Story Choices: A randomised controlled trial

Evaluation Study plan

Evaluator (institution): University of Birmingham, Coventry University

Principal investigator(s): Julia Carroll

Education Endowment Foundation

Evaluation Summary

Project title	Story Choices: A randomised controlled trial
Evaluator *(Institution)*	University of Birmingham and Coventry University
Principal investigator(s)	Julia Carroll
Study plan author(s)	Julia Carroll, Paul Thompson, Sian Alsop
Trial design	Two-arm pupil randomised controlled trial with random allocation at setting level
Trial type	Teacher Choices RCT
Pupil age range and Key stage	Ages 3-4 (Early Years)
Number of settings *(at design stage)*	100 settings
Number of pupils *(at design stage)*	1200 pupils
Primary outcome measure and source	Pupil storybook vocabulary test (bespoke assessment)
Secondary outcome measure and source	Pupil attitudes towards books (researcher developed assessment) LanguageScreen (OxEdandAssessment)

Study plan version history

Version	Date	Reason for revision
1.2 [*latest*]		
1.1		
1.0 [*original*]		N/A

Any changes to the design need to be discussed with the EEF Evaluation Manager prior to any change(s) being finalised. Describe in the table above any agreed changes made to the trial design. Also, please reflect any changes in the body of the document where appropriate.

Contents

Tables

Figures

Study rationale and background

Teacher Choices Trials

Teacher Choices trials explore some of the most common questions educators ask about their practice and the everyday choices they make when planning teaching and supporting pupils. The aim of Teacher Choices research is to investigate the impact of these different day-to-day pedagogical practices on pupil learning and to generate evidence that can be readily applied by educators. Most Teacher Choices trials focus on school age children, but this project will focus on the Early Years (3-4 year-old children).

Teacher Choices projects aim to focus on research questions that are:

- A genuine area of interest to UK education settings
- A choice that could be made by educators in UK education settings
- Straightforward to implement without significant training and resources
- Feasible to implement and assess within the constraints of an EEF trial.

The Scoping Phase

Coventry University were commissioned to carry out a scoping phase on the topic area of Early Language for the Teacher Choices trial in early 2023. We were asked to investigate the feasibility and importance of four potential research questions, as follows:

1) Which approach is the most effective for pre-teaching key vocabulary to a small group of children?
2) Which approach to monitoring/planning guided play is most effective?
3) Which approach to supporting role play in the classroom is the most effective?
4) Which whole class/group approach is the most effective for teaching language skills through stories?

We identified through initial literature searches and educator survey responses that research question 1 (RQ1) was adequately answered within the literature (e.g. Coyne et al, 2007; Senechal, 1997) and was not a significant dilemma for educators. Both educators and researchers agreed that using multiple approaches to teaching were important for vocabulary acquisition.

We agreed through Study Advisory Board (SAB) discussions following initial literature searches that RQ3 was not suitable for addressing using the Teacher Choices approach, because it was difficult to see how fidelity could be measured. For example, Skene et al (2022) documents some of the difficulties in assessing whether play-based intervention had an effect on language outcomes.

This allowed us to focus on RQ2 and RQ4 in further survey work, focus groups and pilot work. Both questions were significant dilemmas of choice for early years educators, but eventually it was agreed that RQ4 was the most feasible question to address using a Teacher Choices methodology.

Evidence supporting the selection and refinement of RQ4

There is extensive evidence supporting the value of storybook reading with preschool children as a method of improving language skills (Dowdall et al, 2020). In fact, almost all preschool group language interventions that we encountered include storybook reading as an element (e.g. Toub et al, 2018; West et al, 2024).

There is also evidence supporting interactive book reading approaches (also known as dialogic approaches) (Whitehurst et al, 1988). However, we did not find any previous research addressing the issue of planned versus spontaneous interactive reading sessions within an Early Years setting. Much of the previous work on storybook reading interventions centres on parent based intervention rather than educator based intervention (Dowdall et al, 2020).

Our educator survey indicated that all respondents recognised the value of storybook reading and they all included some kind of group storybook reading in their weekly routine. Similar proportions of the sample used each of the three different proposed approaches: pre-planning sessions, responding to child choice or a mixture of the two. Hence, this was a genuine dilemma of choice on an important issue in early years settings.

Given the small changes between the conditions, we anticipated small effects on standardised measures of vocabulary. We therefore looked at ways of maximising the sensitivity of the trial to allow us to detect small effects. We decided to do the following:

- Reduce the number of approaches to be compared to two rather than three, increasing power for the remaining comparison.
- Implement a bespoke storybook reading vocabulary measure that should be maximally sensitive to changes in the vocabulary used in storybooks aimed at this age range.

Evidence supporting the feasibility of implementation and the value of the teacher guidance

We carried out pilot studies giving educators a draft of the educator guidance (shown in appendix 1) and asking them to complete the record forms as indicated, working with a small group of around 6 children. We trialled the storybook reading and contingent interaction approaches. Three educators trialled the storybook reading approaches – two for the planned storybook reading approach and two for the child-led storybook reading approach.

The guidance provides general information on interactive storybook reading and links to videos and guidance freely available on the EEF website. This information is provided for educators in both conditions. In the planned approach, educators are asked to plan the books they will read ahead of time and the discussion points around those books, while in the child led approach, educators are asked not to plan which book to read but instead to offer children a choice of books.

Five settings took part in this phase. Feedback on the storybook reading approaches was all positive. Educators found the guidance clear and the videos on the EEF website useful. However, there are some reasons to be slightly cautious about these findings – educators only used the approach for one week, and only with a small group of children rather than a large group. We might expect to find that adherence to the guidance was less consistent over a longer period of time, and with a larger and more diverse group of children.

Feasibility Assessment

The feasibility of RQ2 and RQ4 are summarised in Table 1. On the basis of this assessment, we selected RQ4 as the most feasible research question to address in the full evaluation.

Table 1: Feasibility Assessment

Feasibility Factors	RQ2 Contingent Interactions	RQ4 Storybook Reading
True dilemma of choice	Moderate. Teachers agree it is important but not much discussion about how to structure it.	High. Settings tend to use one approach or another and will argue why their approach is best
Effectiveness of choice approaches	Previous research suggests that scaffolding/guided play is crucial to early learning, but there is a gap in the literature about how to best plan/structure it	Extensive evidence suggests that interactive storybook reading is effective at improving language but there is a gap in the literature on how to plan it.
Teacher interest in question	Moderate. All educators mentioned guided play as important.	Moderate – as detailed above most educators tended to believe that they took the most effective approach to storybook reading in their context
Trial implementation feasibility	Implementation would be relatively straightforward with some simplification	Implementation would be straightforward – almost all settings regularly use storybook sessions and they are clearly delineated from freeflow play
Prevalence of choices in normal practice	High. Most educators said guided play is the most common activity they do	High. Most educators read storybooks in group situations most days
Heterogeneity in choice implementation	Moderate. Some settings have key worker system for monitoring, but there is little planning of interactions	High. Approximately one third of settings surveyed planned their books, one third were child led and one third used a combination of the two approaches.

