



Education
Endowment
Foundation

Story Choices—Teacher Choices

Evaluation report



UNIVERSITY OF
BIRMINGHAM



The Education Endowment Foundation is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, nurseries and colleges to improve teaching and learning for 2 – 19-year-olds through better use of evidence.

We do this by:

- **Summarising evidence.** Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds.
- **Putting evidence to use.** Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were re-endowed with an additional £137m, allowing us to continue our work until at least 2032.

For more information about the EEF or this report please contact:



Education Endowment Foundation

5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP



0207 802 1653



info@eefoundation.org.uk



www.educationendowmentfoundation.org.uk

Contents

Executive summary	4
Introduction.....	6
Methods	14
Impact evaluation results.....	28
Implementation and process evaluation results	52
Conclusion	61
References	66
Appendix A: Security classification of trial findings.....	68
Appendix B: Effect size estimation	70
Further appendices	71

About the evaluator

The project was evaluated by a team from the University of Birmingham in partnership with Coventry University. The trial director and principal investigator was Professor Julia Carroll. The trial manager was Dr Tanvir Ahmed. Dr Sian Alsop led the development of the vocabulary assessment measure and Dr Paul Thompson and Dr Carlo Tramontano acted as trial statisticians.

Professor Julia Carroll and Dr Tanvir Ahmed
School of Education
University of Birmingham
Pritchatts Road
Birmingham, B15 2TT

j.m.carroll@bham.ac.uk

Dr Paul A. Thompson
Intellectual Disabilities Research Institute (IDRIS)
School of Social Policy and Society
University of Birmingham
Pritchatts Road
Birmingham, B15 2TT

p.a.thompson@bham.ac.uk

Acknowledgements

We are grateful to all the staff and children in the nurseries and preschool settings that participated in this evaluation. Their enthusiasm for the research was inspiring. We are especially grateful to settings that took part in the interviews and observations, as well as those who responded to survey requests.

We would like to thank the team at the Education Endowment Foundation, including Lauren Spinner, Beverley Jennings, Faizaan Sami, Jennifer Magness, Tom MacKay, Christine Kelley, and Sarah Tillotson.

Executive summary

The project

This project aimed to contribute to the evidence base around interactive storybook reading practices in the early years. Dedicated interactive storybook reading time is a common and evidence-based practice in early years settings but there is variation in how such sessions are planned and delivered. This project compared the impact on children’s vocabulary and oral language skills of two different but commonly used approaches. In the child-led storybook reading approach, the educator allowed children to select the book to be used and gave prompts spontaneously in response to child interest. In the planned storybook reading approach, the educator selected the book in advance and planned prompts to use during the session. As the benefits of storybook reading are well established, it was not felt that a ‘no treatment’ control was required. The evaluation team developed brief written guidance to send to educators on using their allocated approach and prompt cards with reminders of the approach. Educators were asked to deliver the storybook sessions to groups of children as part of the daily routine of their early years class, to use their setting’s existing storybooks for the trial, and to deliver the storybook session at least three times a week for at least 15 minutes per session over 12 weeks (from February 2025 to May 2025). The children were all due to start reception in the September of 2025 (that is, they were aged 3:0 to 3:11 years in September 2024).

This trial was an Education Endowment Foundation (EEF) funded Teacher Choices trial. Teacher Choices trials aim to test everyday choices educators make when planning their learning sessions and supporting children. Nine hundred and sixty-four children in 82 settings took part in the project, including 58 maintained settings and 23 private, voluntary, and independent (PVI) settings. The project was a two-arm, parallel group, cluster randomised controlled trial (RCT) randomised at the setting level. The primary outcome measure was a bespoke storybook vocabulary assessment developed specifically for this trial, based on the vocabulary commonly found in storybooks for preschool children. The secondary outcome measures were a standardised measure of oral language skills and an attitude to literacy measure. The implementation and process evaluation explored the background characteristics of the settings and educators involved in delivering the sessions, the feasibility of the approaches, dosage, fidelity, and perceived engagement and effectiveness. The trial started in January 2025 and concluded in July 2025.

Table 1: Key conclusions

Key conclusions	
1.	Children in both groups made significant progress over time on the bespoke storybook expressive vocabulary measure. The mean score in the child-led group (32.27) was 1.22 points higher, on average, compared to the mean score in the planned group (31.05). The result has a moderate to high security rating.
2.	Children eligible for EYPP in the child-led group achieved a score (31.25) that was 1.81 points higher in storybook expressive vocabulary, on average, compared to children in the planned group (29.44). These results have a lower security rating than the overall findings as they were based on a smaller sample of children.
3.	Children in the child-led group achieved a score in wider oral language skills (100.07) that was, on average, 3.98 points higher than that in the planned group (96.10).
4.	There was no evidence of a difference in attitudes to literacy between the children in the planned storybook settings, on average, and those in the child-led settings.
5.	Many of the educators in the child-led group felt that the child-led approach led to high child engagement and empowerment around storybook reading, while some of those in the planned approach felt that children were sometimes not very engaged. In addition, children in the child-led condition were, on average, exposed to a wider variety of books, perhaps due to letting the children choose. Educators in the child-led group also felt that that approach led to more opportunities for expressive language use by the children.

EEF Security rating

These findings have a moderate to high security rating. The trial was well-powered, however, the fact that 29.8% of the pupils who started the trial were not included in the final analysis reduced its security.

Additional findings

On average, storybook expressive vocabulary scores were higher at the end of the trial than at baseline for children in both groups. However, the children in the child-led group scored 1.22 points higher in storybook expressive vocabulary at the end of the trial, on average, compared to those in the planned group (ES = -0.19; 95% CI: -0.43, 0.05). As with any study, there is always some uncertainty around the result: the possible difference in impact between the two groups ranged from a 2.72 point advantage in the child-led group to a 0.26 point advantage in the planned group. The result has a moderate to high security rating. This result was consistent across PVI and maintained settings. The findings were also consistent across a broader measure of oral language skills. Children in the child-led group scored 3.53 points higher on the LanguageScreen measure at the end of the trial, on average, compared those in the planned group (ES = -0.30, 95% CI: -0.55, -0.06). The possible difference in impact between the two groups ranged from a 0.77 point to a 6.29-point advantage for the child-led group.

The study predicted that the planned approach would lead to higher language outcomes as children would be exposed to a wider range of books and planned vocabulary teaching. It is therefore somewhat surprising the child-led approach was more effective. Analysis of session records indicated that, contrary to predictions, the child-led approach was associated with exposure to a wider range of books. No differences were found between approaches in attitude towards literacy; however, this may reflect limitations in the measurement rather than a true absence of effect. There were some indications in the feedback from practitioners in the implementation and process evaluation (IPE) that children in the child-led approach were more engaged during sessions than children in the planned approach. There were also suggestions from practitioners in the IPE that the child-led approach gave children good opportunities to use language to express their preferences and ideas and opportunities for children with lower language levels to rehearse and become confident with favourite texts. More research is needed to understand which of these factors is most important in driving the effectiveness of the child-led approach.

Impact

Table 2: Summary of impact on primary and secondary outcomes

Outcome/ group	Effect size (95% confidence interval) Hedges <i>g</i>	EEF security rating	No of children	<i>p</i> -value
Storybook vocabulary measure—all children	-0.19 (-0.43, 0.05)		691	0.110
Storybook vocabulary measure—EYPP children only	-0.60 (-1.43, 0.24)	N/A	94	0.170
Oral language standardised score—all children	-0.30 (-0.55, -0.06)	N/A	671	0.014

Note that here and throughout, the child-led approach was considered 'condition 1' and the planned approach was considered 'condition 2'. Therefore, a negative effect indicates higher scores in the child-led approach than in the planned approach and a 95% confidence interval that crosses zero indicates we cannot be confident that there is an effect.

Introduction

Overview of evaluation design

Context for the project

There are several reasons to value storybook reading as a learning activity in preschool. Books tend to introduce children to a wider range of vocabulary and language structures than spoken language (Nation et al, 2022) and therefore reading books aloud with pre-literate children provides a strong basis for developing their language skills.

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 (for example, 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). For these reasons, the EEF provides guidance in its Early Years Store on the importance of effective storybook reading. However, there is relatively little research around how interactive storybook reading should be implemented within preschool classrooms.

About this project

This project was commissioned because within the Teacher Choices framework the EEF was interested in addressing research around how best to support language development in preschool. The project was preceded by an initial scoping phase in which we considered the viability of four different research topics. This scoping phase is described in more detail in Appendix C.

As part of this scoping phase, we conducted an educator survey that included questions about standard practice around storybook reading in preschool settings. Our survey indicated that all respondents recognised the value of storybook reading and they all included it, in some form, in their weekly routine. Similar proportions of the sample used each of the three different proposed approaches: pre-planned sessions, those 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.

Also within the scoping phase we carried out pilot studies giving educators a draft of the educator guidance (shown in Appendix D) and asking them to complete the record forms as indicated, working with a small group of around six children. Feedback on the storybook reading approaches was all positive. Educators found the guidance clear and the videos demonstrating interactive storybook reading on the EEF website useful, indicating that implementing the approaches would be feasible.

When developing the theory of change and the logic model (Figure 1, page 8), we decided to contrast two parallel approaches to storybook reading rather than an intervention and a ‘business as usual’ control. This was because the survey during the scoping phase indicated that the two approaches were in equally common use and, therefore, it would be difficult to know what ‘business as usual’ would be.

Given that the Teacher Choices commission focused on improving early language outcomes, we wanted language to be our primary outcome measure. Given the small changes between the conditions, we anticipated small effects on standardised measures of oral language skill. We therefore developed a bespoke vocabulary measure which was aligned with the type of vocabulary regularly encountered in storybook reading to use as our primary outcome measure. We included a standardised measure of oral language skill, ‘LanguageScreen’, as our secondary outcome measure.

Within the logic model we hypothesized that the planned approach would lead to higher quality language interactions and would therefore have more of a positive effect on vocabulary and oral language than the child-led approach. In contrast, we hypothesised that the increased autonomy children felt in the child-led approach would help to foster their engagement and

enjoyment of the reading experience, thereby resulting in more positive attitudes towards literacy activities. We therefore included another secondary outcome measuring attitudes towards literacy activities.

Teacher Choice approaches

Planned approach

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 children: as the logic model (Figure 1) highlights, the planned storybook reading approach may be particularly valuable for children from disadvantaged backgrounds.

Child-led approach

As detailed in the logic model, we anticipate that the child-led approach will help to improve children's attitudes towards literacy activities as they will experience greater choice and autonomy around books, which is shown to be an important factor in positive reading attitudes.

Details of the nature of the two contrasting interventions are given in Table 3, page 9.

Figure 1: Story Choices logic model

Boxes in green indicate elements specific to the child-led approach; boxes in yellow indicate elements specific to the planned approach.

