and 61 to the control group not receiving the intervention (details of the randomisation procedure are in Appendix 2). The evaluation PI conducted the randomisation and shared the results with the delivery team to check consistency with the agreed geographical distribution of the settings. The delivery team communicated the outcome of randomisation to the settings the week after it was conducted.

The primary outcome of this efficacy trial is oral language skills. Oral language is a multidimensional general construct that cannot be fully captured through a single measure. Language competency encompasses several components including grammar, vocabulary, phonology, and narrative discourse (Massonnié et al., 2022). The intervention targets oral language development across these domains, with the exception of phonology. While examining each component separately provides valuable insights into which specific skills are enhanced by the intervention, its overall effect on language competency can only be asserted using a measure that integrates these multiple dimensions.

We use therefore a composite index of independent assessments of expressive vocabulary and expressive language skills.

- Expressive vocabulary is measured using the Renfrew Expressive Vocabulary Test (REV) (Renfrew, 2023). The test evaluates a child's ability to correctly name pictures of words arranged by difficulty level. In our pilot the test took an average of 7:55 minutes to complete. The results are scored from 0 to 100.
- Expressive language skills are measured using the *Renfrew Action Picture Test* (*RAPT*) (Renfrew, 2019). RAPT assesses children's grammar skills by asking them to describe a series of illustrated scenes, incorporating words, verbs, and increasingly complex grammatical structures. The children's grammar outcome provides a proxy for quality of connected speech and is scored between 0 and 39. RAPT also measures the amount of relevant information conveyed by the children and their ability to describe actions and details in the pictures. The information score provides an

additional measure of vocabulary knowledge and is scored between 0 and 41. In our pilot, the test took an average of 4 minutes to complete.

The two Renfrew assessments take altogether about 15 minutes to complete and were selected to capture the impact of the intervention on different language skills. More specifically, vocabulary is targeted by the *Word Play* games; comprehension is targeted by the *Story Telling* sessions; and narrative skills are targeted by the *Hexagons* sessions. Plans to administer an additional test -the Renfrew Bus Story assessment (Renfrew, 2010) a reduced version of the same test called "Narrative Task" – were dropped after a pilot testing showed that most children were not able to answer the questions. The pilot also revealed that the assessments results were difficult to score without access to a recording on the assessment. We therefore decided to audio-record all the assessments to allow the assessor to more accurately score the results. This in turn led to the need to obtain parental consents from parents, which had important implications in terms of logistical implementation and sample size.

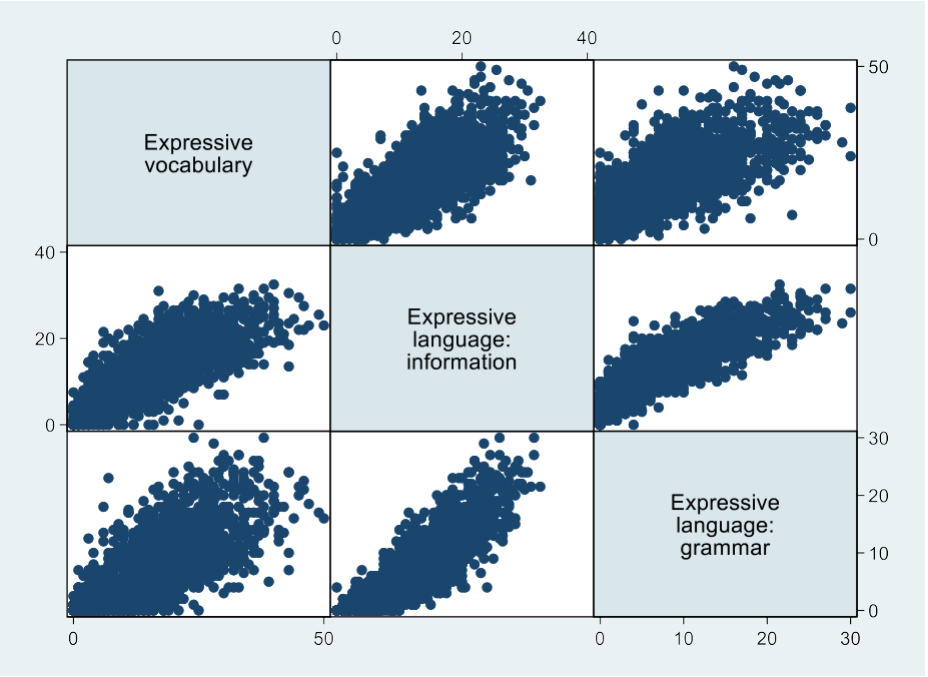
There is no established theory to guide how different indicators of oral language—each representing a distinct dimension—should be combined, and we cannot assign weights to these components based on prior knowledge. Therefore, we will use Principal Component Analysis (PCA) to extract the main underlying factor from the various indicators. This principal component will serve as the basis for constructing an index of oral language skills. PCA is specifically designed to uncover latent constructs, making it particularly suitable in situations like this, where the relative importance of different components is unknown.

The three components of the two Renfrew tests are strongly and linearly correlated to each other as shown in the correlation matrix of Table 2 and in the scatterplot matrix of Figure 1.

Table 2 Correlation matrix of the Renfrew assessments

	Exp. Voc. (REV)	Exp. Lang. information (RAPT)	Exp. Lang. grammar (RAPT)
Exp. Voc. (REV)	*1*		
Exp. Lang. information (RAPT)	*0.7360*	*1*	
Exp. Lang. grammar (RAPT)	*0.6851*	*0.8583*	*1*

Figure 1 Scatterplot matrix of the Renfrew assessments



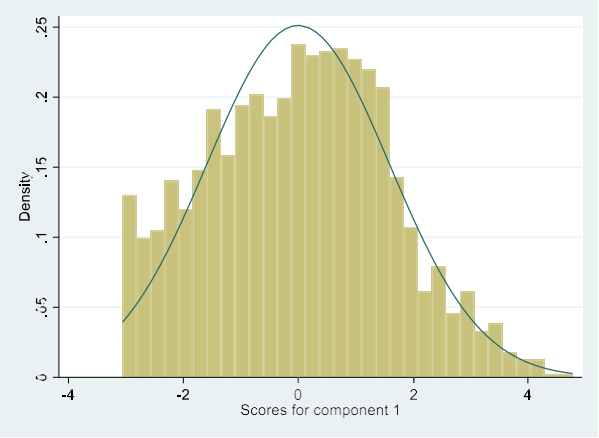
We constructed a composite index of the Renfrew assessments using Principal Component Analysis (PCA) applied to the baseline assessment data. Table 3 presents the eigenvalues and the proportion of variance explained by each component. The first principal component accounts for 84% of the total variance, providing a strong summary measure of the three assessments.

Table 3 Principal components of the Renfrew assessments

	eigenvalues	proportion of variance explained	cumulative
Component 1	2.52	0.84	0.84
Component 2	0.34	0.11	0.95
Component 3	0.14	0.05	1.00

We then standardised the Renfrew assessments to have a mean of zero and a standard deviation of 1. We estimated the following factor loadings/weights for the first principal component: 0.55, 0.60, 0.58. We applied the weights (these are the same as the loadings in this case) to the three standardised scores to obtain an overall PCA score. Since the weights are very similar to each other, the PCA index is very similar to the arithmetic mean. The two are nearly identical with a correlation of 0.97. The chart of Figure 2 shows the distribution of the PCA index. The index is normally distributed, except at the lower tail, where there are more extremely low scores than a normal distribution would predict.

Figure 2 Distribution of the PCA composite index of oral language at baseline



The secondary outcomes of the study are oral language skills measured in their various dimensions by the Renfrew assessments described above, and by the Grammar and Phonology Screen test (*GAPS*) sentence repetition scale (Gardner et al., 2006). *GAPS* is designed for children aged 3 years and 4 months to 6 and a half years. It evaluates fundamental grammatical skills across sentences and words. The test can be administered by both professionals and non-professionals, typically taking 5 to 10 minutes to complete. During our pilot it took an average of 5:20 minutes to complete.