Teacher Choice Approaches – Intervention Description

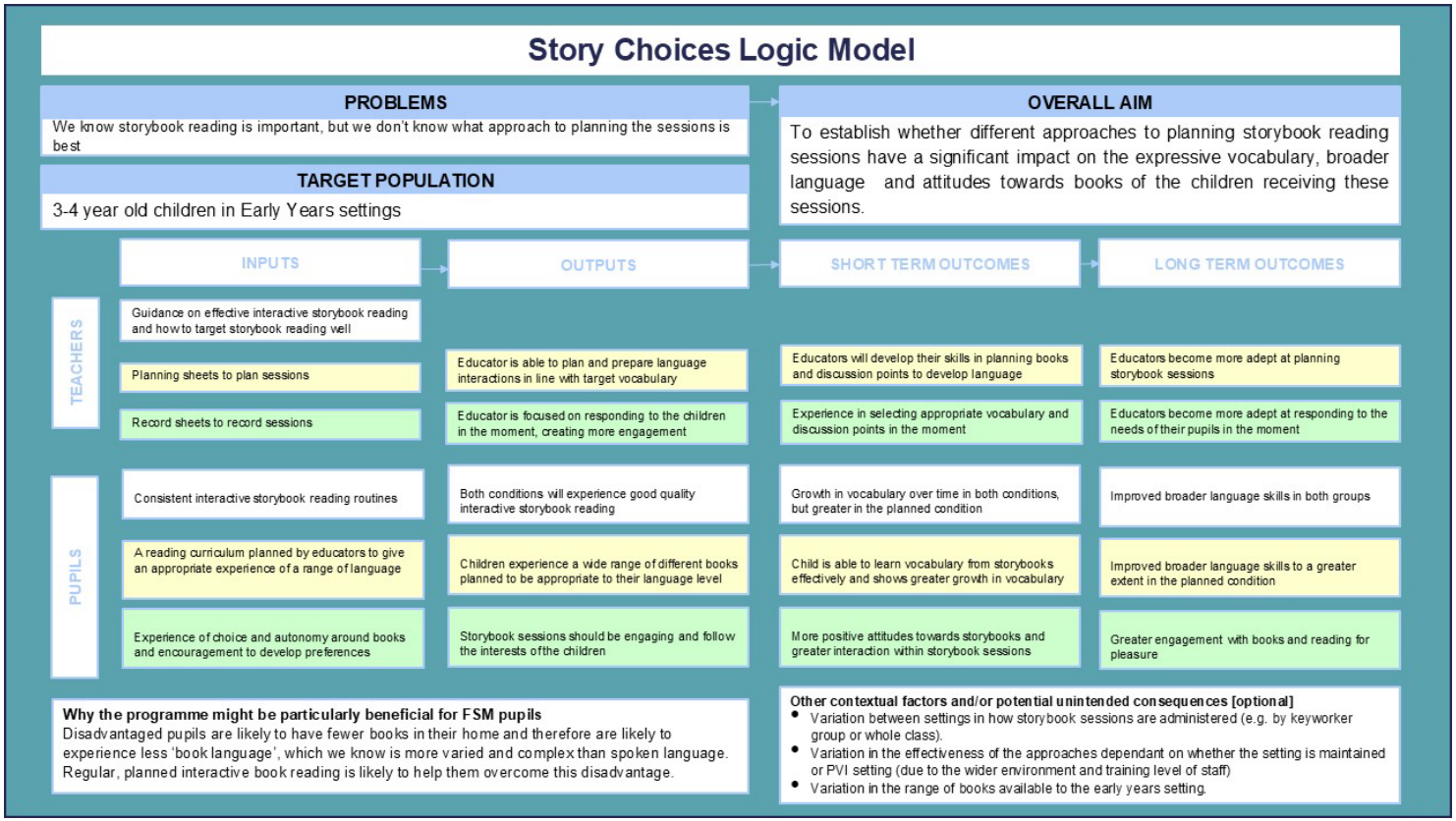
Details of the nature of the two contrasting interventions are given in **Table 2**.

Table 2: Intervention Description

NAME	Planned Storybook Sessions	Responsive (child-led) sessions
WHY (RATIONALE)	Approximately one third of survey respondents use this approach to storybook reading, so it reflects common practice. It allows educators to plan the content and emphasis of the sessions in advance to align to the needs and interests of their group. Extensive research shows the value of dialogic or interactive reading, but previous research has not investigated the value of planning these interactions.	Approximately one third of survey respondents use this approach to storybook reading, so it reflects common practice. It allows children to be actively involved in the storybook session and select a book they are interested in. Research shows that allowing older children to choose the book they read improves motivation and engagement. It is not known the extent to which this will be replicated in prereading children listening to stories in a group.
WHO (RECIPIENTS)	3-4 year old children in Early Years settings	3-4 year old children in Early Years settings
WHAT (MATERIALS)	Early Years settings will be asked to use their existing storybooks for these sessions to effectively reflect day to day practice. Educators will be given guidance on how to plan the sessions, including access to videos of an interactive storybook reading session. They will receive planning sheets which they are asked to complete before the sessions.	Early Years settings will be asked to use their existing storybooks for these sessions to effectively reflect day to day practice. Educators will be given guidance on how to plan the sessions, including access to videos of an interactive storybook reading session. They will receive record sheets which they are asked to complete after each session.
WHAT (PROCEDURES)	Educators will follow their normal routine for group storybook reading (e.g. whole class or key worker groups) except that sessions will be planned beforehand by the educator. The educator is asked to record ahead of time which book they will use and what the key discussion points and vocabulary they will	Educators will follow their normal routine for group storybook reading (e.g. whole class or key worker groups) except that sessions will not be planned beforehand by the educator. Instead the book to be read will be selected by one or more of the children from a choice of at least 5 books. The educator is

	highlight during the session will be. They are given a record form for this purpose. After the session they are asked to complete a brief reflection saying whether the session went as planned.	encouraged to use interactive reading with the book, planning 'in the moment' to develop discussion points and highlight vocabulary depending on the interest of the children.
WHO (PRACTITIONERS)	Early years educators and teachers. They will not receive any specific training but will be asked to follow some written guidance. We intend to collect details about the educators' qualifications, experience, and knowledge of oral language pedagogy.	Early years educators and teachers. They will not receive any specific training but will be asked to follow some written guidance. We intend to collect details about the educators' qualifications, experience, and knowledge of oral language pedagogy.
HOW (DELIVERY GUIDE)	Storybook sessions will be delivered in groups to the children in the class as part of the daily routine of the class.	Storybook sessions will be delivered in groups to the children in the class as part of the daily routine of the class.
WHERE (LOCATION)	In Early Years settings	In Early Years settings
WHEN & HOW MUCH (DOSAGE)	At least three times a week for at least 15 minutes per session over the 12 weeks of the intervention period from Feb 2025 to May 2025.	At least three times a week for at least 15 minutes per session over the 12 weeks of the intervention period from Feb 2025 to May 2025.
TAILORING (ADAPTATION)	Educators can choose the books they use and the discussion points they raise themselves, depending on the topics they are covering in class, the needs of their group or any other factors.	Educators should provide children with a range of books from which to choose. They should use some procedure of their choice to ensure that each child gets an equal opportunity to choose – e.g. selecting a name out of a hat or having the names on a rota.

Figure 1: Logic Model



Impact Evaluation Design

Research questions

Overall Aim:

To establish whether different approaches to planning storybook reading sessions have a significant impact on the expressive vocabulary, broader language and attitudes towards books of the children receiving these sessions.

The **primary research question** is as follows:

> *RQ1: What is the difference in vocabulary knowledge measured by a bespoke vocabulary test of early years children participating in planned storybook reading sessions in comparison to children participating in child-led storybook reading sessions?*

We anticipate that regular interactive storybook reading will improve vocabulary in both arms of the trial. However, we hypothesize it will increase more in the planned storybook approach. As detailed in the logic model, we hypothesize that the planned storybook approach will allow teachers to tailor the teaching to the language needs of the children more directly and therefore lead to a greater increase in vocabulary.

We will also estimate differences in storybook vocabulary in EYPP (Early Years Pupil Premium) pupils, as the logic model (Figure 1) highlights that the planned storybook reading approach may be particularly valuable for children from disadvantaged backgrounds.

> *RQ2: Does the impact of the intervention differ for children eligible for Early Years Pupil Premium?*

As detailed in the logic model, we anticipate that the responsive approach will help to improve pupils' attitudes towards books, as they will experience greater choice and autonomy around books, which is shown to be an important factor in positive reading attitudes. This leads to RQ3.

> *RQ3: What is the difference in attitudes towards books measured by a brief questionnaire measure of early years children participating in planned storybook reading sessions in comparison to children participating in child-led storybook reading sessions?*

As detailed in the logic model, we anticipate that the planned storybook approach will lead to improved performance in a standardised, global measure of language as a result of the increased range of vocabulary and narrative experienced in this approach. This leads to RQ4.

> *RQ4: What is the difference in overall language level measured by the LanguageScreen measure of early years children participating in planned storybook reading sessions in comparison to children participating in child-led storybook reading sessions?*

Further research questions address the ways in which the effects of the intervention may be moderated by factors linked to characteristics of the Early Years setting or how the approaches are implemented. Private, Voluntary and Independent settings differ from maintained settings in multiple ways, including the age range of children attending, the ratios between staff and pupils and the qualifications of the staff. We anticipate that qualified teachers working with pupils within a small age range may be more effective in implementing the approaches effectively. We therefore hypothesize they are likely to be effective in improving language across both approaches, but particularly in the planned approach.

RQ5: Does the impact of the intervention differ according to whether the setting is a PVI or maintained setting?

Implementation process evaluation will be used to address questions around the fidelity of implementation, the perceived value of the approaches by practitioners and the role of contextual factors in the success of the implementation and approaches.

RQ6: Does dosage (measured in terms of total number of sessions) mediate the vocabulary gains shown by pupils in the two approaches?

RQ7: Does fidelity to the programme approach mediate the vocabulary gains shown by pupils in the two approaches?

RQ8: What is the perceived value of the two approaches by practitioners using each approach?

RQ9: What are the potential costs to the setting of each of the two approaches?