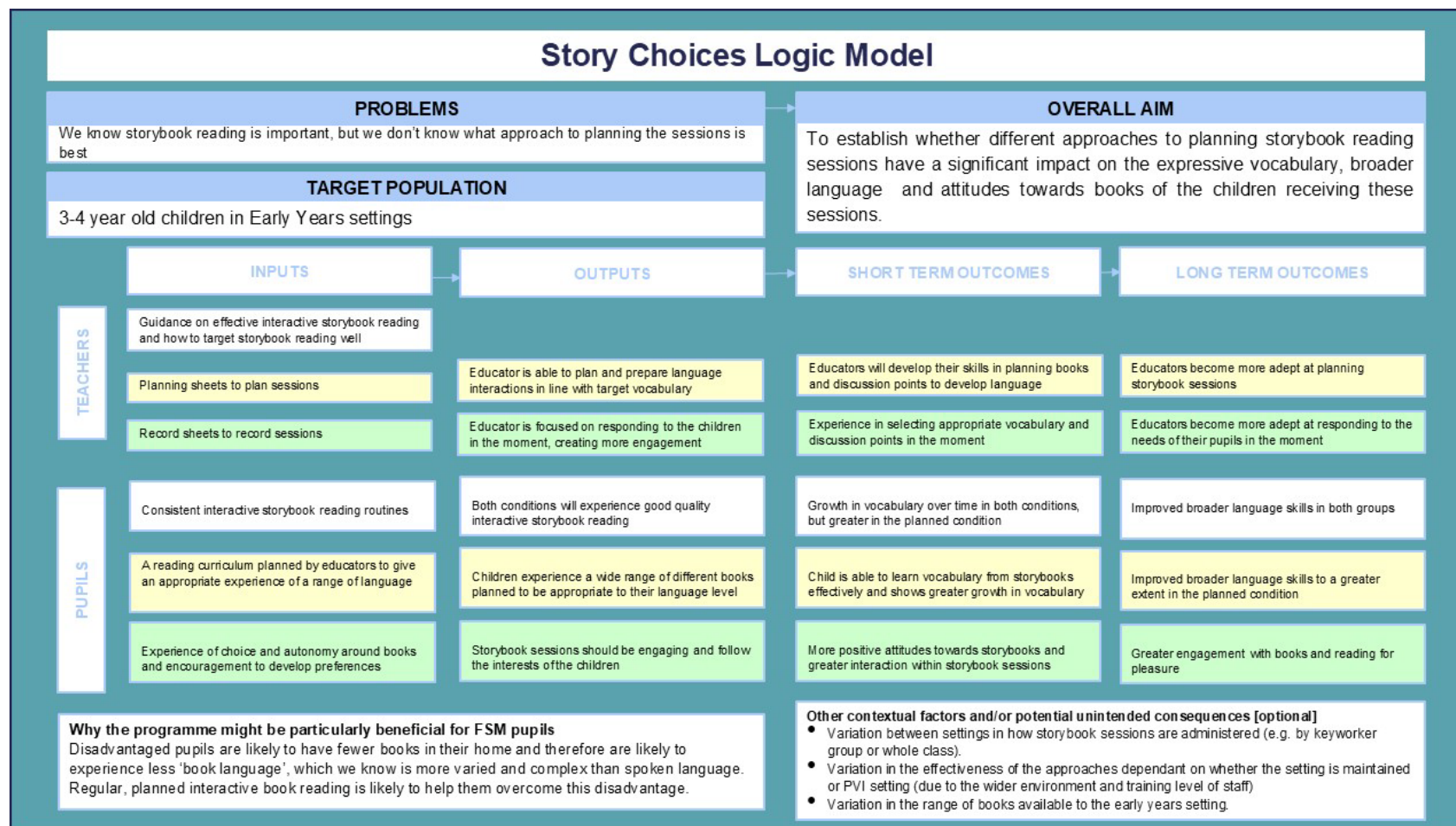


Table 3: Intervention description

Name	Planned storybook sessions	Child-led storybook sessions
Why (rationale)	Approximately one third of survey respondents used 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 used 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 preliterate children listening to stories in a group.
Who (recipients)	Three- to four-year-old children in Early Years settings.	
What (materials)	Early years settings were asked to use their existing storybooks for these sessions to effectively reflect day to day practice. Educators were given guidance on how to plan the sessions, including access to videos of an interactive storybook reading session. They received planning sheets which they were asked to complete before the sessions.	Early years settings were asked to use their existing storybooks for these sessions to effectively reflect day to day practice. Educators were given guidance on how to plan the sessions, including access to videos of an interactive storybook reading session. They received record sheets which they were asked to complete after each session.
What (procedures)	Educators followed their normal routine for group storybook reading (for example, whole-class or key worker groups) except that sessions were planned beforehand by the educator. The educator was asked to record ahead of time which book they will use and what key discussion points and vocabulary they will highlight during the session. They were given a record form for this purpose. After the session they were asked to complete a brief reflection saying whether the session went as planned.	Educators followed their normal routine for group storybook reading (for example, whole-class or key worker groups) except that sessions were not planned beforehand by the educator. Instead, the book to be read was selected by one or more of the children from a choice of at least five. The educator was 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 did not receive any specific training but were asked to follow some written guidance. We collected details about the educators’ qualifications, experience, and knowledge of oral language pedagogy.	
How (delivery guide)	Storybook sessions were delivered in groups to the children in the class as part of the daily routine of the class.	
Where (location)	Early Years settings.	
When and 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.	
Tailoring (adaptation)	Educators could choose the books they used 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 provided children with a range of books from which to choose. They used procedures of their choice to ensure that each child got an equal opportunity to choose—for example, selecting a name out of a hat or having the names on a rota.

Evaluation objectives

The research questions for the study are shown in Table 4. For further details, please see the Research Methods section.

Table 4: Overview of research objectives

Research objective	Research questions	Methodology area	Data collection methods and participants	Data analysis method
Impact Evaluation				
Compare the impact of storybook reading approaches to storybook vocabulary measure	RQ1 (primary): What is the difference in vocabulary knowledge measured by a storybook vocabulary test of early years children participating in planned storybook reading sessions in comparison to children participating in child-led storybook reading sessions?	Impact evaluation	Pupil assessment	Multi level mixed model; intention to treat analysis
Consider whether the impact differs for children eligible for EYPP	RQ2: What is the difference between conditions for children eligible for Early Years Pupil Premium?	Impact evaluation	Pupil assessment	
Compare the impact of storybook reading approaches on attitudes to literacy activities	RQ3: What is the difference in attitudes towards literacy activities between the two conditions?	Impact evaluation	Pupil assessment	
Compare the impact of storybook reading approaches on oral language skills	RQ4: What is the difference between the two conditions in oral language skills measured by the LanguageScreen measure?	Impact evaluation	Pupil assessment	
Compare the impact across PVI and maintained settings	RQ5: Does the difference between the two conditions differ according to whether the setting is a PVI or maintained setting?	Impact Evaluation	Pupil assessment	
Examine the role of setting compliance to the guidance in overall effects	RQ6: Does compliance to the approach mediate the vocabulary gains shown by children in the two approaches?	Implementation compliance	Pupil assessment	
Implementation and Process Evaluation (IPE)				
Implementation	RQ7: How, and how well, are the two conditions implemented in different settings?	IPE	Session logs; interviews and observations	Frequency counts; thematic analysis
Examine the impact of dosage to effectiveness of the approach	RQ8 Does dosage (measured in terms of total number of sessions) mediate the vocabulary gains shown by children in the two approaches?	IPE	Post trial setting survey	Multi-level model
Consider the feasibility and acceptability of the two approaches to the educators	RQ9: what is the perceived value of the two approaches from the practitioners using it?	IPE	Post-trial setting survey; educator survey	Thematic analysis

Consider the role of contextual factors that might mediate the effectiveness of the approach	RQ10: Does the impact of each of the two approaches differ according to other characteristics such as storybook reading routine or range of books in the setting?	IPE	Post trial setting survey; observations	Content analysis; multi-level model; thematic analysis
--	---	-----	---	--

The study plan is published on the EEF website here:

https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/story-choices_study-plan_v.1.0.0.pdf?v=1769605655

Overall aim

The overall aim was to establish whether different approaches to storybook reading sessions have a significant impact on the vocabulary, oral language skills, and attitudes towards literacy activities of the children receiving these sessions.

Ethics and trial registration

The trial was reviewed and approved by the Coventry University ethics committee on 11 May 2023, approval number 151846. After the trial was transferred to the University of Birmingham, the ethical approval was ratified by the University of Birmingham ethics team.

After settings had provided an expression of interest they were sent an information sheet and a Memorandum of Understanding (MOU), which detailed what would be involved in the trial and the responsibilities of the research team and the settings, respectively. Settings were only included in the trial if they signed and returned the MOU. Information provided to educators and parents is in Appendix D, including the MOU for settings, setting and parent information sheets, and the Data Privacy Notice.

After transfer to the University of Birmingham, we informed all settings of the change and the change to how their data would be stored and we asked them to sign and complete a revised MOU.

At the start of pre-testing, settings were provided with an information sheet for parents with an opt-out consent form. Pre-testing was only carried out after settings confirmed that they had sent out the information sheet and that children whose parents had opted out were not included in the trial. Information about data storage, anonymity, and right to withdraw were shared at the start of all online surveys sent to staff with a separate consent page.

Preschool children may not be verbally able to give or refuse consent to take part and so we monitored children during testing for signs that they did not want to continue. Testing ceased if a child was deemed to have withdrawn consent.

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

Data protection

Data was shared and stored securely in accordance with the Data Protection Act (2018) and the U.K. GDPR (General Data Protection Regulation). Coventry University was the data controller for the vocabulary pilot data; the University of Birmingham was data controller for the duration of evaluation trial.

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 approach to deliver 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.

We collected and processed the following information about preschool children taking part in Story Choices:

- first and last name;
- date of birth;
- home postcode;
- gender;
- 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;
- attendance pattern at the setting in general, and at the storybook reading sessions;
- results from a short vocabulary assessment at the start and end of the trial;
- results from a short questionnaire about attitudes towards literacy activities and reading at the start and end of the trial;
- results from an oral language skills measure ([LanguageScreen](#)) at the end of the trial; and
- preschool setting name and unique registration number.

For setting staff—including the key contact person and all staff members leading storybook reading sessions—the following data was collected:

- work setting;
- job title;
- number of years of experience in preschool settings;
- any teaching or childcare qualifications;
- knowledge of oral language teaching approaches; and
- attitudes and views on the guidance materials for the trial.

Data archiving and deletion

After this report is published in July 2026, participant data and information from the educator surveys will be transferred and stored in the EEF archive for future research. The EEF archive is managed by FFT on behalf of the EEF and hosted by the Office of National Statistics (ONS). Names and other meaningful identifiers will be removed before the data is added to the archive. At this point the EEF becomes the data controller and will keep information in its archive for as long as it is needed for research purposes. Please see the EEF's [archive guidance](#) and its [privacy notice](#) for more information on how the EEF processes and will use personal data. The EEF and other research teams will be able to access the data as part of subsequent research through the ONS Approved Researcher Scheme: this scheme is used by the ONS to grant secure access, for statistical research purposes, to data that cannot be published openly, as permitted by the Statistics and Registration Service Act 2007 (SRSA). Audio recordings from teacher interviews were transcribed and deleted within one month of the interview date. All other personal data held by the University of Birmingham will be deleted within one year of publication of this report. This means the personal data will be deleted by October 2027.

Project team

Table 5: The project team

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	University of Birmingham
Carlo Tramontano	Statistician (vocabulary measure development)	Coventry University
Paul Thompson	Statistician (main trial)	University of Birmingham
Anila Ahad	Research assistant	University of Birmingham
Molly Bonnington-Brown	Research assistant	University of Birmingham
Kathryn Cawley	Research assistant	University of Birmingham
Jennie Cook	Research assistant	University of Birmingham
Rachael Neale	Research assistant	University of Birmingham
Fiza Shah	Research assistant	University of Birmingham
Luisa Tarczynski-Bowles	Research assistant	University of Birmingham
Zoe Vickers	Research assistant	University of Birmingham

Study advisory board

Table 6: The 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

Methods

Trial design

Table 7: Trial design

Trial design, including number of arms		Two-arm, parallel group cluster RCT-arm, cluster randomised.
Unit of randomisation		Setting.
Stratification variable(s) (if applicable)		Setting type (PVI versus maintained), region.
Primary outcome	Variable	Storybook vocabulary.
	Measure (instrument, scale, source)	Vocabulary measure focused on storybook language, 0–46, researcher-designed (provided in Appendix E).
Secondary outcome(s)	Variable(s)	1. LanguageScreen Expressive Vocabulary; 2. LanguageScreen Listening Comprehension; 3. LanguageScreen Receptive Vocabulary ; 4. LanguageScreen Sentence Repetition ('SR'); 5. Attitudes towards literacy activities ('ATLA').
	Measure(s) (instrument, scale, source)	LanguageScreen (standard score, OxEdandAssessment.com); revised version of the ATLA measure, researcher designed (provided in Appendix F).
Baseline for primary outcome	Variable	Storybook vocabulary.
	Measure (instrument, scale, source)	Expressive vocabulary measure focused on storybook language, 0–46, researcher-designed (details provided in Appendix E).
Baseline for secondary outcome(s)	Variable	The pre-test ATLA measure will act as baseline for the post-test ATLA measure; the pre-test storybook vocabulary measure will act as the baseline measure for the post-test LanguageScreen measures.
	Measure (instrument, scale, source)	As described above.