In our study we will only use the Grammar component of the assessment (sentence repetition), because phonology is not targeted by the intervention. The grammar component of GAPS includes 11 questions. The child is asked to repeat a short series of sentences after the administrator's prompt, with the support of a visual aid booklet story. The test is scored from 0 to 11.

Sample size calculations overview

	Protocol		Randomisation	
	OVERALL	EYPP	OVERALL	EYPP
Minimum Detectable Effect Size (MDES)	0.17	0.29	0.20	0.36
Pre-test/ post-test correlations level 1 (child)	0.40	0.40	0.40	0.40
level 2 (setting)	-	-	-	-
Intracluster correlations (ICCs) level 2 (setting)	0.10	0.10	0.10	0.10
Alpha	0.05	0.05	0.05	0.05
Power	0.80	0.80	0.8	0.8
One-sided or two-sided?	Two-sided		Two-sided	
Average cluster size	15	3	13	2
Number of settings intervention	65	65	62	62
control	65	65	61	61
total	130	130	123	123
Number of children intervention	975	195	977	168
control	975	195	946	175
total	1,950	390	1,923	343

The sample size calculations were conducted with *Optimal Design Version 3.01*, assuming a significant pre-post intervention correlation, and a moderate correlation between pupils within settings. More specifically the assumptions and parameters were the following:

- Intervention and control groups of equal size
- A statistical level of significance (α) of a two-tail test in the difference between means of 0.05
- Statistical power (β) of 0.80 (representing 80% of chances of finding a statistically significant effect if there is indeed an effect)
- A pre-post intervention correlation between assessment scores of 0.40 (in accordance to results of EEF research (Singh et al., 2023), which found the correlation between pre and post-test English assessments in Early Years Foundation Stage and Key Stage 1 in the range 0.43-0.63 using the NPD data, and in the range 0.54-0.88 using the data from past EEF-funded trials - we adopt the most conservative of these estimates)
- An intra-cluster correlation coefficient (ICC) of 0.10 (EEF research suggests an ICC in early years trials at around 0.06 for studies where English is the outcome measure (Singh et al., 2023) - we adopt a more conservative value of 0.10)
- The protocol assumed a 25% sample loss due to attrition and missing values that would reduce the sample from 20 observations per setting to 15 observation per setting (of which 3 EYPP).

- A proportion of children eligible for EYPP of 20% (in 2023, 14 per cent of children aged 3-4 old not in reception were in receipt of EYPP (DfE, 2023), which we increased to 20% to account for the targeting of disadvantaged areas at recruitment stage)

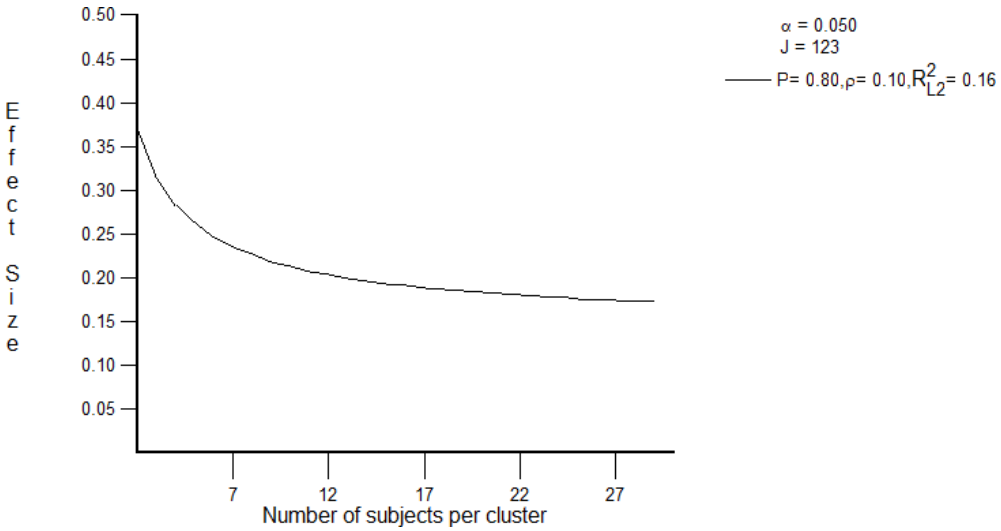
This set of assumptions delivered a minimum detectable effect size (MDES) of 0.17SD for the whole sample and of 0.29 for EYPP children at protocol stage. The study was therefore reasonably powered to detect an impact of the intervention in the recruited settings, but it was not sufficiently powered to detect an impact on the EYPP subgroup. The study was also able to withstand the loss of a significant number of settings. At protocol stage we estimated that a 10% setting level drop-out rate (equivalent to 13 settings) would result in an MDES of 0.20, a 20% drop out rate (26 settings) would result in an MDES of 0.22, a 30% drop out rate (39 settings) in an MDES of 0.23.

Active recruitment started on 29th January of 2024 and officially ended on 29th July 2024. Recruitment was extended over the period from August to September 2024 to allow for the replacement of early withdrawals. A total of 2,412 settings were approached to take part in the study. Of these, 644 settings attended webinars to share information on the project, which reflect initial interest in the intervention. Of these 316sent an expression of interest for participating in the programme and 140 signed a memorandum of understanding. However, 17 settings withdrew from the study by the time data collection started in mid-September 2024. Ultimately, 123 settings shared their pupil data with the evaluation team and joined the study. Of these, 97 were maintained settings and 26 were PVIs. In addition, the average number of children per setting turned out to be 13, rather than 15, as assumed at protocol stage. This implies a reduction in the sample of 18% between protocol and randomisation.

After randomisation we applied the same assumptions used at the protocol stage to the new sample. The power calculations delivered an MDES of 0.20SD for the overall sample and 0.36SD for the EYPP eligible group. This indicates that the study is adequately powered to detect an impact on the overall sample, but it is not adequately powered to detect an impact on the sub-sample of EYPP children. The study is also sufficiently powered to withstand setting dropouts, as the MDES would only marginally increase. We estimate that a loss in settings by 10% (equivalent to 12 settings) would increase the MDES to 0.21SD, a loss by 20% (equivalent to 25 settings) would increase MDES to 0.22, and a loss by 30% (equivalent to 37 settings) would increase MDES to 0.24.

The size of the sample is likely to be smaller at the end of the evaluation because of further missing data at the follow-up, and because of attrition between the two waves. The chart in Figure 3 shows that 10 children per setting represent a critical threshold for our study. Fewer than 10 children per setting would not be able to detect an impact smaller than 0.20 SD. On the positive side, the power estimates presented here are conservative in two ways. First, we assumed no effect by blocking variables. However, the sample was stratified by region, and it is likely that the regional stratification explains some of the variation in test scores, thus improving the precision of the estimates. Second, test scores are computed using audio recordings to improve the quality of the data. Data quality improves statistical power in two ways. First, a lower measurement error implies a smaller sample standard deviation and therefore implies a larger detectable effect size (which is expressed in terms of standard deviation units). Second, a lower measurement error also implies a stronger correlation between pre- and post-intervention assessments, which in turn implies a higher statistical power.

Figure 3 MDES by setting size



Analysis

In this section, we outline our statistical approach to the analysis. In both the primary and secondary outcome analyses, we will estimate the intention-to-treat effect (ITT) of the intervention, i.e. the average impact of the intervention on the population, regardless of programme compliance or treatment dosage. All children assessed at baseline will be included in the analysis, whether or not they actively participated in the intervention or dropped out during implementation. Methods for analysing the impact of the intervention with imperfect compliance are discussed in the "Compliance" section.

Primary outcome analysis

To address the primary research question (RQ1), we will estimate a multi-level model of oral language skills We will assess the impact of the intervention on a standardised composite index of the Renfrew tests by comparing the changes in scores over time between the intervention and the control groups, using an ANCOVA (analysis of covariance) model, in which the post-intervention assessment scores are regressed against the pre-intervention assessment scores. The model includes only the pre-intervention assessment scores, the treatment indicator, and the strata indicators. In the robustness analysis section below, we present a "saturated" model including some pre-intervention characteristics at the child level and at setting level (Rubin, 2008).