Similarly, we anticipate that the impact of the intervention may be moderated by various characteristics of the storybook reading routine (e.g. whether it is carried out in small groups or large groups), and the range of books available in the setting. However, previous research does not allow us to hypothesize what the effects of these different characteristics are, and so we pose two non-directional research questions:

RQ10: Does the impact of the intervention differ according to other characteristics such as storybook reading routine or range of books in the setting?

RQ11: What are the characteristics of settings which show good progress in storybook vocabulary in each of the two approaches?

Design

Table 3 summarises the trial design.

Table 3: Trial Design

Trial design, including number of arms		Two-arm parallel group cluster RCT**-arm, cluster randomised**
Unit of randomisation		setting
Stratification variables (if applicable)		PVI versus Maintained, Region
Primary outcome	**Variable**	Storybook vocabulary
	Measure (instrument, scale, source)	Expressive vocabulary measure focused on storybook language, 0-46, Researcher-designed
Secondary outcome(s)	**Variable(s)**	1. LanguageScreen Expressive Vocabulary 2. LanguageScreen Listening Comprehension 3. LanguageScreen Receptive Vocabulary 4. LanguageScreen Sentence Repetition or SR 5. Attitudes towards books
	Measure(s) (instrument, scale, source)	LanguageScreen (standard score, OxEdandAssessment.com)

		Revised version of the attitudes towards literacy scale, researcher designed
Baseline for primary outcome	**Variable**	Storybook vocabulary
	Measure (instrument, scale, source)	Expressive vocabulary measure focused on storybook language, 0-46, Researcher-designed
Baseline for secondary outcome	**Variable**	The pre-test attitudes towards books measure will act as baseline for the post-test attitudes towards books measure, The pre-test storybook vocabulary will act as the baseline measure for the post-test LanguageScreen measure.
	Measure (instrument, scale, source)	as described above.

The trial will be a two-arm, parallel group cluster randomised controlled study randomised at the level of setting.

The primary outcome measure is a bespoke vocabulary measure, while the secondary outcome measures are a standardised language measure and an attitudes towards books measure.

Participant selection

Participants will be children in group-based Early Years settings in the year before they start reception (ages 3;0 to 3;11 years in September 2024). Settings are eligible to take part if they are eligible to receive government funding for childcare hours, have at least 8 children in this age group who attend for at least 15 hours a week term time, are in the target areas (East Midlands, West Midlands and Greater Manchester), and are not taking part in another research trial relating to language or literacy.

Note that all children in this age range will be entitled to 15 hours of childcare per term time week (averaged across the year in some cases), but some may split the hours between more than establishment, or their parents may choose not to use their entitlement. We assume a minimum of 8 children per setting, and a maximum of 20 children per setting, based on the maximum number of children we can assess in a day.

Recruitment Plan

We aim to recruit 100 settings in total, to ensure that we retain the minimum sample after an estimated 20% attrition (based on previous EEF Early Years trials). Government statistics indicate that more children are educated in the private/voluntary/independent sector at age 3 than in the state maintained sector (approximately 69% in PVI and 31% in state: https://explore-education-statistics.service.gov.uk/data-tables/education-provision-children-under-5). With this in mind, we would like to ensure the PVI sector is well represented. A 50:50 split in the sample would provide greatest power. However, previous research indicates higher likelihood of drop out for PVI settings. We therefore aim to recruit 45 settings in the state sector and 55 settings in the PVI sector, with the aim of retaining at least 40 settings in each at the completion of the study.

We intend to focus recruitment around three areas, which currently have fewer EEF trials in the preschool age group: West Midlands, East Midlands and Greater Manchester. We aim to have approximately one third of the settings from each of these areas.

Outcome measures

Below we provide details on the outcome measures. Key details are summarised in Table 4.

Table 4: Outcome Measures

Baseline measure	**Definition**	**Reference**
Bespoke storybook vocabulary measure	An expressive vocabulary measure focused on storybook language and based on the 'informal definitions' task	Hadley et al (2016)
Attitudes towards books	A **Likert** scale questionnaire based on the attitudes to reading scale	Carroll, Holliman & Weir, (2019)
Primary Outcome		
Storybook vocabulary measure	Repetition of previous measure	
Secondary Outcomes		
LanguageScreen	A standardised measure of overall language level, including expressive and receptive vocabulary and narrative skills	West et al (2024)
Attitudes towards books	Repetition of previous measure	

Primary outcome: Bespoke storybook vocabulary measure

This measure is based on the 'informal definitions task' used by Hadley et al (2016). Children are asked to say or show what they know about each word, and their responses are scored on a scale of 0-2. The measure has 23 items in total, distributed across four subscales: one using dolls to act out word meanings (4 items), one requesting actions or gestures (4 items), one using adjectives (5 items) and one using nouns (7 items). Typically, receptive and expressive vocabulary tasks for this age group are limited to items that can be easily pictured, but this format allows us to assess knowledge of a wide range of different words. Importantly, this task can flexibly assess partial knowledge of a word or concept, unlike typical expressive vocabulary tasks which have a highly limited range of possible responses. For example, in the CELF Preschool expressive vocabulary, children only receive points for correctly naming the picture. One of the items is 'telescope'. Responses such as 'you use it for looking at the sky', 'pirates have one' or 'kaleidoscope' would not receive credit even though they indicate partial knowledge of the function or name of the item pictured. In contrast, in our task we provide the item name and ask for information about the item, and answers indicating that the child knew the function of a telescope would receive credit.

Measure Development

In order to make this measure as sensitive as possible to the language that can be learnt in storybooks over this age range, we have taken the following steps.

- We selected 1000 storybooks that represent a cross-section of the storybooks typically used in Early Years settings including well established classics, books representing a range of different ethnicities and nationalities, books covering key topics such as seasons and annual festivals, friendships and emotional development, and everyday activities.
- We created a corpus of the words used within these 1000 storybooks.

- From this corpus we selected words that met the following criteria:
    - Are particularly characteristic of storybook language, rather than child-directed spoken language
    - Occur in at least 1 in 12.5 books in our sample (therefore should be encountered in at least 4 different book contexts throughout the intervention period)
    - Have an 'age of acquisition' rating suggesting they are typically acquired between ages 2 and 6 years.
    - Are unambiguous and can be defined relatively simply, using words or actions.

Pilot Testing Phase 1

We created a shortlist of 75 words which were tested in a first pilot phase with at least 40 children responding to each word.

From this pilot phase we selected 23 words which met the following criteria:

- Showed good correlation with overall scale score
- Did not have issues raised by the testers (for example, the word 'laugh' did not translate well across different accents, and was therefore removed)
- Showed good levels of discrimination according to the item analysis
- Showed varied levels of difficulty according to the item analysis and percentage correct
- Showed better mean scores in the reception aged children in comparison to the nursery aged children. This criterion was set to ensure that the items were words that were likely to be learnt over the period between nursery and reception.

These 23 words were subjected to a second round of pilot testing to allow us to verify the properties of the measure.

We set the following acceptance criteria for the measure:

- Inter-rater reliability >0.85
- Fewer than 10% of children discontinue/refuse to answer all questions
- Internal reliability >0.80
- Significant difference between nursery and reception on average (effect size >0.3)
- Skewness and kurtosis within acceptable levels
- Correlates at >0.60 with a standardised test of vocabulary

Pilot Testing Phase 2

We tested a revised version of the Storybook Vocabulary task with 23 items. Participants and mean scores are given in Table 5. Alignment with preset acceptance criteria are given in Table 6. Distribution of scores on the task for nursery and reception children are shown in Figure 2.

Table 5: Results from Pilot Testing

Total Valid Cases: 297 (5 cases with no responses were removed)				
			Mean score (standard deviation)	
	Number	**Percentage**	**Storybook vocabulary**	**CELF**
Nursery	125	42.0%	27.18 (9.49)	18.43 (6.96)
Reception	172	58.0%	34.77 (7.77)	24.77 (7.43)
Boy	139	47.0%		
Girl	157	53.0%		
EAL	50	16.8%	25.14 (10.60)	15.21 (7.86)
Non-EAL	247	83.2%	32.79 (8.53)	23.55 (7.12)

Assessment of internal reliability did not indicate that any items should be removed.