The trial was a two-arm, parallel group cluster randomised controlled study randomised at the level of setting. The primary outcome measure was a storybook vocabulary measure, while the secondary outcome measures were a standardised language measure and an attitudes towards literacy activities measure.

Participant selection

Selection criteria were set to allow unrestricted enrolment from a wide range of settings and children. Participants were children in group-based early years settings in the year before they start reception (ages 3:0 to 3:11 years in September 2024, turning four before 1 September 2025). Settings were eligible to take part if they were eligible to receive government funding for childcare hours, have at least eight 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 were not taking part in another research trial relating to language or literacy.

Recruitment was carried out between May and October 2024. Settings were recruited through the EEF website and newsletters alongside postcards sent to eligible settings within the target areas and meetings with key council staff at the relevant local authorities. Interested settings completed an online expression of interest to allow the team to assess their eligibility. Eligible settings were then sent an information sheet and a MOU. Settings signed up to the trial by signing and returning the MOU and providing a key contact name for communication. At the end of the trial, settings received £200 as a thank-you for their participation.

Eighty-four settings signed up to take part in the trial: two dropped out before pre-testing leaving 82 to be randomised. Within the settings, we tested a random sample of children who met eligibility criteria, up to a maximum of 20 in any setting. We used a random number generator matched to a list of children's names for sample selection.

Notably, settings were asked to carry out the storybook reading approaches as part of their regular routine. Sessions were not limited to those children who completed pre-testing and could be delivered either as whole-class sessions or smaller keyworker sessions, according to the setting routine.

Outcome measures

Below we provide details on the outcome measures: key details of these are summarised in Table 8.

Table 8: Outcome measures

Baseline measure	Definition
Storybook vocabulary measure	A vocabulary measure focused on storybook language and based on the 'informal definitions' task
Attitudes towards literacy activities	A Likert scale questionnaire based on the ATLA measure
Primary outcome	
Storybook vocabulary measure	Repetition of previous measure (same set of items used)
Secondary outcomes	
LanguageScreen	A standardised measure of oral language skills, including expressive and receptive vocabulary and narrative skills
Attitudes towards literacy activities	Repetition of previous measure

Primary outcome—storybook vocabulary measure

A bespoke measure was created for this trial—the storybook vocabulary measure—details on the development of which are provided in the study plan and are summarised in Appendix E. The measure is based on the 'informal definitions task' used by Hadley et al. (2016) and uses vocabulary commonly found in storybooks aimed at three- to five-year-old children. This measure was developed to align with the expectation within the logic model that the approaches would have their strongest effect on expressive vocabulary of the words within the storybooks they were reading. Children were asked to say or show what they know about each word, and their responses are scored on a scale of zero to two. The measure has 23 items in total, distributed across four subscales: one using dolls to act out word meanings (four items), one requesting actions or gestures (four items), one using adjectives (five items), and one using nouns (seven items). The measure was administered individually with children by research assistants who were provided with a list of potential correct responses. Research assistants were asked to write down what a child said or gestured for each item to allow checking of the consistency of scoring across testers. Raw scores are used in all analyses. Inter-rater reliability was assessed on a subset of 10% of pre-test responses and was found to be high (ICC = 0.951). Sample-specific internal reliability of the pre-test data was good (Cronbach's Alpha = 0.857).

Secondary outcomes

Attitudes towards literacy activities (ATLA) measure

This measure is provided in Appendix F. This was also administered by the research assistants in a one to one situation. The 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 child-led approach may help to build positive attitudes. This is a brief Likert scale questionnaire based on the work of Carroll, Holliman and 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 added pictures to make the task clearer for young children and added 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 seven literacy-related items rated on a scale of one to four, to give a maximum score of 28. We used 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 ten to 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. Standardised scores were used in the analysis. More details can be found here: <https://oxedandassessment.com/uk/languagescreen/>.

We did not use National Pupil Database data within this study.

Settings were allocated to their assigned approach after the pre-testing was complete. Post-testing for all measures was carried out by research assistants who were blind to the approach used by the setting, with the exception of 22 children across six different settings who were tested by the project staff due to time constraints.

Sample size

Table 12 summarises the calculations for planned sample size alongside the achieved sample size.

Sample size was determined using the PowerUp! Software (Dong and Maynard, 2013), using the function `cra2_r2` (two-level cluster-randomized design where the clusters are randomly assigned to treatment and control). For all of the scenarios listed, we have estimated the MDES based on the following assumptions and a two-level linear mixed effects design with preschool setting at level two and controlling for baseline outcome.

Intraclass 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- and 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, for example, Hadley et al. (2016).

We modelled the MDES for two different scenarios. The first indicated the minimum sample size needed to detect an effect size of 0.3, working on the basis of an average cluster size of eight children (based on the minimum setting size set in recruitment criteria) and calculating attrition at 10% at setting level and 20% at individual level. It was agreed with the EEF that 0.3 was an acceptable MDES in this case. EEF studies are often powered to detect smaller effect sizes (for example, a Hedges *g* of 0.2), however, powering this study to detect this effect size would have resulted in an unrealistically large target sample (over 200 settings), given previous difficulties in recruiting early years settings to EEF trials. It was decided that powering to an effect size of 0.3 and using a bespoke measure focused on the particular age group and outcome, and, therefore, potentially more sensitive to change due to the assigned approaches, was an acceptable compromise. According to the pilot analysis the difference between preschool and reception age children equated to a Hedges *g* effect size of 0.89. On this basis, an effect size of 0.3 would equate to around four months' additional progress, which was felt to be educationally significant.

Table 12 (column two) indicates the MDES for children eligible for Early Years Pupil Premium (EYPP) based on this sample size.¹ 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. The trial is powered to detect an effect size larger than 0.46 on the primary outcome variable for the EYPP subgroup.

Within both scenarios, we did not recruit to target, meaning that we were somewhat underpowered for the target MDES.

Based on these scenarios, we intended to recruit 100 settings to take part and aim to test 12 children per setting. This would have allowed 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. As preschool settings vary widely in the number of children attending, we needed to be flexible in the cluster size at each setting, but too much variation in cluster size would cause a slight reduction in statistical power. We settled on a minimum cluster size of eight and a maximum of 20. In settings where there were more than 20 children, we created a randomised list of children prior to testing and followed the order of the list to avoid potential bias that comes with self-selection or teacher selection of children. We assumed an average of 12 children per setting being tested in this scenario because we estimated that one person can easily assess up to 16 children in a day on the measures, and because we had a minimum setting size of eight.

Randomisation

This trial used stratified blocks, cluster randomisation at the level of preschool setting to identify the impact of two story session approaches—planned and child-led storybook reading sessions—on the vocabulary and language skills of children. A setting was a natural unit of randomisation in this case as it minimized the risk of spillovers and contamination between the two treatment arms and allowed the setting freedom to arrange their group storybook sessions in line with their normal practice.

There was no pure control group setting as we were 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. Randomisation was stratified by the type of setting (PVI versus state maintained) to ensure participation by both types of setting and to investigate potential differences between them. Within each strata, settings were randomly allocated to either treatment approach one (child-led) or treatment approach two (planned). Block sizes were varied to minimise predictability of allocation. The stratifying variable was included in the primary analysis as a control variable (Kahan and Morris, 2012).

We did not randomise in batches. Randomisation was carried out during the week commencing 7 February 2025, creating a list for the whole sample in each stratum to ensure every possible scenario was covered. All randomisation was undertaken by a statistician blind to the identity of the settings that completed pre-testing. Randomisation lists were

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

generated using R version 4.4.1 (14 June 2024) and using R package `blockrand` (Snow, 2020). An R script for the randomisation can be found in Appendix H.

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.

An intention-to-treat (ITT) approach was used (including all randomised settings and children into 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) are calculated for both the null (that is, that without covariates) and conditional (with covariates) models at post-test.

The primary analysis (RQ1) examined mean follow-up vocabulary levels using the storybook vocabulary test scores, adjusting for the respective baseline measure, setting type, and with the covariate of interest specified as a dichotomous condition variable. The analysis for the primary outcome measure used a linear mixed model (LMM) given that the measure was continuously distributed. Piloting of the measure indicated a small amount of skewness. Some amount of skew in the raw outcome measure was 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 indicated that a LMM was the most appropriate model in this instance. Model assumptions were checked and satisfied at the analysis stage. Nested model comparison was 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 were conducted using R version 4.4.1 (14 June 2024) using the R packages `Tidyverse` (Wickham et al., 2019), `lme4` (Bates et al., 2015), `ordinal` (Cristensen, R., 2023), and `lmerTest` (Kuznetsova, A. et al, 2017). The model included a two-level design (child, level 1 [$i=1, \dots, N$] and setting, level 2 [$j=1, \dots, J$], where N = total number of children and J = total number of settings) as follows:

Two-level design:

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

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

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

where

Y_{ik} are the vocabulary test scores;

$X1_{ik}$ are the baseline vocabulary test scores;

INT_k is the condition variable;

$X2_k$ 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;

β_{jk} , $j = 0, \dots, 1$ are the level 1 parameter estimates; and

γ_{0m} , $m = 0, \dots, 3$ are the level 2 parameter estimates.

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

1. linearity—plotting residuals versus predictor(s): if a structure is present, then transformation or an alternate model specification is required (that is, GLM);
2. homogeneity of variance—variance of the residuals across groups is the same; and
3. residuals are approximately normally distributed—plotting QQ plot.

Secondary analysis

Most secondary analyses (RQ3, RQ4) replicated the primary analysis outlined above, specifically for LanguageScreen (overall standardized score) and ATLA measures (total raw score) used as secondary outcome variables.

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

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

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

where

Y_k are the secondary outcome scores at followup;

$X1_{ik}$ are the baseline secondary outcome scores;

INT_k is the condition variable;

$X2_k$ 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;

β_{jk} , $j = 0, \dots, 1$ are the level 1 parameter estimates; and

γ_{0m} , $m = 0, \dots, 3$ are the level 2 parameter estimates.

Confidence intervals for each intervention effect estimate are 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.

Given that the planned distributional assumptions for the ATLA measure (total raw score) were not satisfied, as appropriate, a generalized linear mixed model with alternate link function was used. A Poisson linear mixed model was specified, as follows:

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

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

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

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

Modelling for this measure reports rate ratios—exponentiating the parameter estimates from the model. The rate ratio is an effect size measure reported when the outcome is a count or if the score is positive and has some positive skew with integer responses. This effect size is interpreted as the comparison of incidence of events happening at different times. For example, when considering the secondary outcomes in this study, we compare the difference in counts pre and post intervention. For each outcome, these models include the covariate of interest specified as a dichotomous variable for intervention group allocation. In addition, covariates for the baseline scores for the outcome and the setting type covariate were also included.

Adjustment for multiple comparisons

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 (Holm, 1979). The Holm method, in a stepwise way, computes the significance levels depending on the P-value based rank of hypotheses. For the i^{th} 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 is tested in order from the smallest to largest P values ($H(1), \dots, H(m)$). The comparison will immediately stop when the first $p(i) \geq \alpha'(i)$ is observed ($i = 1, \dots, m$) and hence all remaining hypotheses of $H(j)$ ($j = i, \dots, m$) are directly declared non-significant without requiring individual comparison.

Analysis in the presence of non-compliance

The primary analyses used an intention-to-treat model. We also carried out analyses focusing on children who received Story Choices as intended. We collected compliance data in terms of a post-trial survey indicating whether schools complied with their allocated approach. We then used this variable in the compliance analysis. Sixty-two of the settings completed this survey, of which sixty-one confirmed that they had delivered the intervention as intended on a single self-report question. We therefore included these sixty-one settings as compliant in our compliance analysis.

We used two-stage, least squares instrumental variables (IV) regression to examine the effect of the intervention in those complying with their allocation. The trial allocation will be the instrumental variable in this analysis.