We will estimate the effect size using a multi-level random intercept model. This model is preferable to the OLS fixed effects model discussed in the protocol, because the OLS model does not control for unobserved heterogeneity between clusters.

The multi-level random intercept model is:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \alpha_u + \varepsilon_{iu} \qquad (1)$$

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- T_{ui} is treatment status for child *i* in setting *u* (*T=1* if the child is in the intervention group)
- β_2 is the estimated impact of the intervention
- S_{si} are indicators for the strata used at randomisation
- α_u are setting-level random-effects
- ε_{iu} is a child level error term

The output of the model is divided in two parts (see Appendix 4 for various examples). The first part includes the fixed effects: the coefficient estimate of the pre-intervention assessment score (β_1), the coefficient estimate of intervention variable (β_2), and 10 coefficients of the stratification variables (γ_s). The second part of the output reports the random effects, which include an estimate of the variance of baseline test scores across settings (the random intercept), and an estimate of the variance of scores within settings.

In model (1) the outcome is a composite index of the REV and RAPT assessments For the analysis we use the simple arithmetic mean of the Renfrew test scores at both baseline and follow-up, rather than the PCA index. As shown in the Design overview section, the PCA index produces results that are nearly identical to the arithmetic mean. Using the mean also avoids the complication of outcomes being measured on different scales at baseline and follow-up when relying on PCA.

Secondary outcome analysis

In the secondary outcome analysis, we will estimate the impact of the intervention on the GAPS score and on the individual Renfrew test scores (REV and RAPT). We do not standardise the scores at this stage (we discuss standardisation of the effects of the intervention in the “effect size calculation” section). The analysis of the secondary outcomes will be carried out in the same way as the analysis of the primary outcome. We will use the multi-level random intercept:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \alpha_u + \varepsilon_{iu} \qquad (2)$$

Where:

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- T_{ui} is treatment status for child *i* in setting *u* (*T=1* if the child is in the intervention group)
- β_2 is the estimated impact of the intervention
- S_{si} are indicators for the strata used at randomisation
- α_u are setting-level random-effects
- ε_{iu} is a child level error term

We will separately assess the impact of the intervention on the REV and RAPT and GAPS scores using model (2). As mentioned, oral language is a multidimensional construct, and different components of the intervention target distinct aspects of language. This analysis will

help identify which specific skills were most positively influenced by the intervention. Moreover,

because the activities in the Talking Time programme focus on different language domains, this approach will also shed light on which activities are most effective in improving particular aspects of oral language.

Subgroup analyses

Children from more disadvantaged backgrounds on average perform more poorly in standardised assessments of language skills (Nelson et al. 2011, Law et al., 2018). Attending school in a socially and economically deprived neighbourhood may affect language proficiency, and there is some evidence that children learning English as a second language are at some risk of literacy difficulties (August & Shanahan, 2006; Kieffer, 2008).

As detailed in the protocol, this efficacy study has two further research questions concerning the impact of Talking Time© on children's oral language skills for two subgroups:

- What is the impact of *Talking Time*© on oral language skills of disadvantaged children that are eligible for the Early Years Pupil Premium (EYPP)?
- What is the impact of *Talking Time*© on oral language skills of children with EAL?

For conducting the subgroup analyses we will adopt two approaches: 1. Estimating the intervention's impact separately for children within and outside the subgroup, and 2. Estimating the impact of the intervention by including an interaction term for subgroup status. In both cases, we will apply the same model specification used for the primary analysis, which is a random-effects model with both random intercepts.

The model specification for each approach is described below.

1. Separate regressions

We will estimate the following equation separately for children within the subgroup and outside the subgroup *h*:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \alpha_u + \varepsilon_{iu} \qquad \text{if subgroup=}h(0,1) \qquad (3)$$

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- T_{ui} is treatment status for child *i* in setting *u* (*T=1* if the child is in the intervention group)
- β_2 is the estimated impact of the intervention
- S_{si} are indicators for the strata used at randomisation
- α_u are setting-level random-effects
- ε_{iu} is a child level error term

2. Interaction term model:

In the interaction term model, we will use the full sample of children in the study, including an indicator for subgroup status and the interaction of subgroup status and intervention assignment, as shown in the equation below:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \partial_1 subg_i + \partial_2 subg_i * T_{ui} \quad + \alpha_u + \varepsilon_{iu} \qquad (4)$$

Where:

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- T_{ui} is treatment status for child *i* in setting *u* (*T=1* if the child is in the intervention group)
- β_2 is the estimated impact of the intervention
- S_{si} are indicators for the strata used at randomisation
- $subg_i$ is subgroup status for child *i* ($subg_i$=1 if the child is either eligible for EYPP or with EAL)
- $subg_i * T_{ui}$ is the interaction between subgroup status and treatment status
- α_u are setting-level random-effects
- ε_{iu} is a child level error term

A coefficient (∂_2) for the interaction term that is statistically different from zero implies that the impact of the intervention is different for that specific subgroup. The impact of the intervention for the subgroup is also reported by the coefficient β_2 in model (3). In principle, the impacts for the subgroup effects estimated by the two models should be similar and:

$$\beta_{2(model3)} \approx \beta_{2(model4)} + \partial_{2model4}$$

However, the results of the models (3) and (4) can be different when using multi-level models because the subgroup-specific models use the random-effects structure specific to the particular subgroup. In general, the results of model (4) are preferable and include all information needed to estimate the impact of the intervention on the subgroup.

Additional analyses

Robustness analysis

In our robustness analysis we will compare the results of the multi-level random intercept model (1) to an OLS model and to the multi-level random intercept and random slope model. We will estimate an OLS model with fixed effects, which is widely used in the literature and that was originally proposed in the protocol. This model offers a useful benchmark to which compare the results of the multilevel model.

The OLS model with clustered standard errors is:

$$y_{uit} = \beta_1 y_{uit-1} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \varepsilon_{iu} \qquad (5)$$

where:

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- T_{ui} is treatment status for child *i* in setting *u* (*T=1* if the child is in the intervention group)
- β_2 is the estimated impact of the intervention
- S_{si} are 10 indicators for the strata used at randomisation

- ε_{iu} is a child level error term

The multi-level random intercept and random slope model is:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \alpha_u + \alpha_u T + \varepsilon_{iu} \qquad (6)$$

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- T_{ui} is treatment status for child *i* in setting *u* (*T=1* if the child is in the intervention group)
- β_2 is the estimated impact of the intervention
- S_{si} are 10 indicators for the strata used at randomisation
- α_u are setting-level random-effects
- $\alpha_u T$ are interactions between the random intercept and treatment status
- ε_{iu} is a child level error term

While the random intercept model (1) assumes that project effects do not vary across settings, the random slope model (6) allows the slope to vary across settings. The random slope model allows the impact of the intervention to vary across settings, which may happen, for example, as a result of differences in implementation or in the characteristics of the population. For example, settings have different capacity to implement the interventions, and different population groups may respond in different ways to the intervention.

The output of the random slope model (6) is divided in two parts (see Appendix 4 for various examples). The first part includes the fixed effects: the coefficient estimate of the pre-intervention assessment score (β_1), the coefficient estimate of intervention variable (β_2), and 10 coefficients of the stratification variables (γ_s). The second part of the ouputs reports the random effects, which include an estimate of the variance of baseline test scores across settings (the random intercept), an estimate of the variance of the project effects across setting (the random slopes, i.e. the interaction between the treatment status and settings), and an estimate of the variance of scores within settings. To compare model 1 and model 5 we will conduct a likelihood ratio test to assess whether the random slope and random intercept model (6) is to be preferred to the simple random intercept model (1) (see the example in Appendix 4).