Table 6: Criteria for Acceptance of Vocabulary Measure

Criteria set	Results found	Pass/Fail
Inter-rater reliability >0.85	0.951	Pass
Fewer than 10% of children discontinue/refuse to answer all questions	From 303 cases, 5 cases did not respond at all and 15 had some missing items. Total 20/303 (6.6% of respondents had some missing data)	Pass
Internal reliability >0.80	Cronbach's alpha = 0.827	Pass
Significant difference between nursery and reception on average (effect size >0.3)	There was a statistically significant difference in scores in Story Choices measure by nursery and reception children. Reception children scored higher than nursery children, *mean difference = 7.59 (95% CI, 5.61 to 9.56), t(294) = 7.55, p = .001. Effect size is large (0.89).*	Pass
Skewness and kurtosis within acceptable levels	Skewness: -1.06 (SE 0.14) Kurtosis: 0.70 (SE 0.28) Distribution for the measure for nursery and reception is shown below. There are no signs of a floor or ceiling effect, though there is some skewness, particularly notable in the reception sample. We believe this is acceptable for two reasons: A) The measure is primarily for use with nursery children B) The skewness indicates a tail of lower scores which is a potential consequence of including items accessible to children with lower levels of attainment, and therefore allows sensitivity at these lower levels.	Pass with some reservations
Correlates at >0.60 with a standardised test of vocabulary	Significant positive correlation was found between the scores in Story Choices and CELF measures, *r = .65, p< 0.001*.	Pass

Figure 2: Distribution of the Vocabulary Scores

Note that this measure is aimed at nursery age children, but that as we wanted to be sensitive to change over the final year of nursery, we tested both nursery and reception age children. The distribution for nursery age children is therefore the more important distribution to view. This distribution is approximately normal, with a tail of lower scores. The distribution for reception age children is still normal but is less ideal, showing a slight ceiling effect (3% of participants achieving maximum score of 46) and a longer tail of poorer performance. However, this is unlikely to be an issue in the study analysis as a) we test nursery children rather than reception age children and b) we plan alternative analysis if model assumptions are not met.

Secondary outcomes

Attitudes towards books

This measure allows us to assess whether there is a difference between the groups in their attitudes towards books and reading (RQ3). As detailed in the logic model, we anticipate that the responsive approach may help to build positive attitudes. This is a brief Likert scale questionnaire based on the work of Carroll, Holliman & Weir (2019). The measure used in this paper has two subscales – one focusing on how much children enjoy different aspects of literacy and one focusing on how frequently they take part in literacy activities. We anticipate that the enjoyment scale will be more appropriate than the frequency scale for this younger age group. We have added pictures to make the task clearer for young children, and added some two dummy items with non-literacy related activities to make the purpose of the scale less obvious After pilot work we removed one item (concerning writing) that was less relevant in preschool classrooms. This has 7 literacy-related items rated on a scale of 1-4, to give a maximum score of 28. We intend to use this total raw score in the analysis.

LanguageScreen

While the storybook vocabulary measure is closely aligned with the vocabulary that children will be experiencing through storybooks, it is also important to assess whether the approaches have an effect on a broader measure of language that is well standardised and has been shown to predict later educational attainment. This provides an assessment of the potential long-term outcomes of the measure, as shown in the logic model. The measure demonstrates high internal reliability (Cronbach's alpha = 0.92) in a sample that is very similar to the target sample (West et al, 2024). The measure has four subscales: expressive vocabulary (ability to name pictures; 24 items), listening comprehension (understanding spoken stories: 16 items), receptive vocabulary (matching spoken words and pictures, 23 items) and sentence repetition (repetition of increasingly complex sentences, 14 items). These subscales demonstrate reliabilities between 0.74 and 0.80. We anticipate that storybook reading should influence both receptive and expressive vocabulary and listening comprehension, and therefore we intend to carry out the full measure. The measure is administered using an iPad, and takes around 10-15 minutes per child to administer. Children are scored correct or incorrect for each item with automated discontinuation rules. The app gives raw scores and standardised scores for each subscale and for overall performance in comparison to a representative sample of the population of children of this age. More details can be found here: https://oxedandassessment.com/uk/languagescreen/.

We do not intend to use National Pupil Database data within this study.

Data will be collected by the staff employed on the project (Julia Carroll and Tanvir Ahmed), alongside temporary research assistants employed for the purpose. These research assistants will largely be undergraduate and postgraduate students studying Education or Psychology courses. All assistants will be trained and will be required to have a DBS certificate.

Settings will be allocated to approach after the pre-testing is complete. Post-testing will be carried out by research assistants who are blind to the approach used by the setting.

Sample size

Table 7 summarises the calculations for planned sample size.

Table 7: Sample Size Calculations

		Design		Randomisation	
		Overall	**EYPP**	**Overall**	**EYPP**
Minimum Detectable Effect Size (MDES)		0.30	0.44	0.316	0.46
Pre-test/ post-test correlations	level 1 (child)	0.42	0.42	0.42	0.42
	level 2 (class)	n/a	n/a	n/a	n/a
	level 3 (setting)	0	0	0	0
Intracluster correlations (ICCs)	level 2 (class)	n/a	n/a	n/a	n/a
	level 3 (setting)	0.20	0.20	0.20	0.20
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		10 (12 with 20% attrition)	2	12	2
Number of settings	Condition 1	45	45	41	41
	Condition 2	45	45	41	41
	Total	90 (100 with 10% attrition	90 (100 with 10% attrition	82	82
Number of children	Condition 1	400	52	478	305
	Condition 2	400	52	478	305
	Total	800	104	956	610

Sample size was determined using the PowerUP! Software (Dong & Maynard, 2013). For all of the scenarios listed, we have estimated MDES based on the following assumptions and a two-level linear mixed effects design with preschool setting at level 2 and controlling for baseline outcome:

Intracluster correlations are set at 0.2, based on the findings from the NELI-preschool trial (West et al, 2024), a recent study with a similar target sample, also focused on language development.

Pre- post-test correlations cannot be calculated from this study as they used a latent factor estimation of overall language skill rather than a single measure of vocabulary. We have set this as high (0.42) based on previous work e.g. Hadley et al (2016).

We have modelled MDES for two different scenarios. The first indicates the minimum sample size needed to detect an effect size of 0.3, working on the basis of an average cluster size of 8 children, (based on the minimum setting size we have set in recruitment criteria) and calculating the attrition at 10% at setting level

and 20% at individual level. It was agreed with EEF that 0.3 was an acceptable MDES in this case. Column two indicates the MDES for children eligible for Early Years Pupil Premium (EYPP[1]) based on this sample size. The next two columns indicate MDES for the achieved sample size at pretesting – 82 settings with an average cluster size of 12. We estimate EYPP numbers based on the assumptions above, as we did not have this information collected at pre-test.

Based on these scenarios, we intend to recruit 100 settings to take part and aim to test 8 children per setting. This allows us to maintain the minimum sample size for a MDES of 0.3 even after 10% attrition at the level of the setting and 20% at the level of the individual. We plan that a maximum of 20 children may be assessed in a given setting. In settings where there are more than 20 children we will create a randomised list of children prior to testing and follow the order of the list, to avoid potential bias that comes with self-selection or teacher selection of children.

We have assumed an average of 12 children per setting being tested in this scenario because we have estimated that one person can easily assess children in a day on the measures, and because we have a minimum setting size of 8.

Early Years Pupil Premium data on each child will be part of the data requested from the setting at pre-test.

Randomisation

This trial will use stratified blocks, cluster randomization at the level of pre-school setting to identify the impact of two story session approaches—planned storybook reading sessions and child-led storybook reading sessions—on the vocabulary and language skills of pupils. A setting is a natural unit of randomization in this case as it minimizes the risk of spillovers and contamination between the two treatment arms and allows the setting freedom to arrange their group storybook sessions in line with their normal practice.

There is no pure control group setting as we are interested in understanding which of the two approaches is most effective rather than assessing the impact of a single approach against the counterfactual of no approach. Randomization will be stratified by the type of setting (PVI versus state maintained) to ensure participation by both types of settings and to investigate potential differences between them. Within each strata, settings will be randomly allocated to either treatment approach 1 or treatment approach 2. Block sizes will be varied to minimise predictability of allocation. The stratifying variable will be included in the primary analysis as a control variable (Kahan & Morris, 2012).

We currently do not intend to randomise in batches. Randomisation will be carried out during the week commencing 7th February 2025 by creating a list for the whole sample in each stratum to ensure every possible scenario is covered.

All randomisation will be undertaken by a statistician blind to the identity of the recruited settings.

[1] We have estimated rates of EYPP based on the total number of 3 and 4 year olds eligible for EYPP (108,328) divided by the total numbers of 3 and 4 year old in preschool settings in 2024 (808,800), resulting in an EYPP rate of 13.4%.

Statistical analysis

Before pre-testing we carried out pilot testing of the vocabulary measure (primary outcome measure) and the attitude to books measure (secondary outcome measure). The results of this pilot work for the vocabulary measure are detailed above. They indicated acceptable internal reliability, test-retest reliability, construct validity, and item loadings. There was also a difference between nursery and reception aged children, with a large effect size.