This analysis used a two-stage, least square approach to estimate the model and Lang-Zieger standard errors reported which are robust to clustering. This approach provides a robust but simplified method to estimate the compliance-adjusted

treatment effect, given the higher likelihood of convergence issues (Gaure, 2013). The R packages `ivreg` and `estimatr` were used to implement the two-stage instrumental variable analysis (Jiang and Small, 2014; Fox Kleiber and Zeileis, 2021). Compliance was instrumented by the intervention allocation (Angrist and Imbens, 1995). The stage 1 model is defined as follows:

$$Compliance_k = \beta_0 + \beta_1 INT_k + \beta_2 baseline_k + \beta_3 X1_k + \beta_4 X2_k + \varepsilon_k$$

Predicted values for $\widehat{Compliance}_k$ from the stage 1 model were included in the stage 2 model, as follows:

$$Y_k = \beta_0 + \beta_1 \widehat{compliance}_k + \beta_2 baseline_k + \beta_3 X1_k + \beta_4 X2_k + r_k$$

Missing data analysis

Exploration of the impact of missing data on trial outcomes was conducted by investigating 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 included all covariates and the outcome appearing in the analysis as per recommendation by White, Royston and Wood (2011). In addition, variables that were predictive of missingness are included on the basis of strength of association with response variables as were any variables that explain response or non-response (Van Buuren, Boshuizen and Knook, 1999).

We have summarised the extent of missing data across all outcomes and their respective control variables. A full multiple imputation strategy was used as more than 5% of data in the primary model was missing and MAR was assumed following Little's MCAR test.

We used multiple imputations by chained equations approach via the `mice` package in R (Van Buuren and Groothuis-Oudshoorn, 2011) and generated 20 imputed datasets. We then estimated the intervention effect for each imputed dataset and pooled the results using Rubin's combination rules for standard errors.

We also report levels of missingness for each trial stage in a CONSORT flow diagram (Figure 2, page 29).

Subgroup analyses

To estimate the effect of the intervention for children from disadvantaged backgrounds, the main analysis described above was repeated on a subsample of children eligible for the EYPP. The model was run for the primary outcomes only. We also calculated effect sizes and 95% confidence intervals, as outlined in the primary analysis. To determine which approach was more effective for disadvantaged children, a moderation analysis adjusted the primary analysis with the inclusion of the moderator as a main effect and interaction between moderator and randomised group indicator. For example, the EYPP moderator analysis is as follows:

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

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

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

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

where

Y_{ik} are the vocabulary test scores;

$X1_{ik}$ are the baseline vocabulary test scores;

INT_k is the condition variable;

$X2_k$ are indicators of setting type (strata variable);

μ_{0k} is the random intercept term for setting; and

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} .

Estimation of effect sizes

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

$$\hat{\Delta}_d = \frac{\hat{\beta}_1}{sd_{pooled}}$$

where $\hat{\beta}_1$ 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 . We then take the Cohen's d and apply the correction factor, J , as follows:

$$J = 1 - \frac{3}{4df - 1}$$

$$\hat{\Delta}_g = \left(\frac{\hat{\beta}_1}{sd_{pooled}} \right) \cdot J$$

All parameter estimates from the models will be reported with 95% confidence intervals. We estimate the 95% confidence interval approximately by plugging in the upper and lower CI of the model estimates in to the effect size CIs.

Calculation of the incident rate ratios are directly calculated by exponentiating the parameter estimates from the count models for the ATLA measures (total raw score) directly. Modelling for this measure reports rate ratios (exponentiating the parameter estimates from the model). The rate ratio is an effect size measure reported when the outcome is a count or if the score is positive and has some positive skew with integer responses. This effect size is interpreted as the comparison of incidence of events happening at different times. For example, when considering the secondary outcomes in this study, we compare the difference in counts pre and post intervention. For each outcome, these models include the covariate of

interest specified as a dichotomous variable for intervention group allocation. In addition, covariates for the baseline scores for the outcome and the setting type covariate were also included.

Estimation of ICC

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

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

Where σ^2 is the residual variance and σ^2_{school} is the random intercept variance according to setting (setting level clustering).

We also report the conditional ICC from the primary analysis model:

$$ICC_{conditional} = \frac{\sigma^2_{school(adj)}}{\sigma^2_{(adj)} + \sigma^2_{school(adj)}}$$

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

Additional analyses and robustness checks

In addition to the analyses described above, other moderator and mediator variables were recorded. Moderators are variables that help to determine ‘under what conditions’ the intervention is effective, while mediators help to establish ‘how and why’ the intervention is effective (Kraemer, Kiernan, Essex and Kupfer, 2012). Two moderators were planned at the design phase: educator questioning skills and qualifications (measured in the pre-trial educator survey) and number of different books used across the sessions (measured by session records); these correspond to RQ11. In addition, we conducted further analysis for robustness which excludes schools/children whose outcomes were collected by individuals who were not blind to assigned approach (that is, the core research team). This replicated the primary analysis and was unplanned in the first instance.

To estimate the moderated effect of the intervention for children allowing for the three moderators, a moderation analysis adjusted the primary analysis with the inclusion of the moderator as a main effect and interaction between moderator and randomised group indicator. For example, the ‘number of books in setting’ moderator analysis (Nbooks) is as follows:

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

$$L2: \beta_{0k} = \gamma_{00} + \gamma_{01}X2_k + \gamma_{02}INT_k + \gamma_{02}MOD_k + \gamma_{03}MOD_k * INT_k + \mu_{0k}, \mu_{0k} \sim N(0, \tau^2)$$

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

where

Y_{ik} are the vocabulary test scores;

$X1_{ik}$ are the baseline vocabulary test scores;

INT_k is the condition;

MOD_k is the moderator variable at level 2;

X_{2k} are indicators of setting type (strata variable)

μ_{0k} is the random intercept term for setting; and

τ_{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} .

Implementation and process evaluation

Research questions

RQ7: How, and how well, are the two conditions implemented in different settings?

- Did observations indicate that the settings adhered to the guidance given for each approach? [fidelity]
- Did the observations indicate that the sessions were differentiated for settings in the planned and child-led conditions?
- What changes to the approaches did individual educators and settings make?
- Do these factors vary across the two approaches?

RQ8: What role does dosage (number of sessions attended) play in vocabulary growth in each of the two conditions?

RQ9: What is the perceived value of each of the two approaches from the educators using it?

- Do the children indicate engagement and enjoyment during the sessions?
- Do the educators believe that the approach has led to improved vocabulary knowledge, language skills, or engagement with literacy?
- What are the perceived costs associated with each approach?
- Do the educators intend to continue with their allocated approach after the end of the trial?

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

- Did settings where the sessions were led by a qualified teacher show greater progress than those that were not? This is based on whether any of the educators who completed the educator survey for that setting had a teaching qualification.
- Did the impact of teaching qualifications vary across the two approaches?
- Did settings that had a large number of books show better progress than other settings?
- Did the impact of number of books vary across the two approaches?

Research methods

The IPE used a mixed method approach, described below and summarised in Table 9. Data collection included:

- session records, which provided the number of sessions carried out and the range of books used;
- online pre-trial surveys gathering information on characteristics of the settings and the educators taking part;
- interviews and observations at eight settings; and
- an online post-trial survey.

Session records

Educators were 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. This allows us to assess the total number of sessions delivered, the number of sessions each child attended, and gives broad indicators of the fidelity to approach of the sessions.

Pre-trial setting-level survey

The setting-level survey provided 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 had sessions, whether they were carried out with whole classes or key worker groups, and whether they were integrated with other activities such as story sacks. It is likely that these factors were not 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.

Pre-trial educator survey

All educators leading storybook sessions were 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 on a longer measure used in a recent paper on assessing the quality of dialogic reading (Towson et al., 2024). We aimed to analyse these questions to assess whether they align to questions suggested by the CROWD framework (for example, completion prompts, recall prompts, open ended questions, and questions or distancing prompts) or whether they are, for example, focused only on surface aspects of the picture (for example, 'Where is the dog?'). A scoring scheme was created where four points were given if the response included at least three different types of CROWD questions, three points were given for two different types of CROWD questions, two for at least one CROWD question, and one for providing questions that were not aligned to the CROWD questions. Educators gave a wide range of questions, most of which fitted the CROWD framework. Question classification was carried out by a research assistant blind to the identities of the educators and which approach they were assigned to.

Mean quality scores were slightly higher in the educators in the planned condition, but the difference was not considered significant ($t(58) = -1.75, p = .09$).

Table 9: Details on moderating variables—educator questioning skill and qualified teacher involvement

Approach	N	Questioning skill (SD)	Teacher involved in delivery (% of settings)
Child-led	30	2.57 (0.79)	68.8%
Planned	30	2.92 (0.73)	31.3%

Post-trial settings survey

All settings that took part were 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 were, and whether they intend to continue with the approach. This survey is included in Appendix J.

Case studies

In the last two weeks of the trial, the project team (TA and JC) carried out case study observations of eight settings, working with two PVI settings and two mainstream settings in each of the two approaches. The case study observations consisted of three elements:

- *environmental observation*: gathering information about the classroom environment—for example, size of the space, opportunities for spoken communication, reading, and writing, and the extent to which children engaged with those opportunities;
- *session observation*: watching a storybook reading session being delivered; an observation checklist (example included in Appendix K) was 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; and
- *educator interviews*: we then interviewed the educators at the settings in order to gain a full understanding of their views of the approach; we began 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 was driven by the research questions—understanding how the two intervention approaches were implemented, potential barriers to implementation for each, and the perceived feasibility and value of each intervention approach. As there was no untreated or ‘business as usual’ control condition, analysis needed to focus on each of the two approaches equally and assess whether there were similarities and differences in the feasibility, acceptability, and perceived value of each.

Qualitative data—observations, interviews, and open text questions within the surveys

First, we divided comments according to the condition the setting was allocated to. We then coded the data deductively, aligning comments to the IPE dimensions (implementation, value, and contextual factors). For survey responses, open text comments were thematically examined followed by content-analysis using keyword searches using the NLTK package in Python. For example, when considering implementation, we selected the keywords associated with central elements of implementation of each of the two approaches and searched for mention of those words. For the interviews and observations, we used a deductive thematic approach.

Quantitative data—surveys and session records

Quantitative data from the pre-trial setting survey, pre-trial educator survey, post-trial setting survey, and session records were analysed using descriptive statistics, including percentage responses or mean responses as appropriate. They were compared across the two conditions. RQs which included mediation or association were assessed statistically using correlations and *t*-tests. Table 10 gives an overview of the dimensions addressed by the IPE.

Table 10: IPE methods overview

IPE dimension	RQ addressed	Session records	Surveys	Interviews	Observations	Sample size and sampling criteria	Data analysis methods
Implementation	RQ7	X		X		All settings (80)	Thematic analysis and descriptive statistics
Dosage	RQ8	X	X	X	X		
Feasibility and acceptability	RQ9		X	X			
Cost	RQ10		X	X			
Context/moderators	RQ11		X	X	X		Multiple regression

Timeline

Table 11: 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
Dec 2024	First draft of study plan complete	Julia Carroll
Sep-Oct 2024	Phase 1 piloting of vocabulary measure	Tanvir Ahmed
Oct-Dec 2024	Phase 2 piloting of vocabulary measure	Julia Carroll
31 Oct 2024	Complete setting recruitment	Tanvir Ahmed
Jan-Feb 2025	Pre-testing	Tanvir Ahmed
Feb-May 2025	Trial delivery	Julia Carroll
1 Jun 2025	Study plan complete and trial registered on OSF	Julia Carroll
May-Jun 2025	Process evaluation data collection	Julia Carroll
Jun-Jul 2025	Post-testing	Tanvir Ahmed
Aug-Oct 2025	Data analysis and evaluation	Julia Carroll
Dec 2025	First draft report delivered	Julia Carroll

Impact evaluation results

Participant flow including losses and exclusions

Eighty-four settings agreed to take part and completed MOUs (Memoranda of Understanding): two were excluded due to not having enough eligible children to take part, 969 children were pre-tested in 82 different settings, and 41 settings (children $n = 502$) were allocated to the child-led condition and 41 settings ($n = 476$) were allocated to the planned condition. Four settings withdrew after randomisation, three from the child-led condition ($n = 29$) and one from the planned condition ($n = 14$). This left 78 settings, 38 in the child-led condition and 40 in the planned condition.