Finally, we will run "saturated" versions of the models 1 through 6 to include pre-intervention child-level and setting-level characteristics:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \sum_{k=1}^{n} \theta_k X_{kit_{-1}} + \sum_{j=1}^{m} \vartheta_j Z_{jit_{-1}} + \alpha_u + \varepsilon_{iu} \qquad (1a)$$

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \sum_{k=1}^{n} \theta_k X_{kit_{-1}} + \sum_{j=1}^{m} \vartheta_j Z_{jit_{-1}} + \alpha_u + \varepsilon_{iu} \qquad (2a)$$

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \sum_{k=1}^{n} \theta_k X_{kit_{-1}} + \sum_{j=1}^{m} \vartheta_j Z_{jit_{-1}} + \alpha_u + \varepsilon_{iu}$$

if subgroup=*h(0,1)* (3a)

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \partial_1 subg_i + \partial_2 subg_i * T_{ui} + \sum_{k=1}^{n} \theta_k X_{kit_{-1}} + \sum_{j=1}^{m}$$

$\vartheta_j Z_{ji}$ $+ \alpha_u +$
$t-1$ ε_{iu}

(4a)

$$y_{uit} = \beta_1 y_{uit-1} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \sum_{k=1}^{n} \theta_k X_{kit-1} + \sum_{j=1}^{m} \vartheta_j Z_{jit-1} + \varepsilon_{iu} \quad (5a)$$

$$y_{uit} = \beta_1 y_{uit-1} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \alpha_u + \alpha_u T + \sum_{k=1}^{n} \theta_k X_{kit-1} + \sum_{j=1}^{m} \vartheta_j Z_{jit-1} + \varepsilon_{iu} \quad (6a)$$

where coefficients and variables have the same meaning as in the models from 1 through 6 with the inclusion of:

- X_{kit-1} are *n* child-level variables recorded at the baseline: age in months, gender, EYPP status, and EAL status. These variables are predictors of test scores and were employed in our restricted randomisation. Age, gender, and EYPP/EAL were collected from the settings at the baseline
- Z_{jit-1} are *m* setting-level variables recorded at the baseline. They include the IDACI index[1] and an indicator of staff capacity.

The goal of including these additional variables is to improve the balance between the units assigned to the project and to the control group, and to increase the precision of the estimates (Rubin, 2008).

Eligibility threshold

At baseline, settings were asked to complete a spreadsheet detailing each child's expected attendance by day and session (morning/afternoon). At the endline, they will be asked to update this information. In addition, the delivery team will collect attendance data specifically related to *Talking Time*© sessions. This attendance data will be used to conduct additional robustness analyses. We will report the estimated effects of the intervention on the primary outcomes for the subgroup of children meeting the 15-hour per week attendance threshold and compare the results to those obtained for the full sample.

Imbalance at baseline

We will assess baseline balance by comparing the means of continuous variables (standardised when required) and percentages of binary variables of: primary outcomes and other setting-level and child-level characteristics. We will report count and percentages in the case of categorical variables and means and standard deviations in the case of continuous variables. In the case of dichotomous variables, we will report the per cent difference between the means in the two arms. In the case of continuous variables, we will report the standardised difference in the means between the two arms. We will calculate standardised scores using the pooled standard deviation:

$$sd_{pooled} = \sqrt{\frac{sd_1^2(n_1-1) + sd_2^2(n_2-1)}{n_1+n_2-2}} \quad (7)$$

[1] The IDACI index is an area-level index of economic disadvantage. It is produced by The Ministry of Housing, Communities and Local Government (MHCLG) and its most recent release was in 2019. It is calculated for small geographical units which contain approximately 1,500 residents or 650 households.

It measures the proportion of children aged 0-15 that live in deprived households in a given area.

Where sd^2_1 and sd^2_2 and n_1 and n_2 are the variances and the sample sizes of the project and control group respectively. The standardised variables are computed as:

$$std_{var} = \frac{y_i - \bar{y}}{sd_{pooled}}$$

We will then report the sample difference in the standardised scores and the standard error of the difference.

We will use the table below (from the EEF Evaluation Report template) to report the results, and we will include values for the following variables: child's gender, child's age, EYPP status, EAL status, setting-level IDACI index, language screen score, and GAPS, REV, and RAPT scores. The covariates were selected among the available child-level and setting-level data, which predict the outcomes and that were used in our restricted randomisation.

Setting-level (categorical)	National-level mean	Intervention group		Control group		
		n/N (missing)	Count (%)	n/N (missing)	Count (%)	Difference
IDACI index						
Setting-level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Standardised difference
Pupil-level (categorical)		n/N (missing)	Count (%)	n/N (missing)	Count (%)	Difference
Female child						
EYPP						
EAL						
Pupil-level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Standardised difference
Child's age in months						
GAPS score						
REV score						
RAPT score						

Missing data

We define missing data as data that we intended to collect but that for various reasons we did not collect. In this section, we focus on missing data related to the outcome variable (test scores) at either baseline or endline, as this is the most likely source of missingness. In contrast baseline setting covariates and demographic characteristics of children, such as age, gender, EYPP and EAL status, are relatively easy to obtain and are not expected to be missing in significant numbers.

We identify two problems with missing data:

- **Statistical power**. Missing data reduce sample size and the statistical power of the study. This is particularly true if the proportion of missing data is large. For example, a reduction of the sample by more than 10% will have implications on the size of the minimum detectable effect size.

- **Internal validity**. Non-random differences between the intervention and the control group produced by missingness can lead to biased estimates of the project effect. For example, if children eligible for EYPP are systematically missing from the intervention group, and if EYPP eligibility is correlated with test scores, then the impact of the intervention will not be correctly estimated.

In general, if the extent of missing data is very small (say under 5%), all the problems above have a minor impact on the impact estimates. On the other hand, the larger is missingness, the bigger the two problems above become. There is no general rule to tell what it is an acceptable level of missingness, and for simplicity, we will use here the 5% threshold recommended by the EEF (2022)

Datasets with missing data can be analysed and interpreted in different ways depending on our knowledge or assumptions about the mechanism generating missingness. Normally a distinction is made between three types of missing data (Rubin, 1976): missing completely at random (MCAR); missing at random (MAR), and missing not at random (MNAR).

Data are MCAR, when the missingness is totally uncorrelated with the outcome variable. For example, some assessors may be less efficient at conducting the tests and might complete fewer tests in the allocated time. It is safe to assume that the allocation of assessors to setting is uncorrelated with the outcome of the assessments. In this case the analysis of the data is unaffected by missingness.

Data are MAR when there is an association between missingness and observable characteristics, but after conditioning on these characteristics, missingness is random. For example, missingness of post-intervention test scores could be related to EYPP eligibility, but conditional on EYPP (that is within groups of EYPP and non-EYPP children) missingness is random. In these circumstances the results of the analysis are potentially invalid, but we can safely analyse the data if we are able to control for EYPP eligibility.

Data are MNAR when there is an association between missingness and unobservable characteristics that are correlated with the outcome. For example, for some unknown reasons, children in the intervention group that are expecting to perform poorly in the test, may systematically refuse to be tested. In this case the impact analysis would overestimate the

effect of the intervention, because poorly performing children would be systematically absent from the intervention group. Since we cannot observe children's expectations (or predict them) we are not able to condition the analysis on the variables affecting the outcome as in the MCAR case, and we need to employ more sophisticated methods.

Our approach to deal with missing data will depend on whether the missing data can be defined as MCAR, MAR, or MNAR. Unfortunately, there is no test that can tell us whether data are MCAR or MAR.

Concretely, we will start by assessing the extent of missingness in the data. If missingness is larger than 5% we will follow two courses of action. The first consists of analysing the data to discern any pattern in missingness. The second, depending on the type of missingness identified, consists of adjusting the analysis under different assumptions about the missingness of the data and then compare the results.