Imbalance at Baseline

Characteristics of each trial arm group will be summarised descriptively, both as randomised and as analysed in the primary analysis. However, no formal statistical comparisons will be undertaken (Senn, 1994). Continuous measures will be reported as a mean, standard deviation (SD), minimum and maximum, while categorical data will be reported as a count and percentage. We will use the following characteristics:

- Setting characteristics: type of setting, size
- Setting-level pupil characteristics: proportion of pupils eligible for Early Years Pupil Premium (EYPP)
- Pupil characteristics: age, gender
- Current approach to storybook reading in the setting

Primary analysis

An intention-to-treat (ITT) approach will be used (including all randomised settings and children in the analysis), specifying random intercepts to account for the between-setting variability, and reporting standard errors of the parameter estimates. ICCs (at the setting level) will be calculated for the null model (i.e. that without covariates) at post-test.

The primary analysis (RQ1) will examine mean follow-up vocabulary levels, using the bespoke vocabulary test scores, adjusting for the respective baseline measure, setting type, and with the covariate of interest specified as a dichotomous treatment/control variable.

The analysis for the primary outcome measure will use a linear mixed model (LMM), given that the measure is continuously distributed. Piloting of the measure indicates a small amount of skewness. Some amount of skew in the raw outcome measure can be permitted, provided that the assumption of the normality of residuals is satisfied (assessed by looking at the residual diagnostic plots). The range of possible scores indicates that a LMM should be the most appropriate model in this instance. Model assumptions will be checked at the analysis stage and, if necessary, a generalised linear mixed model with appropriate link function can be changed to permit analysis of sufficiently skewed data with heterogeneous residuals (Dobson & Barnett, 2018). Nested model comparison will be based on likelihood ratio tests (Chi square) and both Bayesian and Akaike's Information Criteria (BIC and AIC respectively), with lower indices indicating the preferred model (Harrell, 2001). All analyses will be conducted in R (version 4.4.1 2024-06-14), using the R packages: Tidyverse, lme4, ordinal, lmerTest. The model will include a two-level design (child - level 1[i=1,...,N]; setting – level 2 [j=1,…,J]; where N= total number of children and J = total number of settings) as follows:

Two-level design:

$$\boldsymbol{L1}: Y_{ik} = \beta_{0k} + \beta_{1k} X1_{ik} + r_{ik},\ r_{ik} \sim N\left(0, {\sigma^2}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0k} = \gamma_{00} + \gamma_{20}X2_k + \gamma_{30}INT_k + \mu_{0k}, \qquad \mu_{0k} \sim N(0, \tau_2^2)$$

$$\beta_{1k} = \gamma_{10}$$

where, Y are the vocabulary test scores; X1 are the baseline vocabulary test scores; INT is the treatment/control variable; X2 are indicator of setting type (strata variable). μ_{0k} is the random intercept term for setting. r_{ik} is the individual-level random error term, accounting for student-specific variations within settings.

Assumptions for the linear mixed model will be checked as follows:

1. Linearity – plotting residuals vs predictor(s). If a structure is present, then transformation or an alternate model specification is required (i.e. GLM).
2. Homogeniety of variance – variance of the residuals across groups is the same. There is scope to fit models allowing for heterogeneous groups, but the setup is different (Generalized linear mixed model - GLMM).
3. Residuals are approximately normally distributed – plotting QQ plot

If distributional assumptions are not satisfied, as appropriate, a generalized linear mixed model with alternate link function will be used.

The distributions of the primary outcomes will be assessed prior to conducting the analysis, if variables are skewed, then a Poisson mixed model will be specified, as follow:

$$\boldsymbol{L1}: g(Y_{ik}) = \beta_{0k} + \beta_{1k}X1_{ik} + r_{ik}, \; r_{ik} \sim N\left(0, \sigma^2{}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0k} = \gamma_{00} + \gamma_{20}X2_k + \gamma_{30}INT_k + \mu_{0k}, \qquad \mu_{0k} \sim N(0, \tau_2^2)$$

$$\beta_{1k} = \gamma_{10}$$

Note: $g(Y_{ik}) = \log_e$, where $g(Y_{ik})$ is the log link function for the primary outcome measure.

Alternatively, data transformation could be used but use of the GLMM is preferable.

Effect size for all outcome measures will be reported using Hedges' g (adjusted mean difference) (Hedges, 2007). The two-level linear mixed model has a sample estimate of the effect size equivalent to Hedges' g with 95% confidence interval defined as:

$$\hat{\Delta}_g = \frac{\widehat{\beta_1}}{sd},$$

where $\widehat{\beta_1}$ is the adjusted mean difference in outcome score between trial arms and pooled sd is as follows:

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

Where sd_1^2 is the variance of treatment group; and correspondingly, $sd_2^2$is the variance of control group; $n_1$is the number of individuals in treatment group, and similarly for control group, n_2.

All parameter estimates from the models will be reported with 95% confidence intervals.

We report the ICC from the two-level model which is defined as:

$$\boldsymbol{ICC} = \frac{\boldsymbol{\sigma^2}_{\boldsymbol{school}}}{\boldsymbol{\sigma^2} + \boldsymbol{\sigma^2}_{\boldsymbol{school}}}$$

Where $\boldsymbol{\sigma^2}$ is the residual variance, and $\boldsymbol{\sigma^2}_{\boldsymbol{school}}$ is the random intercept variance according to setting (setting level clustering).

Secondary analysis

The secondary analysis **(RQ2)** will replicate the primary analysis outlined above, with LanguageScreen (overall standardized score) and Attitude to Reading measures (total raw score) used as secondary outcome variables.

$$\boldsymbol{L1}: Y_{ik} = \beta_{0k} + \beta_{1k} X1_{ik} + r_{ik},\ \ r_{ik} \sim N\left(0, \sigma^2{}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0k} = \gamma_{00} + \gamma_{20} X2_k + \gamma_{30} INT_k + \mu_{0k}, \qquad \mu_{0k} \sim N(0, \tau_2^2)$$

$$\beta_{1k} = \gamma_{10}$$

where, Y are the secondary outcome scores at followup; X1 are the baseline secondary outcome scores; INT is the treatment/control variable; X2 are indicator of setting type (strata variable). μ_{0k} is the random intercept term for setting. r_{ik} is the individual-level random error term, accounting for student-specific variations within settings.

Confidence intervals for each effect size will be computed by multiplying the standard errors of each pairwise mean difference by the 2.5th percentile of a Student's t-distribution with the number of degrees of freedom associated with the sample size..

Adjustment of multiplicity

The overall type I error rate for testing between trial arms for the primary endpoint will be controlled at the 2-sided 0.05 significance level. Secondary analyses will control the family-wise error rate using the Holm method.

The Holm method, in a stepwise way, computes the significance levels depending on the P-value based rank of hypotheses. For the ith ordered hypothesis H(i), the specifically adjusted significance level is computed:

$$\alpha'(i) = \frac{\alpha}{m - i + 1}$$

where m is the number of hypothesis tests.

The observed P value $p(i)$ of hypothesis *H(i)* is then compared with its corresponding $\alpha'(i)$ for statistical inference; and each hypothesis will be tested in order from the smallest to largest P values $\left(\mathrm{H}(1),\ \ldots,\ \mathrm{H}(\mathrm{m})\right)$.

The comparison will immediately stop when the first $p(i) \geq \alpha'(i)$ is observed $(i = 1,\ldots,m)$ and hence all remaining hypotheses of $H(j)$ $(j = i,\ldots,m)$ are directly declared non-significant without requiring individual comparison.

Sub-group analyses

To estimate the effect of the intervention for children from disadvantaged backgrounds, the main analysis described above will be repeated on a subsample of children eligible for the Early Years Pupil Premium (EYPP). The model will be run for the primary outcomes only. We will also calculate effect sizes and 95% confidence intervals, as outlined in the primary analysis.

To determine which approach is more effective for disadvantaged children, a moderation analysis will adjust the primary analysis with the inclusion of the moderator as a main effect and interaction between moderator and randomised group indicator. For example, the FSM moderator analysis is as follows:

$$\boldsymbol{L1}: Y_{ik} = \beta_{0k} + \beta_{1k}X1_{ik} + \beta_{2k}FSM_{ik} + r_{ik},\ \ r_{ik} \sim N\left(0, {\sigma^2}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0k} = \gamma_{00} + \gamma_{02}X2_k + \gamma_{03}INT_k + \mu_{0k}, \qquad \mu_{0k} \sim N(0, \tau_2^2)$$

$$\beta_{1k} = \gamma_{10}$$

$$\beta_{2k} = \gamma_{20} + \gamma_{21}INT_k$$

where, Y are the vocabulary test scores; X1 are the baseline vocabulary test scores; INT is the treatment/control variable; X2 are indicator of setting type (strata variable). μ_{0k} is the random intercept term for setting. r_{ik} is the individual-level random error term, accounting for student-specific variations within settings. The cross-level interaction comes when the two-level notation is expanded out, γ_{21}.