At post-test it was found that some of the children ($n = 70$: child led, 42; planned, 28) did not meet the specified age criteria and should be excluded from analysis. Four settings (two from each condition) did not provide dates of birth for their children and therefore these children were withdrawn from analysis ($n = 43$). A sensitivity analysis was added to assess whether withdrawing these children and settings altered group differences on the primary outcome. A further 139 children did not complete the primary outcome measure. Most of these children were absent on the day of testing or had left the setting. This left a post-test sample of $n = 358$ for the responsive condition and $n = 329$ for the planned condition.



Table 12: Minimum detectable effect size at different stages

		Design	Randomisation		Analysis	
		Overall	Overall	EYPP*	Overall	EYPP
MDES		0.30	0.316	0.46	0.3	0.375
Pre-test/post-test correlations	Level 1 (pupil)	0.42	0.42	0.42	0.60	0.64
	Level 2 (class)	n/a	n/a	n/a	n/a	n/a
	Level 3 (setting)	0	0	0	0	0
Intraclass correlations (ICCs)	Level 2 (class)	n/a	n/a	n/a	n/a	n/a
	Level 3 (setting)	0.20	0.20	0.20	0.158	0.173
Alpha		0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		10 (12 to allow for 20% attrition)	12	2	12	3
Number of settings	Responsive	45	41	41	36	36
	Planned	45	41	41	38	38
	Total:	90 (100 to allow for 10% attrition)	82	82	74	74
Number of children	Responsive	400	502	62	358	42
	Planned	400	476	63	329	55
	Total:	800	978	125	687	97

* Number estimated based on design stage assumptions as EYPP was not provided at pre-test.

Table 12 shows the estimated MDES figures at the design, randomisation, and analysis stages. For the latter, the MDES estimate uses the achieved sample for the primary outcome (n = 687 children, n = 74 settings) and subgroups. The ICC for the class level and the pre- and post-test correlations have been updated. The achieved sample was powered to detect effect sizes of 0.247 for all children, 0.335 for FSM-eligible children.

Attrition

Between randomisation and post-test, attrition at the setting level was 4.8% (four settings). Overall attrition at child level was 30%, with similar levels in the child-led and planned conditions: 29% and 31%, respectively. As described above, 70 children were removed from analysis because they did not meet age-based inclusion criteria and 139 were absent at post-test.

Table 13: Child-level attrition from the trial (primary outcome)

		Choice condition 1 (child-led)	Choice condition 2 (planned)	Total
Number of children	Randomised	502	476	978
	Analysed	358	329	687
Pupil attrition (from randomisation to analysis)	Number	144	147	291
	Percentage	28.7%	30.9%	29.8%

Child and setting characteristics

This section describes the characteristics of participating settings and their existing storybook reading practices at baseline, prior to intervention implementation. The findings are based on the setting pre-trial survey completed by 67 settings (84.8% of the sample) and the educator pre-trial survey returned by educators in 64 settings (81.0% of the sample). For the former, we received pre-trial survey responses from 67 of the 79 settings. Within this sample, we received responses from 22 of the 23 PVIs and 45 of the 56 maintained settings and from 36 settings completing the planned approach and 31 completing the child-led approach. Response rates are higher in the PVI settings than in the maintained settings and similar across the two allocated approaches. Given the high participation rates, we made the assumption that these responses represented the sample well.

The trial included 79 settings, of which 56 were maintained and 23 were private voluntary and independently (PVI) operated provisions. Settings varied considerably in size. The maximum number of children attending each session ranged from eight (the lowest threshold to participate in the trial) to 132. It suggests that the settings responding to the survey encompassed both small, targeted provisions and large institutional settings catering children of different age groups.

All settings catered for three- and four-year-old children, as was expected given the inclusion criteria. A smaller number also catered for younger children ($n = 26$). Overall, the settings are similar across the two conditions, as would be expected given the random assignment. There are no significant differences in IMD decile, average number of children, or average number of educators. However, in addition to the range of child attendees, the number of educators varied from two to 39.

Characteristics of the settings across the two conditions are shown in Table 14.

Table 14: Baseline characteristics of groups at analysis

Setting-level (categorical)	National-level mean	Child-led		Planned		Effect size
		n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Type of setting (number of PVIs)	66% of children in PVI settings	105/387	27.1%	79/355	22.3%	
Setting-level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	
Index of Multiple Deprivation (IMD)	5.5	314 (73)	4.50 (2.52)	314 (41)	4.26 (2.82)	-0.09
Number of children in setting			36.97 (24.67)		35.26 (20.07)	-0.08

Number of staff in setting			7.84 (7.59)		8.09 (7.59)	-0.03
Child-level (categorical)		n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Gender—male		206/387 (0)	53.2	177/355 (1)	49.86	
Gender—female		181/387 (0)	46.8	177/355 (1)	49.86	
Eligibility for Pupil Premium	13%	42/387 (44)	10.9	55/355 (58)	15.49	
Special Educational Needs		33/387 (25)	8.5	22/355 (22)	6.2	
EAL status		86/387 (6)	22.2	77/355 (14)	21.7	
Child-level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	
Storybook vocabulary measure	Not available	387 (0)	27.17 (8.16)	355(0)	27.12 (7.86)	-0.01

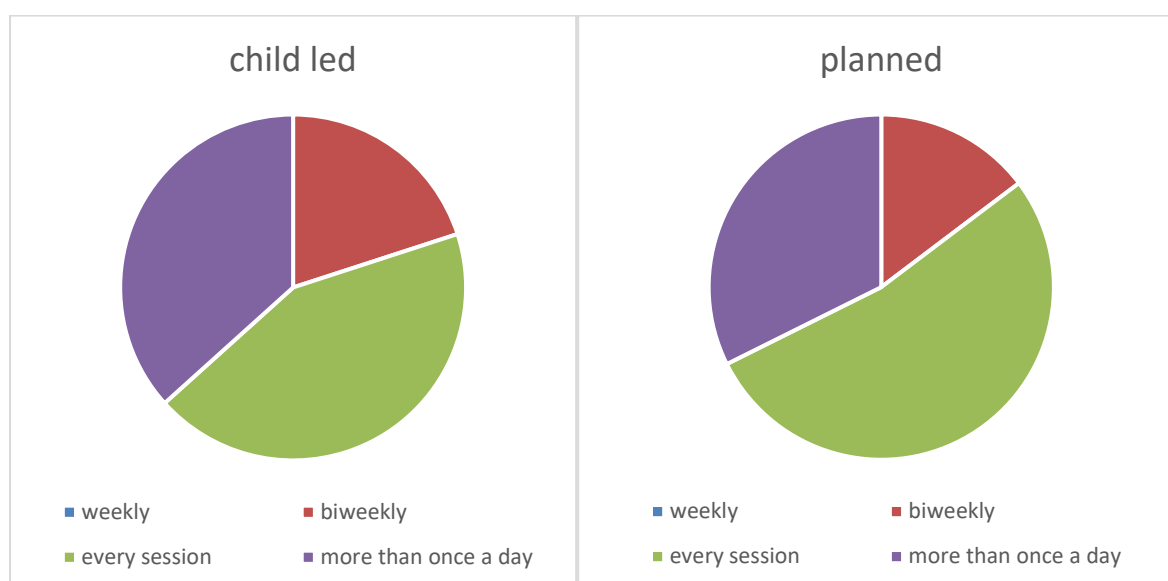
The mean pre-test scores are shown at the bottom of Table 14. The very small effect size indicates no difference between the two conditions at pre-test.

Usual practice

Existing storybook reading practices

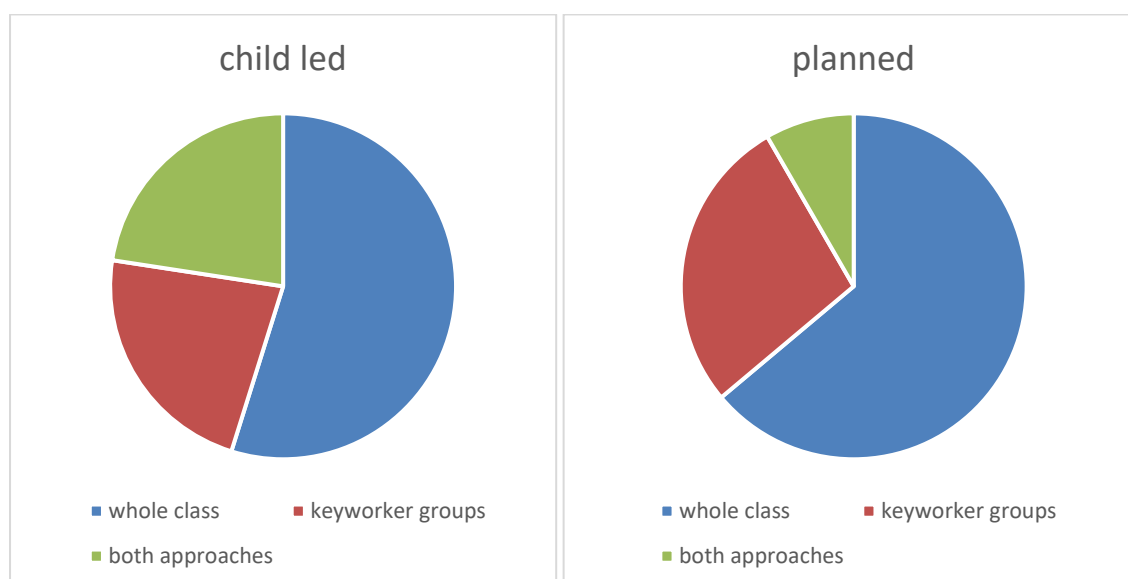
Settings reported in the pre-trial survey that storybook reading was already an established practice, with considerable variation in frequency. The most common pattern was reading once during each session (44%) followed by more than once per session (35%). A smaller proportion organised storybook reading sessions at least twice a week (16%) and no settings used storybook reading less than twice a week. Frequencies were similar across both approaches, as shown in Figure 3

Figure 3: Frequency of storybook reading, by approach



With regard to organising the storybook sessions, the majority of settings (59%) conducted these with the whole class together. A quarter (25%) used multiple keyworker groups for storybook reading, while 16% made other arrangements. Frequencies were similar across both approaches, as shown in Figure 4

Figure 4: Groupings for storybook reading



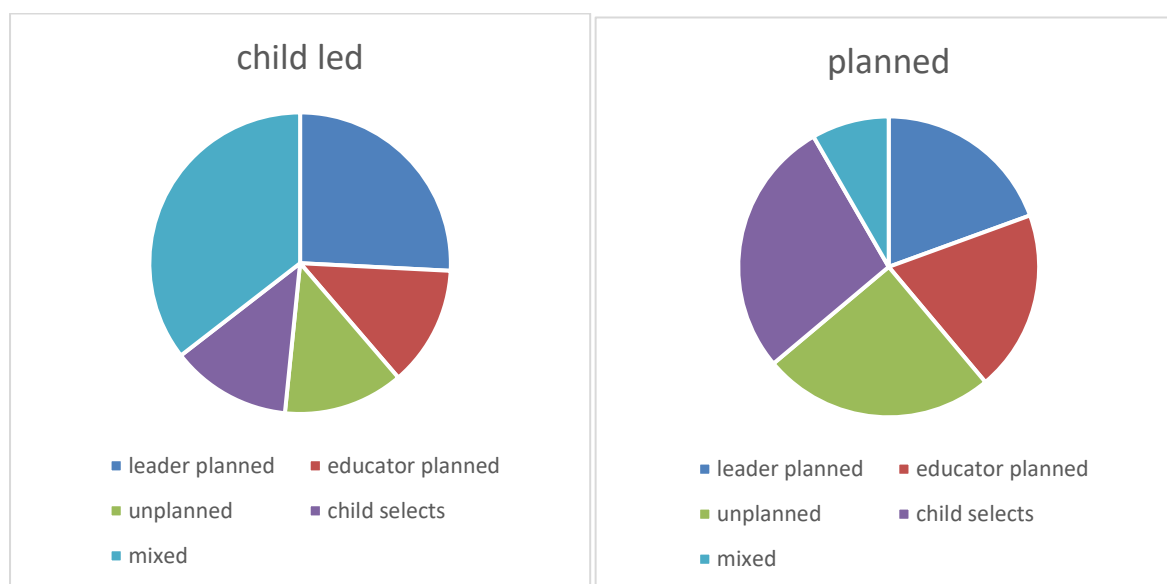
Settings commonly integrated other activities with storybook reading. The most prevalent practices were linking books to a topic or theme (reported by 62 settings, representing 21% of total responses to this multiple-response question), using songs linked to stories (19%), employing puppets or dolls for story enactment (19%), and using toys or pictures linked to stories (19%). Costumes for acting out stories and videos linked to stories were less common (11% each).