In relation to the first course of action we will do the following:

- Compute the number of missing cases at the baseline and at the endline for the outcome variable
- Tabulate the reasons for not taking the assessment as reported by the assessor conducting the assessment
- Compare the proportion of missing cases in the project and in the control group at the baseline and at the endline
- Identify variables correlated with missingness by running a logit regression of missingness on relevant variables at the baseline and at the endline

$$log\left(\frac{P(y-1)}{1-P(y-1)}\right) = \beta_0 + \beta_1 T_{ui} + \sum_{k=1}^{n} \beta_k X_{ki} \quad (8)$$

Where:

- *y* is an indicator variable which is equal to 1 if the observation is missing and 0 otherwise
- *P* is the probability that the observation is missing
- T_{ui} is treatment status of child *i* in setting *u*, and β_1 is the change in the log-odds if the child is in the treatment group
- X_{ki} is a set of *k* child-level variables (gender, age in months, EYPP status, EAL status) and setting-level variables (IDACI index)

If we find no correlations between the covariates and missingness, and if we find no differences in missingness in the project and the control group, we may tentatively assume that the data are MCAR. If we find a correlation between the covariates and missingness but no differences in missingness between the project and the control group, we may tentatively conclude that the covariates fully explain differential missingness, and that data are MAR. If we find correlations between covariates and missingness, and differences in missingness in the project and the control group, then we may tentatively assume that the data are MNAR. In this case we will further explore the nature of missingness in the project and in the control groups separately, to see whether the determinants of missingness differ in the two groups. We will employ the following models, in which variables and coefficients have the same meaning as those described in model (7).

$$log\left(\frac{P(y-1)}{1-P(y-1)}\right) = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki} \qquad \text{if treatment=1}$$

$$log\left(\frac{P(y-1)}{1-P(y-1)}\right) = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki} \qquad \text{if treatment=0}$$

After running the two models separately we can test whether the coefficients of the two models are different one-by-one, and if they are all simultaneously different. If we find that the coefficient estimates differ systematically between the project and the control group, we have more reasons to tentative conclude that data are MNAR.

After conducting the analysis above we will treat the data as either MCAR, MAR, or MNAR and we will carry out the primary analysis in the following way:

1. Assuming data are MCAR, we will carry out no adjustment and will simply omit the missing cases from the analysis. We will delete cases with missing values and estimate primary outcomes using the ANCOVA model (1).

2. Assuming data are MAR, we will omit the missing cases from the data, and we will adjust the analysis conditioning for the variables determining missingness by including these variables in model (1). In addition, we will employ multiple imputation (Rubin, 1996), using the available data to predict the missing observations, and thus exploring the uncertainty of estimates resulting from missing data.

3. Assuming data are MNAR, we will employ a Heckman selection model (Maddala, 1983) to correct for the selection bias produced by the missing variables. We will adopt this approach only after thoroughly assessing the plausibility of the MAR assumptions as discussed above, and only if a valid instrument is found (see below). In addition, we will conduct the sensitivity analyses recommended by Van Buuren for non-ignorable models (2018).

The Heckman selection model is perfectly suited to address selection bias in datasets with missing data (Muñoz et al., 2023). The model adjusts estimates for selection bias produced by both observable and unobservable characteristics. It requires the estimation of two equations. The first equation (the selection equation) estimates the probability of being in the sample (not missing) given observed characteristics. The results of the selection equation are employed to calculate an inverse Mills' ratio. The second equation estimates the impact of the intervention on the outcome including the inverse of the Mills' ratio among the explanatory variables. The coefficient of the Mill's ratio in the outcome equation, at a time tests the presence of selection bias and corrects the estimates for selection bias.

The estimation of the selection equation is a key element of the Heckman model. The ability to adjust for selection bias depends on our ability to explain missingness and to approximate a random experiment (the 'exclusion restriction'). To meet the exclusion restriction, the selection model should preferably include at least one variable which is (preferably strongly) correlated with missingness but uncorrelated with the outcome. This instrumental variable introduces in the model an element of randomness that allows model identification. Unfortunately, credible instrumental variables that meet the exclusion restriction are hard to find. In our analysis we will explore the characteristics of data collection to identify potential instruments, such as the day of assessment or the assigned assessors, that can be correlated with missingness but not with the outcome.

In the first step we estimate the probability that the observation is not missing using a probit model (the selection equation):

$$P(y_i = 1) = \beta_0 + \beta_1 T_{ui} + \sum_{k=1}^{n} \beta_k X_{ki}$$

Where:

- $y_i = 1$ if the observation is not missing and $y_i = 0$ if the observation is missing
- T_{ui} is treatment status of child *i* in setting *u*, and β_1 is the change in probability of the observation being missing if the child is in the treatment group
- X_{ki} is a set of child-level variables (gender, age in months, EYPP status, EAL status) and setting-level variables (IDACI index) – the variables will ideally include a variable that is correlated with missingness but not with the outcome (we will experiment with assessors and day of assessment)

After estimating the model above, we use the predicted values to calculate for each observation the inverse Mills ratio:

$$\lambda_i = \frac{\varphi(P(\hat{y_i} = 1))}{\Phi(P(\hat{y_i} = 1))}$$

In the second step we estimate the multi-level outcome model:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \beta_3 \lambda_i + \sum_{k=1}^{n} \beta_k X_{kit_{-1}} + \sum_{j=1}^{m} \gamma_j Z_{jit_{-1}} + \alpha_u + \alpha_u T + \varepsilon_{iu} \qquad (9)$$

Where the variables are the same as those included in the "saturated" models with the inclusion of the inverse Mills ratio:

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- T_{ui} is treatment status for child *i* in setting *u* (*T*=1 if the child is in the intervention group)
- β_2 is the estimated impact of the intervention
- λ_i is the inverse Mills' ratio obtained from the selection equation
- $X_{kit_{-1}}$ are *n* child-level variables recorded at the baseline: age in months, gender, FSM status, EAL status.
- $Z_{jit_{-1}}$ are *m* setting-level variables recorded at the baseline: the strata used at randomisation, and the IDACI index.

- α_u are setting-level random-effects
- $\alpha_u T$ are interactions between the random intercept and treatment status
- ε_{iu} is a child level error term

The inclusion of the Inverse Mills ratio tests the presence of selection bias due to missingness in the data, and it adjusts the estimation of the impact of the intervention for selection bias due to missingness.

Floor effects

The administration of the RAPT and REV tests did not present any particular issues at baseline. We examined the proportion of zero scores in the sample and compared these with the proportion expected under a normal distribution (calculated as $\Phi\left(\frac{0-\mu}{\sigma}\right)$). The observed and expected values are as follows: for RAPT information, 3.2% observed versus 2.4% expected: for RAPT grammar 10.6% observed versus 8.6% expected; and for REV 2.3% observed against 3.1% expected. These deviations are minor and suggest that the proportion of zeros scores does not substantially depart from what would be expected under normality. Importantly, these zero scores appear to be genuine and not the result of test termination by the child or by the assessor. They are not censored observations. As such, we retained all zero scores in the construction of the composite PCA index of the RAPT and REV assessments, without applying any further adjustments.

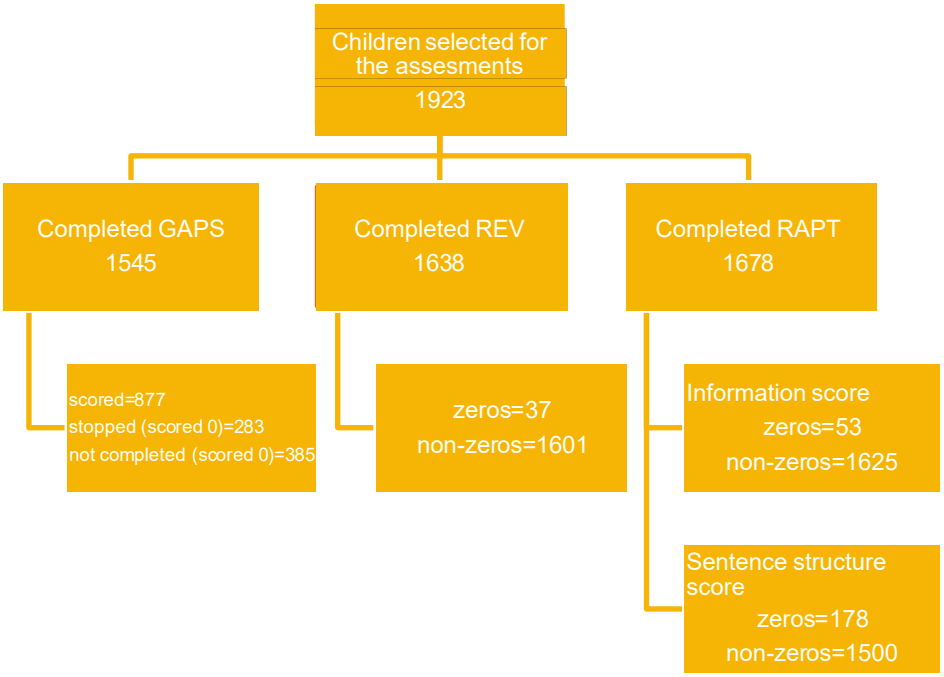
On the other hand, the administration of the GAPS test proved difficult, and we collected data with many zero values and missing values. This poses some serious of problems in terms of statistical power, bias of the estimates, and data analysis. In the following we discuss how we plan to address the issues related to floor effects in the GAPS tests