Sensitivity analyses

Exploring the impact of missing data on trial outcomes by investigating likely missing data mechanisms and re-fitting the primary outcome within a multiple imputation framework (including exploring MAR and MNAR mechanisms via delta-based controlled multiple imputation). Imputation variables for the model will include all covariates and the outcome appearing in the analysis as per recommendation by White, Royston & Wood (2011). In addition, variables that are predictive of missingness are included on the basis of strength of association with response variables. Also, any variables that explain response or non-response (Van Buuren, Boshuizen & Knook, 1999).

We will summarise the extent of missing data in all outcomes and their respective control variables. A full multiple imputation strategy will be used if more than 5% of data in the primary model is missing. Alternatively, we will impute if more than 10% of data for a single variable is missing. MCAR will be determined using Little's MCAR test.

We will use the multiple imputation by chained equations approach via the mice package in R (Van Buuren and Groothuis-Oudshoorn, 2011) and generate 5 imputed datasets. We will then estimate the intervention effect for each imputed dataset and pool the results using Rubin's combination rules for standard errors.

We will report levels of missingness for each trial stage in a CONSORT flow diagram.

Compliance

The main analyses will use an Intention-to-Treat model. We can also carry out analyses focusing on pupils who received Story Choices as intended, based on the number of sessions attended.

We intend to collect compliance data in terms of a register indicating which children took part in each storybook reading session, and what book was read. We will use this data to calculate the number of sessions attended by each pupil and the number of sessions delivered in each setting – i.e. compliance at the level of the pupil and at the level of the setting. We intend to set non-compliance at the level of setting at fewer than 24 sessions completed – this equates to less than 66% of planned sessions completed.

Exploring the impact of different levels of intervention receipt on outcomes. We will use two-stage least squares instrumental variables (IV) regression to examine the effect of the intervention in those who receive varying levels of it. The trial allocation will be the instrumental variable in this analysis.

Adherence analysis will use a Two-Stage Least Square approach to estimate the model and Huber-White standard errors reported which are robust to clustering. The R packages 'ivpack' and 'ivreg' will be used to implement the two-stage instrumental variable analysis (Jiang & Small, 2014; Fox Kleiber, & Zeileis, 2021). Compliance (session adherence, i.e. number of sessions) will be instrumented by the intervention allocation (Angrist & Imbens, 1995). The stage 1 model is defined as follows:

$$Compliance_k = \beta_0 + \beta_1 TX_k + \varepsilon_{jk}$$

Predicted values for, $Compliance_k$, from the stage 1 model will be included in the stage 2 model, as follows:

$$Y_{ik} = \beta_0 + \beta_1 \widehat{compliance}_k + \beta_2 baseline_{ik} + \beta_3 Custody_k + \beta_3 VIQ_k + r_{ik}$$

Implementation and process evaluation (IPE) design

Research questions

- RQ5: Does dosage (measured in terms of total number of sessions attended) mediate the vocabulary gains shown by pupils in the two approaches?

The influence of dosage on the effectiveness of the intervention will be addressed as part of the impact evaluation.

Further research questions address the ways in which the effects of the intervention may be moderated by factors linked to characteristics of the Early Years setting or how the approaches are implemented. Private, Voluntary and Independent settings differ from maintained settings in multiple ways, including the age range of children attending, the ratios between staff and pupils and the qualifications of the staff. We anticipate that qualified teachers working with pupils within a small age range may be more effective in implementing the approaches effectively. We therefore hypothesize they are likely to be effective in improving language across both approaches, but particularly in the planned approach.

- RQ6: Does the impact of the intervention differ according to whether the setting is a PVI or maintained setting?

Implementation process evaluation will be used to address questions around the fidelity of implementation, the perceived value of the approaches by practitioners and the role of contextual factors in the success of the implementation and approaches.

- RQ7: Does fidelity to the programme approach mediate the vocabulary gains shown by pupils in the two approaches?
- RQ8: What is the perceived value of the two approaches by practitioners using each approach?
- RQ9: What are the potential costs to the setting of each of the two approaches?

Similarly, we anticipate that the impact of the intervention may be moderated by various characteristics of the storybook reading routine (e.g. whether it is carried out in small groups or large groups), and the range of books available in the setting. However, previous research does not allow us to hypothesize what the effects of these different characteristics are, and so we pose two non-directional research questions:

- RQ10: Does the impact of the intervention differ according to the experience and skills of the educator delivering the storybook sessions?
- RQ11: Does the impact of the intervention differ according to other characteristics such as storybook reading routine or range of books in the setting?
- RQ12: Do the range of books used vary according to which approach is used, and if so, does this moderate the effectiveness of the approach?
- RQ13: What are the characteristics of settings which show good progress in storybook vocabulary in each of the two approaches?

Research methods

The IPE uses a mixed method approach as described below and summarised in Table 8.

Session Records

Educators are asked to complete a brief record of each session including which children attended, which book was read and, for the planned approach, information on the planned vocabulary and discussion points.

Setting level survey (pre-trial)

The setting level survey will provide information on the characteristics of the setting (size of the setting, age range of children) and their standard approach to group storybook reading sessions, including how often they have sessions, whether they are carried out with whole classes or key worker groups, and whether they are integrated with other activities such as story sacks. It is likely that these factors will not be independent of one another – for example, large PVI settings with a wide age range of children may be more likely to have multiple key worker groups for their reading sessions. We intend to assess associations between these categories in order to understand whether there are certain 'types' of settings. This is likely to be by using visualisations such as Venn diagrams, followed by inferential statistics such as chi2 tests of association as appropriate. We will then use this classification of types of settings to guide the analysis of the characteristics of effective settings (RQ13).

Pre-trial Educator Survey

All educators leading storybook sessions will be asked to complete a brief pre-trial survey assessing their qualifications and years of experience in a childcare setting.

We have reviewed the literature for an appropriate measure of preschool educator knowledge of language development or interactive reading with but have not found anything suitable. As a brief proxy measure, we include a page from a well-known children's storybook (*We're Going on a Bear Hunt*) and ask educators to suggest two or three questions that would be suitable to include in a session reading this book. This measure is based upon a longer measure used in a recent paper on assessing the quality of dialogic reading (Towson et al., 2024). We will analyse these questions to assess whether they align to questions suggested by the CROWD framework (e.g. Completion prompts, recall prompts, open ended questions, Wh questions or distancing prompts) or whether they are, for example, focused only on surface aspects of the picture (e.g. where is the dog?). We will review and refine the coding scheme for these responses once we have collected the data. This will be used as a brief measure of storybook reading skill, which will be used as a potential moderator variable in the analyses.

Post-trial settings survey

All settings who have taken part will be asked to complete a post-trial survey giving feedback on the approaches, including whether they consistently followed the approach, whether they thought it was useful and engaging for the children, what the costs to the setting was, and whether they intend to continue with the approach.

Case studies

In the last two weeks of the trial, we will carry out case study observations of eight settings. We intend to work with two PVI settings and two mainstream settings in each of the two approaches. Note that timing means that we cannot select settings based on success in the programme. Instead we will select settings in order to provide a variety of situations, based on the pre-trial setting survey. The case study observations will consist of three elements:

Environmental observation – gathering information about the classroom environment – e.g. size of the space, opportunities for spoken communication, reading and writing, and the extent to which children engage with those opportunities. A key element of this observation will be a review of the number and range of books available for group storybook sessions.

One or more session observations – watching a storybook reading session being delivered. If sessions are delivered in key worker groups, we will observe as many different key workers as possible. An observation checklist will be used to help to understand how storybook choice is managed, the extent to which interactive storybook reading is carried out, the use of planning materials, and the engagement of the children taking part.

*Educator interviews***.** We then plan to interview the educators at the settings in order to gain a full understanding of their views of the approach. We will begin by linking this to the session observation, asking the educator to explain their choices and reflect on their practice in that specific instance, then using this to ask broader questions about how they have used and interpreted the educator guidelines in their setting.

Analysis

Our analytical focus is driven by the research questions – understanding how the programme was implemented and perceived value of the programme, and also understanding the potential moderators of progress in different settings.

Session records

The session records allow us to assess dosage for each individual pupil and fidelity for each setting. They also allow us to assess the range of books which are used in each setting and the extent to which the different approaches lead to different book choices. The session records will be analysed to provide:

- Total number of sessions attended for each pupil
- Total number of books used in each setting
- Frequency of book repetition in each setting

These data will be analysed using descriptive data including frequency counts.