Book selection processes

Diverse approaches were used by settings in selecting books for storybook reading sessions. The most common approach was centralized planning, where the EYFS lead decided which books would be read and when (22%). In 18% of the settings, individual educators planned ahead and in 21% educators selected books for reading on the day based on their own judgement and a variety of other methods. Child-directed book selection was less common, with 12% of settings allowing

children to choose from a wide range of books and 7% offering children a few pre-selected options. Frequencies were similar across settings allocated to the two approaches, as shown in Figure 5.

Figure 5: Approach to planning used prior to the trial—by allocated approach.



Outcomes and analysis

Primary analysis

This analysis addresses RQ1:

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

Table 15 reports the model parameter estimates from the primary outcome analysis using a linear mixed model. Given the ANCOVA design of the model, the intercept represents the expected value (or mean) of storybook vocabulary test scores at post-test when storybook vocabulary test scores at baseline and setting type covariates are equal to zero and PVI respectively. The adjusted mean difference in storybook expressive vocabulary scores between the planned and child led groups is -1.23 (95% CI: -2.72, 0.26; $t(73.85618) = -1.620$; $p = .106$). This translates to a Hedges g effect size of -0.19 (95% CI: -0.43, 0.05; see Table 15). The negative direction of the effect indicates that children in the planned group scored lower in storybook expressive vocabulary, on average, than children in the child-led group. The possible difference in impact between the two groups ranged from a 2.72 point advantage in the child-led group to 0.26 point advantage in the planned group. Given that our estimated confidence intervals contain zero there is no conclusive evidence that child-led and planned approaches produced a comparatively different impact on storybook expressive vocabulary.

The unconditional ICC (95% CI) associated with setting for the primary outcome, storybook vocabulary measure is 0.158 (0.093, 0.220) and the conditional ICC (95% CI) is 0.158 (0.087, 0.218).

Figure 6 shows the histogram of the storybook vocabulary measure at post-randomisation follow-up. The distribution appears marginally skewed, however, model assumption checks on the residuals from the primary analysis appear to show no issues.

Figure 6: Histogram of the storybook vocabulary measure at post-test, split by trial arm

Red: child-led condition; blue: planned condition.

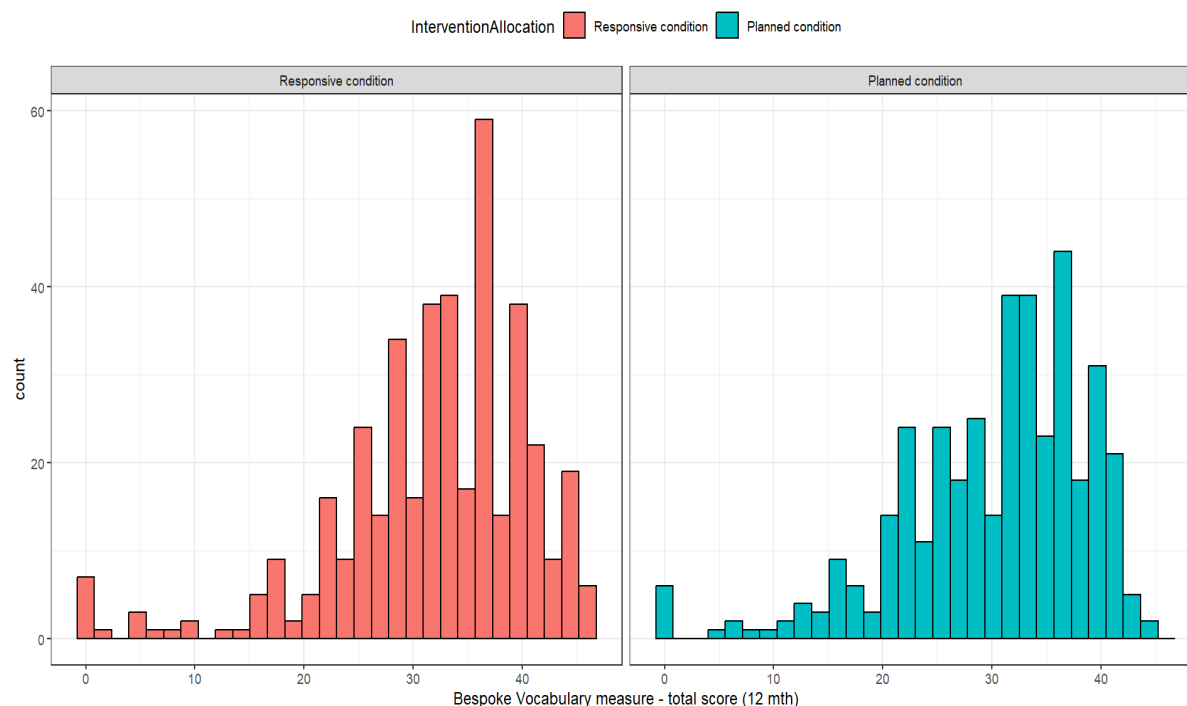


Table 15: Primary analysis for storybook vocabulary test scores at post-randomisation follow-up using a two-level linear mixed model

Coefficient	Primary analysis (two-level)	
	Estimates	CI
Intercept	15.44 ***	13.10, 17.77
Outcome baseline	0.60 ***	0.54, 0.66
Setting type—maintained	1.15	-0.54, 2.82
Intervention—planned	-1.23	-2.72, 0.26
Random effects		
σ^2	34.60	
τ_{00}	6.47 Setting_ID	
ICC	0.16	
N	74 Setting_ID	
Observations	691	
Marginal R ² / conditional R ²	0.364 / 0.464	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 16: Primary analysis summary

	Adjusted means				Effect size		
	Choice condition 1 (child-led)		Choice condition 2 (planned)				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (condition 1 and 2)	Hedges <i>g</i> (95% CI)	<i>p</i> -value
Post-test storybook vocabulary test score	412 (14%)	32.27 (31.14, 33.42)	390 (19%)	31.05 (29.90, 32.20)	802 (412, 390)	-0.19 (-0.43, 0.05)	0.110

Secondary analysis

This section addresses RQ3 and RQ4.

RQ3: What is the difference in attitudes towards literacy between the two conditions?

RQ4: What is the difference between the two conditions in oral language skills measured by the LanguageScreen measure?

Secondary outcome analyses have been organized into separate subsections for clarity. Each secondary outcome analysis reports the study plan specified analysis.

Table 17: Secondary analysis summary

	Adjusted means				Effect size			
	Choice condition 1 (child-led)		Choice condition 2 (planned)					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (condition 1 & 2)	Hedges <i>g</i> (95% CI)	<i>p</i> - value	<i>p</i> -value (Holm adjusted)
Post-test Language Screen—total score	341 (29%)	101.2 (99.1, 103.4)	330 (31%)	97.7 (95.6, 99.9)	671 (341, 330)	-0.30 (-0.55, -0.06)	0.014	0.084
Post-test Language Screen— expressive vocabulary	341 (29%)	101.3 (99.2, 103)	330 (31%)	98.4 (96.3,100)	671 (341, 330)	-0.26 (-0.50, - 0.01)	0.04	0.160
Post-test Language Screen— listening comprehension	341 (29%)	100.9 (98.8, 103)	330 (31%)	97.5 (95.4, 99.6)	671 (341, 330)	-0.28 (-0.49, - 0.06)	0.015	0.084
Post-test Language Screen— receptive vocabulary	341 (29%)	97.5 (95.3, 99.7)	341 (29%)	94.6 (92.3, 96.9)	671 (341, 330)	-0.22 (-0.44, 0)	0.053	0.160
Post-test Language Screen—	341 (29%)	104 (102, 106)	341 (29%)	102 (100, 104)	671 (341, 330)	-0.12 (-0.32, 0.09)	0.278	0.556

sentence completion								
Post-test ATLA measure	350 (27.5%)	5.28 (4.55, 6.01)	380 (20.1%)	5.03 (4.31, 5.74)	730 (350, 380)	0.99 (0.84, 1.17)	0.930	0.930

Figure 7 shows the histograms of all secondary outcomes at post-test. The ATLA raw score has been transformed to correct the left or negative skew to allow fitting of a suitable model to the data.

Figure 7: Histograms of all secondary outcomes at post-test, split by condition

Red: child-led condition; blue: planned condition.



Detailed results for each of the secondary analysis measures are presented below.

ATLA measure

The ATLA measure was considerably left skewed and showed a ceiling effect. We transformed the data by reverse coding and thus the ceiling effect became a floor effect with an inflated number of zero scores. To model this data correctly, we fitted a GLMM (Poisson) model. The Poisson model estimates are presented as incidence rate ratios, which are the appropriate standard effect size measure for count models (as per study plan specification). An incidence rate ratio of 1.0 indicates equal rates in both groups, a rate ratio greater than 1.0 indicates an increased rate in the intervention arm.

In all models reported in Table 18 we do not find evidence of a difference between the child-led and planned conditions (IRR: 0.99; 95% CI: 0.84, 1.17; $p = .930$).

The unconditional ICC (95% CI) associated with setting for the secondary outcome, ATLA measure (Poisson), is 0.438 (0.327, 0.525) and the conditional ICC (95% CI) is 0.397 (0.285, 0.481).

Table 18: Secondary outcome analysis for ATLA measure at post-test using a two-level linear mixed model (Poisson family)

	Secondary outcome analysis (two-level Poisson)	
Coefficient	Incidence Rate Ratios	CI
Intercept	3.92***	3.23, 4.76
Outcome baseline	1.04***	1.04, 1.05
Setting type—maintained	1.03	0.85, 1.25
Intervention—planned	0.96	0.80, 1.14
Random effects		
σ^2	0	
τ_{00}	0.12 Setting_ID	
ICC	0.4	
N	70 Setting_ID	
Observations	616	
Marginal R ² / conditional R ²	0.261 / 1.000	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

LanguageScreen—Standardized Total Score

Table 19 reports the model parameter estimates from the LanguageScreen – Standardized Total Score analysis using a linear mixed model. Given the ANCOVA design of the model, the intercept represents the expected value (or mean) of standardized LanguageScreen – Total Score at post-tests when adjusting for baseline measure of the storybook vocabulary outcome and setting type covariates are equal to zero and PVI respectively. The adjusted mean difference in endline LanguageScreen standardised total scores between the planned group and child-led groups is -3.53 (95% CI: $-6.29, -0.77$; $t(63.62329) = -2.515$; $p = .01442$; Holm adjusted $p = .084$). This translates to a Hedges g effect size of -0.30 (95% CI: $-0.55, -0.06$; Table 19). The negative direction of the effect indicates that children in the planned group scored lower in wider oral language, on average, compared to those in the child-led group. The possible difference in impact between the two groups ranged from a 0.77 point to a 6.29 point advantage for the child-led group.

The unconditional ICC (95% CI) associated with setting for the secondary outcome LanguageScreen – Standardized Total Score is 0.206 (0.099, 0.274) and the conditional ICC (95% CI) is 0.134 (0.059, 0.197).