Adjusting for floor effects in the GAPS assessment

The administration of the GAPS assessment proved to be difficult, and many children were not able to complete the test at the baseline. Figure 4 shows the completion rates for the three assessments (GAPS, REV, and RAPT). Initially, a total of 1,923 children were selected for the three assessments. However, only 1,545 completed the GAPS test (80.3%), 1,638 completed the REV test (85.2%), and 1678 completed the RAPT test (87.3%). Reasons for not completing the tests included: absence; the child having to leave; the child being distressed; the child being disengaged; the child being unable to provide a verbal response; and child's refusal. Although child's absence was a common reason for not completing the tests, child refusal and inability to provide a response represented 47% of non-completion. Note that the overall completion rate across the three tests was the lowest for the GAPS assessment, meaning that several children took the REV and the RAPT test but struggled or refuse to take the GAPS test. In other words, many of the missing cases for the GAPS test at this stage are likely to reflect inability to answer the questions rather than random factors such as absence.

Some of the children that completed the tests obtained a score of zero. The fractions of children with zero scores are relatively low in the case of the REV test (2.3%) and of the RAPT test (3.2% for the information component of the test, and 10.6% for the sentence structure component). However, the fraction of children with zero scores for the GAPS test is very large (43.2%). There are two main reasons for this large fraction of zeros. First, the assessment administrators were advised to adopt a stopping rule, whereby the GAPS test would stop in case the child failed to provide a correct answer to the first three questions. There were 283 such cases corresponding to 18.3% of all children assessed, and explaining 42.4% of the zero values. The remaining reasons for not completing the GAPS test included: child with EAL; child struggle to complete; external distraction; child is shy; child with send. Among these reasons, the child struggling to understand and to complete the test represents the largest fraction of zero values, signalling that inability to answer the questions was a key factor on non-completion. In all these cases the child failing to complete the test was assigned a score of zero.

Figure 4, flow diagram of assessment completion



Children that stopped taking the GAPS assessment or that did not complete the assessment for the reasons stated above were assigned a score of zero. This score is censored because it is unlikely that if the children had been given the opportunity to complete the test to the end they would have scored zero. This is true for a) the children that struggled to take the test and b) the children that stopped after failing three questions. Note that the GAPS assessment does not increase in difficulty as it progresses, because it is designed to assess different competencies. GAPS includes sentences that vary in length, grammatical difficulty, and phonology to assess different dimensions of oral skills but are not arranged from easiest to hardest. The ordering of the items is meant to sample a range of oral language skills across the assessment. Hence, children that stopped after the third question could have obtained a score larger than zero had they continued. The observed zeros in the GAPS test are therefore mostly censored, in the sense that they reflect scores larger than zero in most cases.

Censoring of the GAPS score poses two potential problems to the estimation of the treatment effects:

- Statistical power: in case we were to set the zero values to missing
- Bias: in case the treatment affects missing score in a way that is correlated with the outcome

At the follow-up all children that completed **at least one** of the three assessments will be assessed again using all three assessments. This comprises 1,705 children. We expect that at the follow up fewer children will obtain zero scores in the GAPS test for two reasons. First,

zero scores are strongly correlated with age. A regression of scores on age (in months) obtains a coefficient of 0.25 suggesting that after 9 months the average score in the study sample should increase by about 2 points (out of 11). A regression of the probability of scoring more than zero on age (in months), obtains a coefficient of 0.03, suggesting that after 9 months at least 25% of the children that scored zero at baseline should obtain a score larger than zero. Second, at the follow up we will no longer employ the stopping rule, which accounted for more than 40% of zero values at the baseline.

To plan the analysis of GAPS scores we consider two scenarios. In the first scenario follow-up data include few zero scores, in the second scenario follow-up data include again many zero scores. If the follow-up data include a limited number of zero scores, the estimation of treatment effects is likely to be unbiased. Zero scores at the baseline (either because of the stopping rule or other causes) are orthogonal to treatment assignment because at the time of the baseline assessments, assessors and setting staff were unaware of treatment status. This is illustrated in Table 5 which shows that the proportion of zero scores due to not completing the test and to the stopping rule respectively are very similar in the project and in the control group at baseline.

Table 4 Characteristics of GAPS assessments results

GAPS	Project	Control
Not completed (scored 0)	0.242	0.256
Stopping rule (scored 0)	0.192	0.172
Score (non-zero values)	4.649	4.601
St. deviation	3.960	3.888
Z-score (non-zero values)	1.184	1.172
St. deviation	1.009	0.990

In the second scenario we collect a large number of non-zero scores at the follow-up. In this scenario we have two options:

- Including all observations with zero values in the analysis without further adjustments
- Using a model adjusting for censored observations

The first option is not desirable because it implies a reduction in sample size (the missing values) and ignores censoring of the data (the zero values). We favour the second option which uses the full sample and accounts for censoring. In particular, we will use a *tobit* model (Greene, 2018). Tobit models are specifically designed for addressing censored observations and can be easily employed within a mixed-effect model.

In a Tobit regression we model the outcome as arising from a latent continuous variable. The tobit multilevel random intercept model is exactly like the random intercept model (1) but applied to a latent variable y^*_{uit}

$$y^*_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 T_{ui} + \sum_{s=1}^{10} \gamma_s S_{si} + \alpha_u + \varepsilon_{iu} \qquad (10)$$

Where the observed score is:

$$y_{uit} = \begin{cases} y^*_{uit}, if\ y^*_{uit} > 0 \\ 0, if\ y_{uit} \leq 0 \end{cases}$$

Unlike OLS, which treats zeros as true observations, the Tobit model yields consistent estimates by incorporating both the censored and uncensored parts of the likelihood function.

Compliance

The ITT analysis does not assess the impact of the intervention on children receiving the intervention. Instead, it assesses the impact of the assignment to the intervention, regardless of whether the child receives the intervention or not. If we are interested in the impact of the intervention on children receiving the intervention, we need the estimate the conditional average causal effect (CACE), also known as local average treatment effect (LATE). This is the impact of the intervention on children that received the intervention after been offered the intervention (the compliers).

The first step in estimating the CACE effect is defining compliance. Our definition is based on the number of *Talking Time*© sessions attended. The delivery team expected that a meaningful impact of the intervention would require children to attend at least 50% of sessions. Accordingly, we use 50% of session attendance as the threshold measure of minimum compliance. Additionally, we will estimate the impact of the intervention under full compliance, i.e. for the fraction of children attending all the sessions.

We will consider two measures of compliance:

- full compliance (children attending all the sessions)
- partial compliance (children attending at least 50% of sessions)

Assessing the impact of the intervention on compliers is challenging because compliers are likely to differ systematically from children that are offered the intervention but do not receive it (non-compliers). Ideally, we would compare compliers to individuals in the control group who would have complied had they been given the opportunity to receive the treatment.

Under two key assumptions, we can apply an instrumental variables (IV) approach to estimate the CACE effect on the compliers (Imbens & Rubin, 2015). The first assumption is that treatment assignment is unconfounded -meaning that we can estimate both the effect on treatment assignment on the outcome (the ITT effect) and the probability of receiving the treatment in the intervention group. This assumption holds by design, as the intervention is randomly assigned. The second assumption, known as the 'exclusion restriction', states that treatment assignment should have no impact on outcomes for non-compliers. While this assumption is often enforced through double-blinding, in its absence, validity may be questioned. In our study, where treatment is assigned at the cluster level, the assumption is more plausible; however, we cannot entirely rule out the possibility that practitioners in the control group may modify their behaviour in ways that influence children's outcomes.