In addition to this quantitative data, we will be able to assess qualitatively the fidelity of each setting to the approach (RQ7) and the range of vocabulary and discussion topics addressed in the planned approach. This will be used to address RQ12.

Setting Survey Data

The setting level survey will provide information on the characteristics of each setting. It is likely that these factors will not be independent of one another – for example, large PVI settings with a wide age range of children may be more likely to have multiple key worker groups for their reading sessions. We intend to assess associations between these categories in order to understand whether there are certain 'types' of settings. This is likely to be by using visualisations such as Venn diagrams, followed by inferential statistics such as chi2 tests of association as appropriate. We will then use this classification of types of settings to guide the analysis of the characteristics of effective settings (RQ13).

Educator survey data

We intend to create a single 'educator knowledge' variable consisting of the experience and qualifications of each educator, plus a measure of storybook reading skill based on the example questions the educator provided. We will assess whether this variable is associated with pupil progress using regression analysis (RQ10).

Post-trial quantitative survey data

Quantitative measures of educator satisfaction will be calculated for the two approaches separately. In order to understand whether satisfaction varies between the two approaches, t-tests will be carried out.

Setting and session observations, interviews and open-ended survey questions.

Observation data will be coded according to preset criteria aligned to the research questions. Interview data will be transcribed. All qualitative data will be analysed using nVivo, in order to allow us to draw themes across the different types of data. We will initially code all data deductively based on the research questions, coding these to multiple questions where relevant. We will then undertake a thematic analysis for each research question by inductively coding the relevant text.

Triangulation of qualitative and quantitative data

To integrate data, we will create an analysis framework outlining the links between the research questions, quantitative data, and themes from the qualitative data. For the eight case study settings, we will combine

the qualitative and quantitative data to construct short case studies with the aim of addressing RQs 11, 12 and 13.

Table 8: IPE Main Elements

IPE dimension	RQ addressed	Research methods	Data collection methods	Sample size and sampling criteria	Data analysis methods
Fidelity	RQ7	**Records**	Session records	All settings (100)	Frequency counts
Dosage	RQ6	**Records**	Session registers	All settings (100)	Frequency counts
Perceived impact	RQ8	**Survey**	Post-trial survey	All educators taking part (c. 150)	Descriptive statistics and thematic analysis
Cost	RQ9	**Survey**	Post-trial survey	All educators taking part (c. 150)	Content analysis
Context/ moderators	RQ11, RQ12, RQ13	**Survey, interviews, observation**	Pre-trial survey	All educators taking part (c. 150)	Descriptive statistics
			Post-trial survey	All educators taking part (c. 150)	Regression analysis
			Post-trial interviews	8 settings, 4 in each approach, 2 PVI and 2 maintained	Thematic analysis
Setting context		**Survey**	Setting survey	All settings (100)	Frequency counts

Ethics and registration

Ethical approval has been granted by Coventry University and ratified by the University of Birmingham. We will gain written opt-in consent from the educators taking part, opt out consent from the parents of the pupils and verbal assent from the pupils themselves at the point of testing. We believe that opt out consent is appropriate in this case because the pupils will be taking part in activities that are part of a standard Early Years routine.

We will contact settings and ask them to express an interest in the trial, either by contacting us by phone or email, or by completing an expression of interest form. We will speak to a representative of the setting over the phone, during a webinar or face to face. If they are willing to take part we will ask them to complete an MOU agreeing to take part in the trial.

The trial is registered at the Open Science Framework (https://osf.io/d54f7).

Data protection

Please see the** Data Protection Statement **for EEF Evaluations.

Data will be shared and stored securely, in accordance with the Data Protection Act (2018) and UK GDPR (General Data Protection Regulation). The University of Birmingham are data controllers for the duration of this trial

We will collect and process the following information about preschool pupils taking part in Story Choices:

- First and last name
- Date of birth
- Home postcode
- Gender
- Children's Unique Pupil Number (a code number to allow us to link to the child's record in the Department for Education's national pupil database)
- Whether the child is eligible for Early Years Pupil Premium
- Whether the child has English as an additional language
- Whether the child has any registered special educational needs or Speech, Language and Communication Needs
- Attendance pattern at the setting in general, and at the storybook reading sessions
- Which primary school the child is likely to attend
- Results from a short vocabulary assessment at the start and end of the trial
- Results from a short questionnaire about attitudes towards books and reading at the start and end of the trial
- Results from a brief language measure (LanguageScreen) at the end of the trial
- Preschool setting name and unique registration number

The University of Birmingham will use this data to conduct an evaluation of the Story Choices trial. As part of this evaluation, we are running a trial in approximately 100 preschool settings in England. This trial will involve staff in the settings following one of two different storybook reading routines in their classrooms. The University of Birmingham will randomly allocate settings to one of two approaches, one where the books read are preplanned by the staff and one where they are chosen by the children. We will assess whether these conditions make any difference to the children's vocabulary and attitudes towards books.

For setting staff (including the key contact person and all staff members leading storybook reading sessions) the following data will be collected:

- Name
- Email address
- Job title
- Number of years of experience in preschool settings

- Any teaching or childcare qualifications
- Knowledge of oral language teaching approaches
- Attitudes and views on the guidance materials for the trial

This data will be used to deliver the trial efficiently and to evaluate the effectiveness of the storybook reading approaches.

We will not use any identifiable staff, pupil, or school data and information in any report that we publish about the Story Choices project.

What is the lawful basis for this data collection and processing?

Under the UK General Data Protection Regulation (UK GDPR), the lawful basis we rely on for processing this information is:

Legitimate Interest. Our lawful basis for processing the personal data listed above is legitimate interests (as per Article 6 (1) (f) of the GDPR) and we have considered that staff and pupil interests and fundamental rights do not override those legitimate interests. It is necessary in the University of Birmingham's 'legitimate interests' to process the personal and special category data identified above in order to learn more about the most effective way to plan storybook reading in Early Years settings. The research project fulfils the University's core business aims including undertaking research, evaluation and information activities in sectors that will deliver social impact.

What will happen to the data?

Most of the personal information we process is provided to us by the child's Early Years setting. We collect this information for the following reasons:

- To take account of the children's age and background characteristics in our statistical analysis
- To allow us to link the children's data collected before and after the trial
- To allow the data to be linked in the future to information held on them on the Department for Education's National Pupil Database (NPD). This is a database containing information about the educational progress of all children within the English state school system.

We will assess children's language using the LanguageScreen assessment which is administered by a researcher in settings using the LanguageScreen app. The data from this app if uploaded to OxEd, the external company which manages the app, and they will provide your data to the research team via their secure LanguageScreen webpage. OxEd will use the data collected using the LanguageScreen app to improve their assessment and for further research into language development; you can opt out of OxEd processing data for this purpose by contacting the research team. Further information about Language Screen can be found on their website: https://oxedandassessment.com/uk/languagescreen/ and their privacy notice is available here: https://media.oxedandassessment.com/assets/OxEdPrivStatement.pdf

At the end of the project we will share this information with FFT Education (FFT). FFT will hold the data until participating children would be expected to appear in the NPD, then match with the NPD and replace these identifiers with the Pupil Matching Reference (PMR). The anonymised dataset containing the PMR (but no direct identifiers) will then be archived in the Education Endowment Foundation Data Archive. The Education Endowment Foundation are data controllers for this archive.