Table 19: Secondary outcome analysis for standardized LanguageScreen – Total Score at post-test using a two-level linear mixed model

	Secondary outcome analysis (two-level)	
Coefficient	Estimates	CI
Intercept	81.97 ***	77.45, 86.49
Outcome baseline	0.80 ***	0.67, 0.92
Setting type—maintained	-5.07 **	-8.26, -1.88
Intervention—planned	-3.53 *	-6.29, -0.77
Random effects		
σ^2	117.73	
τ_{00}	18.17 Setting_ID	
ICC	0.13	
N	70 Setting_ID	
Observations	586	
Marginal R ² / conditional R ²	0.263 / 0.362	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

LanguageScreen—Standardized Expressive Vocabulary score

Table 20 reports the model parameter estimates from the LanguageScreen – Standardized Expressive Vocabulary score analysis using a linear mixed model. Given the ANCOVA design of the model, the intercept represents the expected value (or mean) of the score at post-test when adjusting for baseline measure of the storybook vocabulary outcome and setting type covariates are equal to zero and PVI respectively. The adjusted mean difference in endline LanguageScreen standardised expressive vocabulary scores between the planned group and child-led groups is -2.92 (95% CI: -5.61, -0.23; $t(65.30947) = -2.132$; $p = .0367$; Holm adjusted $p = .160$). This translates to a Hedges g effect size of -0.26, (95% CI: -0.50, 0.01; Table 20). The negative direction of the effect indicates that children in the planned group scored lower in expressive vocabulary, on average, compared to children in the child-led group. The possible difference between the groups ranges from a 0.23 point to a 5.61 point advantage for the child-led group. Given that our estimated confidence intervals cross zero there is no conclusive evidence that the child-led and planned approaches produced a comparatively different impact on expressive vocabulary.

The unconditional ICC (95% CI) associated with setting for the secondary outcome LanguageScreen – Standardized Expressive Vocabulary score is 0.209 (0.120, 0.282) and the conditional ICC (95% CI) is 0.132 (0.062, 0.213).

Table 20: Secondary outcome analysis for standardized LanguageScreen – Expressive Vocabulary Score at post-randomisation follow-up using a two-level linear mixed model

	Secondary outcome analysis (two-level)	
Coefficient	Estimates	CI
Intercept	85.41 ***	81.00, 89.83
Outcome baseline	0.68 ***	0.56, 0.80
Setting type—maintained	-5.70 ***	-8.81, -2.59
Intervention—planned	-2.92 *	-5.61, -0.23
Random effects		
σ^2	112.72	
τ_{00}	17.21 Setting_ID	
ICC	0.13	
N	70 Setting_ID	
Observations	586	
Marginal R ² / conditional R ²	0.227 / 0.329	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

LanguageScreen—Standardized Listening Comprehension score

Table 21 reports the model parameter estimates from the LanguageScreen – Standardized Listening Comprehension score analysis using a linear mixed model. Given the ANCOVA design of the model, the intercept represents the expected value (or mean) of the score at post-test when adjusting for baseline measure of the storybook vocabulary outcome and setting type covariates are equal to zero and PVI respectively. The adjusted mean difference in endline LanguageScreen standardised listening comprehension scores between the planned group and child-led groups is 3.40 (95% CI: -6.05, -0.74; $t(64.39562) = -2.513$; $p = .0145$; Holm adjusted $p = .084$). This translates to a Hedges g effect size of -0.28, (95% CI: -0.49, -0.06; Table 21). The negative direction of the effect indicates that children in the planned approach scored lower in expressive vocabulary, on average, compared to children in the child-led approach. The possible difference ranged from a 0.74 point to a 6.05 point advantage for the child-led group.

The unconditional ICC (95% CI) associated with setting for the secondary outcome LanguageScreen – Standardized Listening score is 0.123 (0.055, 0.192), and the conditional ICC (95% CI) is 0.088 (0.035, 0.149).

Table 21: Secondary outcome analysis for standardized LanguageScreen – Listening Comprehension Score at post-randomisation follow-up using a two-level linear mixed model

	Secondary outcome analysis (two-level)	
Coefficient	Estimates	CI
Intercept	86.74 ***	82.13, 91.36
Outcome baseline	0.55 ***	0.42, 0.68
Setting type—maintained	-1.62	-4.71, -1.47
Intervention—planned	-3.40 *	-6.05, -0.74
Random effects		
σ^2	138.42	
τ_{00}	13.41 Setting_ID	
ICC	0.09	
N	70 Setting_ID	
Observations	586	
Marginal R ² / conditional R ²	0.130 / 0.207	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

LanguageScreen—Standardized Receptive Vocabulary score

Table 22 reports the model parameter estimates from the LanguageScreen- Standardized Receptive Vocabulary score analysis using a linear mixed model. Given the ANCOVA design of the model, the intercept represents the expected value (or mean) of the score at post-test when adjusting for baseline measure of the storybook vocabulary outcome and setting type covariates are equal to zero and PVI respectively. The adjusted mean difference in endline LanguageScreen standardised receptive vocabulary scores between the planned and child-led groups, controlling for baseline storybook expressive vocabulary test score and setting type, is -2.91 (95% CI: -5.80, -0.02; $t(60.94696) = -1.976$; $p = .0527$; Holm adjusted $p = .160$). This translates to a Hedges g effect size of -0.22 (95% CI: -0.44, 0; Table 22). The negative direction of the effect indicates that children in the child-led group showed greater progress than children in the planned group. The possible difference in impact between the two groups ranged from an advantage of 0.02 point to 5.80 points for the child-led approach.

The unconditional ICC (95% CI) associated with setting for the secondary outcome LanguageScreen – Standardized Receptive Score is 0.144 (0.071, 0.238) and the conditional ICC (95% CI) is 0.096 (0.033, 0.169).

Table 22: Secondary outcome analysis for standardized LanguageScreen – Receptive Vocabulary Score at post-test using a two-level linear mixed model

	Secondary outcome analysis (two-level)	
Coefficient	Estimates	CI
Intercept	83.20 ***	78.23, 88.17
Outcome baseline	0.60 ***	0.46, 0.74
Setting type—maintained	-4.42 *	-7.77, -1.06
Intervention—planned	-2.91	-5.80, -0.02
Random effects		
σ^2	157.24	
τ_{00}	16.66 Setting_ID	
ICC	0.10	
N	70 Setting_ID	
Observations	586	
Marginal R ² / conditional R ²	0.144 / 0.226	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

LanguageScreen—Standardized Sentence Completion score

Table 23 reports the model parameter estimates from the LanguageScreen- Standardized Sentence Completion score analysis using a linear mixed model. Given the ANCOVA design of the model, the intercept represents the expected value (or mean) of the score at post-test when adjusting for baseline measure of the storybook vocabulary outcome and setting type covariates are equal to zero and PVI respectively. The adjusted mean difference in endline LanguageScreen standardised sentence completion scores between the planned and child-led approaches, controlling for baseline storybook expressive vocabulary test score and setting type, -1.45 (95% CI: -4.05, 1.15; $t(50.8568) = -1.097$; $p = .2777$; Holm adjusted $p = .556$). This translates to a Hedges g effect size of = -0.12 (95% CI: -0.32, 0.09; Table 23). The negative direction of the effect indicates that children in the planned group scored lower on sentence completion, on average, compared to children in the child-led group. Given that our estimated confidence intervals contain zero, there is no conclusive evidence that the child-led and planned approaches produced a comparatively different impact on sentence completion. The possible difference between the two approaches ranged from a 4.05 point advantage for the child-led approach to a 1.15 point advantage for the planned approach.

The unconditional ICC (95% CI) associated with setting for the secondary outcome LanguageScreen – Standardized Sentence Completion Score is 0.134 (0.064, 0.196), and the conditional ICC (95% CI) is 0.072 (0.012, 0.134).

Table 23: Secondary outcome analysis for standardized LanguageScreen – Sentence Completion Score at post-randomisation follow-up using a two-level linear mixed model

	Secondary outcome analysis (two-level)	
Coefficient	Estimates	CI
Intercept	85.42 ***	80.79, 90.05
Outcome baseline	0.73 ***	0.60, 0.87

Setting type—maintained]	-3.44 *	-6.47, -0.41
Intervention—planned]	-1.45	-4.05, 1.15
Random effects		
σ^2	145.98	
τ_{00}	11.32 Setting_ID	
ICC	0.07	
N	70 Setting_ID	
Observations	586	
Marginal R ² / conditional R ²	0.190 / 0.249	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Summary of the main findings

While both choice approaches are effective in improving storybook expressive language, there is a difference in storybook expressive vocabulary scores between the child-led and planned groups at the end of the trial. Children in the child-led group achieved a higher score in storybook expressive vocabulary, on average, compared to children in the planned group—an effect size of -0.19 (95 % CI: -0.43 to 0.05). Given the confidence intervals of this effect cross zero there is no conclusive evidence that either approach produced a comparatively different impact on storybook expressive vocabulary. The possible difference in impact between the two groups ranged from a 2.72 point advantage in the child-led group to a 0.26 point advantage in the planned group.

There is a similar difference between the two groups on the LanguageScreen measure, again showing that the children in the child-led group achieved a higher score in wider oral language skills, on average, compared to children in the planned group. This is consistent across the different LanguageScreen subtest measures: Overall Score ($g = -0.30$), Expressive Vocabulary ($g = -0.26$), Listening Comprehension ($g = -0.28$), Receptive Vocabulary, ($g = -0.22$) and Sentence Completion ($g = -0.12$).

The consistent pattern of higher baseline storybook expressive vocabulary and wider oral language scores in favour of the child-led group suggests that there are small but consistent advantages for the child-led group in language skills across different measures. This is the opposite pattern to the one hypothesised, where planned sessions were predicted to lead to greater language improvement. Possible reasons for this will be considered in the rest of the report.

There was no sign of a difference between conditions in the ATLA measure (IRR = 0.99, $p = .93$).

Analysis in the presence of non-compliance

The impact of non-compliance relates to RQ6.

RQ6: Does compliance to the choice approach mediate the vocabulary gains shown by children in the two conditions?

An instrumental variables approach was used to evaluate whether compliance to allocated intervention altered the results of the primary analysis of Storybook Vocabulary Measure—total score at post-test. The intervention allocation—the trial arm—was used as the instrumental variable and adjusted estimates of the intervention effect were estimated under the compliance scenario (Table 24) and compared to our primary intention-to-treat analysis.

Comparison of estimates shows that the magnitudes of the estimates remain relatively unchanged and there is no change in the pattern of results for both the difference between interventions and setting type, under the CACE model:

- trial arm, $\beta = -1.18$ —95% CI: -3.02, 0.65; $t(46.99) = -1.291$; $p = .2029$; Hedges $g = -0.11$; 95% CI: -0.28, -0.06; and
- setting type, $\beta = 1.31$ —95% CI: -0.59, 3.21; $t(22.81) = 1.423$; $p = .1682$; Hedges $g = 0.12$; 95% CI: -0.05, 0.70.

Compliance data was based on information from the post-trial survey, which was only completed by 61 of the 78 settings.

Table 24: Compliance adjusted effect between intention-to-treat and CACE outcome analyses for storybook vocabulary measure at post-randomisation follow-up using a two-level linear mixed model and instrumental variable analysis with robust standard errors respectively

	Primary outcome analysis: two-level ITT		Primary outcome analysis: single-level CACE	
Coefficient	Estimates	CI	Estimates	CI
Intercept	15.44 ***	13.10, 17.77	16.27 ***	13.95, 18.60
Outcome Baseline	0.60 ***	0.54, 0.66	0.58 ***	0.51, 0.65
Setting Type [Maintained]	1.15	-0.54, 2.84	1.31*	-0.59, 3.21
Intervention [planned]	-1.23	-2.72, 0.26	-1.18 *	-3.02, 0.66
Random effects				
σ^2	34.60			
τ_{00}	6.47 Setting_ID			
ICC	0.16			
N	74 Setting_ID			
Observations	691		543	
Marginal R^2 / conditional R^2	0.364 / 0.464		0.347 / 0.344	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Missing data analysis

The first step in the imputation process was to investigate if any variables were associated with a binary indicator of missingness in the primary outcome. We dummy coded a binary (1 = missing, 0 = not) missingness indicator and used this as the outcome variable in a logistic regression with predictors:

- storybook vocabulary measure at baseline;
- type of setting;
- eligibility for Pupil Premium;
- gender;
- SEND status; and
- English as an Additional Language (EAL) status.

Table 25 presents results from this regression and appeared to show that none of these variables were associated with missingness in the primary outcome.