The CACE effect can be expressed estimated as the ratio of the impact of the intervention assignment on the outcome (ITT) to the impact of assignment on intervention receipt. Since the ITT will be estimated in the primary analysis, and the probability of receiving the intervention (i.e., the proportion of compliers) can be calculated from the data, this estimation is feasible within our study design:

$$CACE = \frac{ITT}{P(comp|T)}$$

To account for data heterogeneity, CACE is often estimated using a two-step instrumental variable regression approach. We will implement this method within a multilevel modelling

framework to account for the clustered structure of the data and improve precision of our estimates:

1. First, we will estimate a regression of receipt (compliance) on treatment assignment and other explanatory variables using a multilevel model:

$$comp_{ui} = \beta_1 T_{ui} + \sum_{k=1}^{n} \beta_k X_{kit_{-1}} + \alpha_u + \alpha_u T + \varepsilon_i \qquad (10)$$

   - $comp_{ui}$ is compliance of the parent of child *i* in setting *u*
   - T_{ui} is treatment status for child *i* in setting *u* (*T=1* if the child is in the intervention group)
   - β_2 is the estimated impact of the intervention
   - $X_{kit_{-1}}$ are *k* child-level variables recorded at the baseline (age in months, gender, EYPP status, EAL status) and setting-level variables (IDACI index) at the baseline
   - α_u are setting-level random-effects, which affects compliance
   - $\alpha_u T$ are interactions between the random intercept and treatment status

2. We will calculate the predicted values from model (6) above

3. We will estimate the primary outcome equation using a multilevel model and including a predicted term for compliance rather than treatment status:

$$y_{uit} = \beta_1 y_{uit_{-1}} + \beta_2 \hat{comp}_{ui} + \sum_{k=1}^{n} \beta_k X_{kit_{-1}} + \alpha_u + \alpha_u \hat{comp}_{ui} + \varepsilon_i \qquad (11)$$

Where:

- y_{uit} is the test score of child *i* in setting *u* at the endline *t*
- β_1 is the correlation between the scores for the same child before (t_{-1}) and after the intervention (*t*)
- $\hat{comp}_{ui}$ is predicted compliance for child *i* in setting *u*
- β_2 is the estimated CACE effect
- $X_{kit_{-1}}$ are *k* child-level variables recorded at the baseline (age in months, gender, EYPP status, EAL status) and setting-level variables (IDACI index) at the baseline
- α_u are setting-level random-effects
- $\alpha_u T$ are interactions between the random intercept and treatment status
- ε_i is a child level error term

Multiple hypotheses testing

The secondary analysis evaluates the impact of *Talking Time*© on expressive vocabulary, expressive language (which includes two separate scores), and GAPS for two subgroups (children with EAL, and children eligible for EYPP). We therefore estimate a total of eight different effect sizes.

When many statistical tests are conducted at the same time, there is a risk of finding statistically significant effects simply as the result of chance. Using a 95% confidence threshold, when conducting a statistical test there is a 5% chance of finding a statistically significant effect and a probability of 95% of finding a non-statistically significant effect because of random chance. If we conduct *n* independent tests, there is a probability $P_{ns} = 0.95^n$ of not finding a single statistically significant effect, and a probability $P_s = 1 - P_{ns}$ of finding at least

one statistically significant effect. With eight independent tests this probability is equal to 1-0.95^8=0.34. There is therefore about a one in three chance that even if the intervention has no impact whatsoever, we will find at least one statistically significant effect in our secondary analysis.

To account for multiple testing for secondary outcomes, we will adopt a more conservative approach to statistical significance. One commonly used method is the Bonferroni adjustment, which divides the 5% statistical significance threshold by the number of tests conducted. For eight tests, the threshold becomes 0.006, and only those tests with p-values lower than 0.006 are accepted as statistically significant. This procedure however is very conservative and tends to find a very small number of statistically significant results. We will use instead a False Discovery Rate adjustment (Efron & Hastie, 2016). In this procedure, we order the p-values for each of the eight tests from smaller to largest, where *i* is the index order of the tests by the ascending order of its p-value. We then calculate the Benjamini-Hochberg significance threshold for each test as:

$$tr_i = \frac{i}{N} q$$

Where *i* is the ranking order of each p-value as defined above, *N* is the total number of tests, and *q* is the chosen statistical significance threshold (a choice of *q*=0.10 is common).

We will then compare the observed p-values to the adjusted threshold and find the maximum *i* for which the observed p-value is smaller than its adjusted threshold tr_i. All effect sizes up to this maximum value of *i* will be accepted as statistically significant.

Intra-cluster correlations (ICCs)

Observations within clusters are correlated, and this has implications for the calculation of standard errors and statistical power. In a two-level data structure (children clustered within settings) we can calculate the intra-cluster correlation (ICC) using the formula:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \qquad (12)$$

where σ_b^2 is the variance between clusters and σ_w^2 is the variance within clusters. The ICC is the proportion of the variance between clusters over the total variability of the variable considered. The variances in (7) are estimated by the multilevel models and the ICC is easily computed.

Effect size calculation

The impact estimates obtained using the ANCOVA regression model (1) will be converted into standardised effect sizes in order to facilitate the comparability of effects across studies. The effect size will be standardised by the unconditional pooled standard deviation of the project and control groups as recommended by the EEF statistical guidance (EEF, 2022b):

$$ES = \frac{(\bar{Y}_t - \bar{Y}_c)_{adjusted}}{sd_{unconditional}}$$

where $(\bar{Y}_t - \bar{Y}_c)_{adjusted}$ denotes the ANCOVA difference in means between the treatment and the control group adjusted for pre-test score and other stratification variables specified in the regression model described in the previous section. The pooled unconditional standard deviation is:

$$S_{unconditional} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

References

August, D., & Shanahan, T. (2006). *Developing literacy in second language learners: Report of the national literacy panel on language minority children and youth* (D. August & T. Shanahan, Eds.). Erlbaum.

Dockrell, J. E., Howell, P., Leung, D., & Fugard, A. J. B. (2017). Children with speech language and communication needs in england: challenges for Practice. *Frontiers in Education*, *2*. https://doi.org/10.3389/FEDUC.2017.00035

Dockrell, J. E., Stuart, M., & King, D. (2010). Supporting early oral language skills for English language learners in inner city preschool provision. *British Journal of Educational Psychology*, *80*(4). https://doi.org/10.1348/000709910X493080

EEF. (2022). *Statistical Analysis Guidance for EEF Evaluations*. EEF.

Efron, B., & Hastie, T. (2016). Computer Age Statistical Inference. In *Computer Age Statistical Inference*. https://doi.org/10.1017/cbo9781316576533

Eldridge, S. M., Ashby, D., & Kerry, S. (2006). Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, *35*(5). https://doi.org/10.1093/ije/dyl129

Gardner, H., Froud, K., McClelland, A., & Van Der Lely, H. K. J. (2006). Development of the Grammar and Phonology Screening (GAPS) test to assess key markers of specific language and literacy difficulties in young children. *International Journal of Language and Communication Disorders*, *41*(5). https://doi.org/10.1080/13682820500442644

Hassinger-Das, B., Toub, T. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). A matter of principle: Applying language science to the classroom and beyond. *Translational Issues in Psychological Science*, *3*(1). https://doi.org/10.1037/tps0000085

Hayes, R. J., & Moulton, L. H. (2017). *Cluster Randomised Trials*. CRC Press.

Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences. In *Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Kieffer, M. J. (2008). Catching Up or Falling Behind? Initial English Proficiency, Concentrated Poverty, and the Reading Growth of Language Minority Learners in the United States. *Journal of Educational Psychology*, *100*(4). https://doi.org/10.1037/0022-0663.100.4.851

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

Massonnié, J., Llaurado, A., Sumner, E., & Dockrell, J. E. (2022). Oral language at school entry: dimensionality of speaking and listening skills. *Oxford Review of Education*. https://doi.org/10.1080/03054985.2021.2013189

Morra Pellegrino, M. L., & Scopesi, A. (1990). Structure and function of baby talk in a day-care centre. *Journal of Child Language*, *17*(1). https://doi.org/10.1017/S030500090001312X

Muñoz, J., Hufstedler, H., Gustafson, P., Bärnighausen, T., De Jong, V. M. T., & Debray, T. P. A. (2023). Dealing with missing data using the Heckman selection model: Methods primer for epidemiologists. *International Journal of Epidemiology*, *52*(1). https://doi.org/10.1093/ije/dyac237

Renfrew, C. (2010). *The Bus Story*. Winslow.

Renfrew, C. (2019). *Renfrew Action Picture Test* (5th Edition). Speechmark Publishing Ltd.

Renfrew, C. (2023). *Renfrew Expressive Vocabulary Test* (5th Edition). Speechmark Publishing Ltd.

Rowe, M. L. (2022). Environmental influences on early language and literacy development: Social policy and educational implications. In *Advances in Child Development and Behavior* (Vol. 63). https://doi.org/10.1016/bs.acdb.2022.04.001

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language*, *47*(1). https://doi.org/10.1017/S0305000919000655

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3). https://doi.org/10.1093/biomet/63.3.581

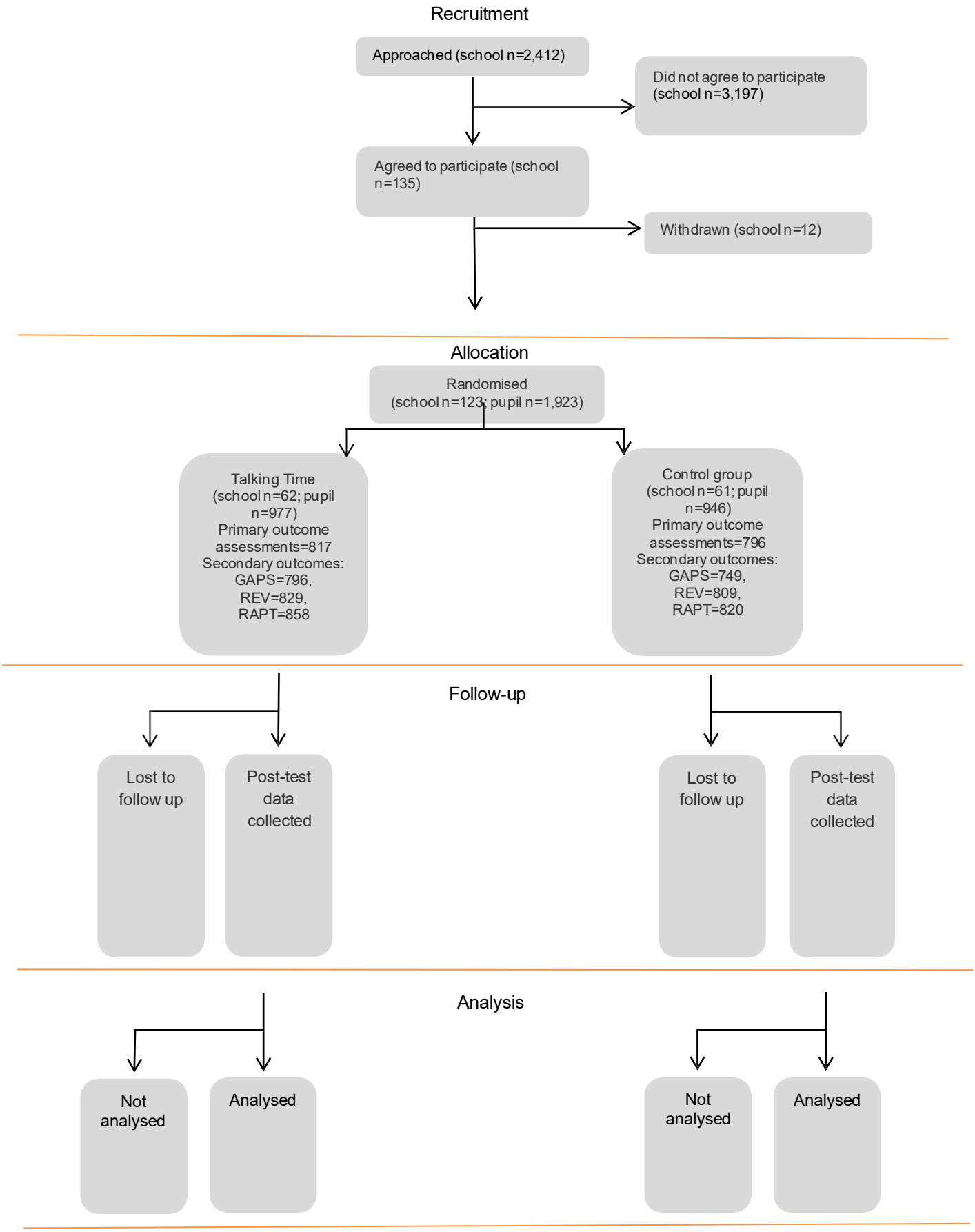
Rubin, D. B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, *91*(434). https://doi.org/10.1080/01621459.1996.10476908

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, *2*(3). https://doi.org/10.1214/08-AOAS187

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Maturi, T., Einbeck, J., Higgins, S., Culliney, M., & Demack, S. (2023). *Improving power calculations in educational trials*. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/methodological-research-and-innovations/improving-power-calculations-in-educational-trials

Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd edition). CRC press.

Appendix 1: Participants flow diagram



Appendix 2: Randomisation algorithms

Randomisation was conducted in the following way:

- We assigned a random number to each setting drawing from a uniform distribution
- We sorted the settings by the random number within each stratum
- We randomly assigned the first setting either to the intervention group or to the control group drawing a number from 0 to 1 from a uniform distribution (if the number=>0.5 the first observation was assigned to the intervention)
- We assigned every other setting within each stratum either to the project or control arm following the sorting order (each setting being assigned to the opposite arm of the previous observation)

In order to avoid the occurrence of a "bad" randomisation we adopted a restricted randomisation procedure (Hayes & Moulton, 2017). The goal of the procedure was to select a randomisation that minimises the difference in key characteristics correlated with the outcome. We considered the following characteristics: proportion of children eligible for EYPP in the setting, proportion of children with EAL, and the setting-level IDACI index of deprivation. It was included because to accommodate logistical issues faced by the delivery team, the sample is drawn from a large number of small strata, which increases the chances of conducting a "bad" randomisation.

Restricted randomisation was conducted in the following way:

- We repeated the randomisation described above 1,000 times
- Each time we regressed the random allocation against the three characteristics selected (proportion of EYPP, proportion of EAL, and IDACI) and we recorded the R-square of the regression
- We selected among the 1,000 random allocations, the allocation with the lowest R-square (showing the highest degree of independence between allocation and determining factors)

The code used is reported below:

```
/* RESTRICTED RANDOMISATION*/

global X="idaciscore EYPP EAL"                  // balancing variables

cap program drop randsel                        // restricted randomisation program
program define randsel
use randsample, clear
local seed=777+`1'                              // set new seed at each draw to keep same results
set seed `seed'
qui generate rand=uniform()                     // generate a random number
sort re_sstrata rand                            // sort observations by the random number within strata
local start = round(runiform(0, 1))             // set a random start (0 or 1)
local end=`start'+1
egen sel`1' = seq(), from(`start') to(`end')    // assign every other observation to the intervention group
qui replace sel`1'=0 if sel`1'==2
drop rand
sort settingname
qui save sel`1', replace                        // save the random allocation in a separate file
```

```
use randsample, clear                                  // open original file
sort settingname
qui merge settingname using sel`1'                     // merge to the saved random allocation
qui drop _merge
qui reg sel`1' $X                                      // regress the random selection on the covariates
di in green "selection "`1' "   "   in yellow "R2="e(r2)   // display the R-square
qui save randsample, replace
erase sel`1'.dta
end

forvalues i=1(1)1000 {                                 //run the program a 1000 times
randsel `i'
}
```