Personnel

Table 9: Personnel on the project

Name	Role	Institutional Affiliation
Julia Carroll	Project lead	University of Birmingham
Sian Alsop	Project co-lead, with responsibility for vocabulary measure development	Coventry University
Tanvir Ahmed	Project co-ordinator	Coventry University
Carlo Tramontano	Statistician (vocabulary measure development)	Coventry University
Paul Thompson	Statistician (main trial)	University of Birmingham

Table 10: Study Advisory Board

Name	Role and Affiliation
Fliss James	Director, East London Research School
Claudine Bowyer-Crane	Professor, University of Sheffield
Courtenay Norbury	Professor, University College London
Louisa Reeves	Director of Policy and Evidence, Speech and Language UK
Konstantinos Skordos	Pedagogy manager, LEYF nurseries

Risks

Table 11: Risk Analysis

Risk	Likelihood	Impact	Mitigation
Settings not keen to be randomised	unlikely	moderate	We have framed the project as a comparison between two approaches rather than an intervention and a control condition. Scoping phase indicated the approaches are similar to everyday practice in many settings Settings will be made aware of the randomisation early in recruitment
Insufficient settings recruited to the trial	May happen	severe	Detailed recruitment plan with significant time and resources devoted to it Expression of interest process Monitoring recruitment
Setting attrition	May happen	Moderate	Clear initial and ongoing communication with settings Minimise burden on settings Provide incentive at the end of the trial Over-recruit to allow for this
Pupil attrition	Likely	mild	Keep trial period within one school year Over-recruit to allow for this

Non-random attrition of settings (e.g. PVI settings more likely to drop out)	May happen	Mild	Over-recruit to allow for this. Ensure PVIs understand the expectations and discuss how it can fit in to their routine
Change in staff in settings	May happen	moderate	Record contact details from multiple staff members. Check in regularly with settings to see if there are changes.
Change in staff in evaluation team	unlikely	mild	Ensure resources are shared and all staff have a good understanding of the processes and timeline
Educators do not follow guidance	May happen	moderate	Monitor fidelity through the trial Make expectations clear Check in regularly with settings to see if there are any problems
Delay to assessment development	unlikely	moderate	Build in possibility of additional casual RA support to speed up testing
Not possible to develop valid assessment	Unlikely	Severe	Plan assessment development well in advance of pretesting
Children do not respond as expected to the assessment	Unlikely	Moderate	Pilot measures before pretesting
Floor or ceiling effects in the measures	unlikely	Mild	Pilot measures before pretesting, and add items if needed

Timeline

Table 12: Setup and evaluation timeline

Dates	Activity	Staff responsible/ leading
Feb-May 2024	Set up meetings, create documentation, recruit research fellow	Julia Carroll
May 2024	Launch setting recruitment	Julia Carroll
May 2024	Create vocabulary database	Sian Alsop
December 2024	First draft of study plan complete	Julia Carroll

Sept. – Oct. 2024	Phase 1 piloting of vocabulary measure	Tanvir Ahmed
Oct.-Dec. 2024	Phase 2 piloting of vocabulary measure	Julia Carroll
31 October 2024	Complete setting recruitment	Tanvir Ahmed
Jan. – Feb 2025	Pre-testing	Tanvir Ahmed
Feb-May 2025	Trial delivery	Julia Carroll
1 June 2025	Study plan complete and trial registered on OSF	Julia Carroll
May-June 2025	Process evaluation data collection	Julia Carroll
June-July 2025	Post-testing	Tanvir Ahmed
August-October 2025	Data analysis and evaluation	Julia Carroll
November 2025	First draft report delivered	Julia Carroll

References

Angrist, J. D. and Imbens, G. W., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity, *Journal of the American Statistical Association,* 90(430), 431-442, 10.1080/01621459.1995.10476535

Carroll, J.M., Holliman, A.J., Weir, F. and Baroody, A.E., 2019. Literacy interest, home literacy environment and emergent literacy skills in preschoolers. *Journal of Research in Reading*, *42*(1), pp.150-161.

Coyne, M.D., McCoach, D.B. and Kapp, S., 2007. Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disability Quarterly*, *30*(2), pp.74-88.

Dobson, A.J., & Barnett, A.G. (2018). An Introduction to Generalized Linear Models (4th ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315182780

Dong, N., and Maynard, R. A., 2013. PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67.
doi: 10.1080/19345747.2012.673143.

Dowdall, N., Melendez-Torres, G.J., Murray, L., Gardner, F., Hartford, L. and Cooper, P.J., 2020. Shared picture book reading interventions for child language development: A systematic review and meta-analysis. *Child Development*, *91*(2), pp.e383-e399.

Fox, J., Kleiber, C. and Zeileis, A., 2021. ivreg: Instrumental-Variables Regression by '2SLS', '2SM', or '2SMM', with Diagnostics. R package version 0.6-1. https://CRAN.R-project.org/package=ivreg

Hadley, E.B., Dickinson, D.K., Hirsh-Pasek, K., Golinkoff, R.M. and Nesbitt, K.T., 2016. Examining the acquisition of vocabulary knowledge depth among preschool students. *Reading Research Quarterly*, *51*(2), pp.181-198.

Harrell, F.E. (2001) Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York. http://dx.doi.org/10.1007/978-1-4757-3462-1

Hedges, L. V., 2007. Effect Sizes in Cluster-Randomized Designs. Journal of Educational and Behavioral Statistics, 32(4), 341–370. https://doi.org/10.3102/1076998606298043

Jiang, Y. and Small, D., 2014. ivpack: Instrumental Variable Estimation. R package version 1.2. https://CRAN.R-project.org/package=ivpack

Kahan, B. C., and Morris, T. P., 2012. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ (Clinical research ed.)*, *345*, e5840. https://doi.org/10.1136/bmj.e5840

Sénéchal, M., 1997. The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *Journal of Child Language*, *24*(1), pp.123-138.

Senn, S., 1994. Testing for baseline balance in clinical trials. *Statistical Medicine, 13*: 1715-1726. https://doi.org/10.1002/sim.4780131703

Skene, K., O' Farrelly, C.M., Byrne, E.M., Kirby, N., Stevens, E.C. and Ramchandani, P.G., 2022. Can guidance during play enhance children's learning and development in educational contexts? A systematic review and meta-analysis. *Child Development*, *93*(4), pp.1162-1180.

Towson, J.A., Macy, M., Abarca, D.L., Myers, K. and FitzPatrick, E., 2023. Examining teachers' use of dialogic reading strategies following a multiple component professional development intervention. *Early Childhood Education Journal*, *52*, 1751–1763. https://doi.org/10.1007/s10643-023-01526-3

Toub, T.S., Hassinger-Das, B., Nesbitt, K.T., Ilgaz, H., Weisberg, D.S., Hirsh-Pasek, K., Golinkoff, R.M., Nicolopoulou, A. and Dickinson, D.K., 2018. The language of play: Developing preschool vocabulary through play following shared book-reading. *Early Childhood Research Quarterly*, *45*, 1-17.

van Buuren, S., Boshuizen, H. C., and Knook, D. L., 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*(6), 681–694. https://doi.org/10.1002/(sici)1097-0258(19990330)18:6<681::aid-sim71>3.0.co;2-r
van Buuren, S., and Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45(3)*, 1-67. DOI: 10.18637/jss.v045.i03.

West, G., Lervåg, A., Birchenough, J. M., Korell, C., Rios Diaz, M., Duta, M., ... and Hulme, C., 2024. Oral language enrichment in preschool improves children's language skills: a cluster randomised controlled trial. *Journal of Child Psychology and Psychiatry, 65(8),* 1087-1097. https://doi.org/10.1111/jcpp.13947

White, I. R., Royston, P., and Wood, A. M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Whitehurst, G.J., Falco, F.L., Lonigan, C.J., Fischel, J.E., DeBaryshe, B.D., Valdez-Menchaca, M.C. and Caulfield, M., 1988. Accelerating language development through picture book reading. *Developmental Psychology*, *24*(4), p.552.

Appendix

Table 13: Key uncertainties at evaluation stage

Choice definition & theoretical uncertainties	Assessment after scoping phase	Learning objectives for full evaluation
What does the research say about planning interactive storybook reading sessions?	There is extensive evidence that interactive reading is effective in improving literacy outcomes. However, we could not find any evidence addressing the role of planning in interactive reading.	Does planning storybook sessions make a significant difference to language outcomes of preschool children?
What is 'business as usual' with regard to group storybook reading sessions?	Educator survey indicated that approximately 1/3 of settings planned their sessions, 1/3 were spontaneous and 1/3 used a mixed approach	To what extent do the approach guidelines mimic what settings would be doing anyway? To what extent do the educators find the approaches easy to implement?
Implementation uncertainties		
How can we ensure that educators are using interactive reading in their sessions?	Educators largely say that they are carrying out interactive reading, but experts widely believe that there is great variation in practice. We could not find a brief measure of storybook reading practice suitable for a range of early years educators.	In order to assess individual differences in educator knowledge, we take a brief measure of their experience and qualifications and a brief measure of planning appropriate storybook reading prompts. Are these measures associated with individual pupil outcomes?

Evaluation & methodological uncertainties		
How can we sensitively measure improvement in storybook related vocabulary?	Expressive vocabulary measures tend to be more sensitive to change than receptive vocabulary measures when looking at storybook reading intervention research. We would like the measure to be closely aligned to the vocabulary the children are learning. However, expressive vocabulary tests for this age group tend to be picture naming tasks, which test only name knowledge for easily pictured items.	Does the storybook vocabulary measure efficiently and reliably measure change in vocabulary knowledge in this age group?
How can we manage the differences between PVI and maintained settings?	There are many differences between PVI and maintained settings that may influence outcomes. We will try to measure some of these differences.	Do the characteristics of the settings have an influence pupil outcomes?