Table 25: Logistic regression with missing data status of storybook vocabulary measure at post-test as primary outcome and predictors of missingness as per SAP

	Storybook vocabulary score (missing or not) outcome		
Coefficient	Odds ratios	CI	<i>p</i>
Intercept	0.04 ***	0.01, 0.16	<0.001
Pre-test total score	1.02	0.97, 1.07	0.406
Type of setting	0.73	0.34, 1.63	0.426
Eligibility for Pupil Premium(Y/N)	0.88	0.25, 2.41	0.823
Gender	1.02	0.50, 2.08	0.950
EAL Status	0.89	0.32, 2.16	0.806
Special Educational Needs(Y/N)	1.19	0.27, 3.63	0.791
Observations	651		
R ² Tjur	0.002		

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Following guidance from van Buuren (2018) and Austin et al. (2021) we included the variables included in the primary analysis as default. Therefore, we proceeded to impute missing scores of the storybook vocabulary measure, LanguageScreen subscales and total score, and ATLA measure using the primary adjustment set of covariates from the primary analysis and any available secondary outcome data as they are likely correlated to the primary outcome and with each other (Plumpton et al., 2016).

Table 26 shows the comparison of the primary analysis model with the pooled results from 20 imputed datasets. The pattern of findings is unaltered in the pooled results compared to the primary analysis but there is a noticeable difference in the value of the setting type parameter.

Convergence was assessed by plotting the trace plots and through visual inspection determining stationarity (see Appendix I). The number of iterations was also increased from the default to ensure stability, from 20 to 100. In terms of effect sizes and associated uncertainty, we report a Hedges g of - 0.20 (95% CI: -0.46, 0.04), which when compared to the primary analysis effect size Hedges g , - 0.19 (95% CI: -0.43, 0.05), is almost identical. Considering that the CI contains zero we cannot rule out that no intervention effect is present.

Table 26: Missing data sensitivity analyses comparing results from primary analysis to pooled imputed results from 20 imputed datasets

	Primary outcome analysis (two-level ITT)		Imputed analysis (pooled results)	
Coefficient	Estimates	CI	Estimates	CI
Intercept	15.44 ***	13.10, 17.77	13.99 ***	11.67, 16.30
Outcome baseline	0.60 ***	0.54, 0.66	0.65 ***	0.58, 0.72
Setting type—maintained	1.15	-0.54, 2.84	0.59	-0.97, 2.14
Intervention—planned	-1.23	-2.72, 0.26	-1.11	-2.47, 0.25
Random Effects				
σ^2	34.60		43.902	
τ_{00}	6.47 Setting_ID		5.31 Setting_ID	
ICC	0.16		0.11	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Subgroup analyses

In the primary outcome, planned subgroup analyses were conducted with children that were eligible for Pupil Premium (EYPP). We undertook two analyses: EYPP individuals only as a subgroup (that is, only those EYPP children—column 2, Table 27) and, secondly, as a moderator analysis, so all individuals in the trial are included and an interaction term is included in the model to specifically quantify whether EYPP moderates treatment effect (column 3, Table 27). We report the EYPP subgroup analyses against the estimates from the primary outcomes analysis for convenience.

The adjusted mean difference in expressive vocabulary scores for the sample of children eligible for EYPP is -2.30, (95% CI [-5.54, 0.94], $t(28.00035) = -1.410$, $p = 0.170$, Hedges' $g = -0.60$, 95%CI -1.43, 0.24). The negative direction of the effect indicates that children eligible for EYPP in the child-led group scored higher in expressive vocabulary, on average, compared to those in the planned group. The possible difference in impact between the two groups ranged from a 5.54 point advantage for the child-led group to a 0.94 point advantage for the planned group. The wide range of scores and much smaller sample of children means that the evidence is inconclusive. There was no evidence that eligibility for EYPP moderated the effectiveness of the two approaches ($\beta = -1.02$, 95% CI: -3.84, 1.81; $t(604.38799) = -0.707$; $p = .480$, Hedges' $g = -0.12$, 95% CI: -0.44, 0.20). The unconditional ICC (95% CI) associated with setting for the primary outcome, storybook vocabulary measure, is 0.158 (0.093, 0.220); for the conditional ICC it is 0.173 (0.102, 0.255).

Table 27: Subgroup analyses for children that are eligible for Pupil Premium comparing primary analysis, primary analysis with only those eligible for EYPP, and EYPP as a moderator in the primary analysis

	Primary analysis		EYPP individuals only		EYPP interaction	
Coefficients	Estimates	CI	Estimates	CI	Estimates	CI
Intercept	15.44 ***	13.10, 17.77	12.56 ***	6.46, 18.65	15.64 ***	13.20, 18.08
Baseline	0.60 ***	0.54, 0.66	0.68 ***	0.52, 0.84	0.59 ***	0.53, 0.65
Intervention	-1.23	-2.72, 0.26	-2.30	-5.54, 0.94	-0.84	-2.50, 0.83
Setting Type	1.15	-0.54, 2.84	1.89	-2.92, 6.70	1.09	-0.76, 2.93
EYPP					-0.12	-2.21, 1.97
Intervention x EYPP					-1.02	-3.84, 1.81

Random effects			
σ^2	34.60	32.40	32.46
τ_{00}	6.47 SettingID	7.91 SettingID	6.78 SettingID
ICC	0.16	0.20	0.17
N	74 SettingID	35 SettingID	63 SettingID
Observations	691	94	614
Marginal R ² / conditional R ²	0.364 / 0.464	0.432 / 0.543	0.378 / 0.486

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Additional analyses and robustness checks

We include an additional analysis of children for whom date of birth was not recorded. The primary analysis was rerun including children who did not meet the age inclusion criterion (including children for whom dates of birth were not available and children who were too young to meet criteria). Results are shown in Table 28. The size and direction of the effects showed very little change with this inclusion.

Table 28: Subgroup analyses for children for whom date of birth was not provided

	Unconditional primary analysis		Conditional primary analysis		Conditional primary analysis (inc DoB unknown)	
Coefficients	Estimates	CI	Estimates	CI	Estimates	CI
Intercept	16.27 ***	14.29, 18.26	15.44 ***	13.10, 17.77	13.84 ***	11.44, 16.23
Baseline	0.60 ***	0.54, 0.66	0.60 ***	0.54, 0.66	0.65 ***	0.58, 0.71
Intervention	-1.21	-2.70, 0.28	-1.23	-2.72, 0.26	-1.26	-2.76, 0.23
Setting type			1.15	-0.54, 2.84	0.93	-0.79, 2.66
Random effects						
σ^2	34.63		34.60		41.86	
τ_{00}	6.51		6.47		6.40	
ICC	0.16		0.16		0.13	
N	74		74		78	
Observations	691		691		738	
Marginal R ² / conditional R ²	0.361 / 0.462		0.364 / 0.464		0.364 / 0.448	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

As detailed above, we have two further moderator analyses assessing whether effects were significantly moderated by the skills and experience of the educator. We include these for both the primary and secondary outcome measures (Storybook Vocabulary and LanguageScreen total score).

Table 29: Storybook Vocabulary—teacher-involved moderation analysis

	Primary analysis			Teacher Involve interaction		
Predictors	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>
(Intercept)	15.44	13.10, 17.77	<0.001	17.22	14.20, 20.25	<0.001
Baseline	0.60	0.54, 0.66	<0.001	0.57	0.51, 0.64	<0.001
Intervention allocation	-1.23	-2.72, 0.26	0.106	-2.07	-4.58, 0.43	0.104
Type of setting	1.15	-0.54, 2.84	0.182	0.87	-1.07, 2.81	0.379
Is a teacher involved in delivery of the approach?				-1.32	-3.82, 1.18	0.301
Intervention Allocation *Teacher Involved				1.85	-1.71, 5.41	0.307
Random effects						
σ^2	34.60			29.06		
τ_{00}	6.47 _{SettingID}			6.95 _{SettingID}		
ICC	0.16			0.19		
N	74 _{SettingID}			59 _{SettingID}		
Observations	691			569		
Marginal R ² / conditional R ²	0.364 / 0.464			0.392 / 0.509		

Whether or not a qualified teacher is involved with the storybook sessions is not a significant moderator of the primary outcome ($\beta = 1.85$, 95% CI: -1.71, 5.41; $t(54.95849) = 1.022$; $p = .311$; Hedges $g = 0.18$, 95% CI: -0.16, 0.50). The unconditional ICC (95% CI) associated with setting for the primary outcome, storybook vocabulary measure, is 0.117 (0.064, 0.191) and that for the conditional ICC is 0.193 (0.115, 0.295).

Table 30: LanguageScreen score—teacher-involved moderation analysis

	LanguageScreen analysis			TeacherInvolve interaction		
Predictors	Estimates	CI	p	Estimates	CI	p
(Intercept)	81.97	77.45, 86.49	<0.001	82.73	76.53, 88.92	<0.001
Baseline	0.80	0.67, 0.92	<0.001	0.78	0.65, 0.92	<0.001
Type of setting	-5.07	-8.26, -1.88	0.002	-4.91	-8.95, -0.86	0.018
Intervention allocation	-3.53	-6.29, -0.77	0.012	-3.53	-8.59, 1.53	0.171
Is a teacher involved in delivery of the approach?				-1.71	-6.75, 3.34	0.507
Intervention Allocation * Teacher Involved				0.96	-6.04, 7.96	0.788
Random effects						
σ^2	117.73			120.50		
τ_{00}	18.17 SettingID			22.33 SettingID		
ICC	0.13			0.16		
N	70 SettingID			55 SettingID		
Observations	586			476		
Marginal R ² / conditional R ²	0.263 / 0.362			0.247 / 0.365		

Whether or not a qualified teacher is involved with the storybook sessions is not a significant moderator of the secondary outcome, ($\beta = 0.96$, 95% CI: -6.04, 7.96; $t(46.07) = 0.270$; $p = .789$; Hedges $g = 0.04$, 95% CI: -0.32, 0.42). The unconditional ICC (95% CI) associated with setting for the primary outcome, storybook vocabulary measure, is 0.118 (0.048, 0.192) and that for the conditional ICC is 0.156 (0.069, 0.251).

Educator questioning skill

Next we considered whether educator questioning skill, as measured in the pre-trial educator survey, was a significant moderator of the primary or secondary outcome.

Table 31: Storybook Vocabulary score with questioning skill moderation

	Primary analysis			Questioning skill interaction		
Predictors	Estimates	CI	p	Estimates	CI	p
(Intercept)	15.44	13.10, 17.77	<0.001	14.09	9.73, 18.45	<0.001
Baseline	0.60	0.54, 0.66	<0.001	0.61	0.54, 0.67	<0.001
InterventionAllocation1	-1.23	-2.72, 0.26	0.106	2.37	-3.52, 8.25	0.430
Type of setting	1.15	-0.54, 2.84	0.182	0.63	-1.26, 2.52	0.512
Educator questioning Skill				0.47	-0.94, 1.88	0.514
InterventionAllocation: quest skill				-1.12	-3.21, 0.97	0.294
Random effects						
σ^2	34.60			29.87		
τ_{00}	6.47 SettingID			5.29 SettingID		

ICC	0.16	0.15
N	74 SettingID	55 SettingID
Observations	691	546
Marginal R ² / conditional R ²	0.364 / 0.464	0.409 / 0.498

Educator questioning skill was not a significant moderator of the primary outcome ($\beta = -1.12$, 95% CI: -3.21, 0.97; $t(47.75154) = -1.051$; $p = .298$; Hedges $g = -0.18$, 95% CI: -0.52, 0.16). The unconditional ICC (95% CI) associated with setting for the primary outcome, storybook vocabulary measure, is 0.089 (0.047, 0.145), and that for the conditional ICC is 0.151 (0.081, 0.233).

Table 32: LanguageScreen score with Questioning skill moderation

	LanguageScreen analysis			Questioning skill interaction		
Predictors	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>
(Intercept)	81.97	77.45, 86.49	<0.001	86.42	77.14, 95.69	<0.001
Baseline	0.80	0.67, 0.92	<0.001	0.82	0.69, 0.96	<0.001
Type of setting	-5.07	-8.26, -1.88	0.002	-5.44	-9.61, -1.26	0.011
InterventionAllocation1	-3.53	-6.29, -0.77	0.012	-2.37	-14.42, 9.68	0.699
Educator questioning skill				-2.30	-5.21, 0.60	0.120
InterventionAllocation: quest skill				0.21	-4.04, 4.47	0.921
Random effects						
σ^2	117.73			118.59		
τ_{00}	18.17 SettingID			20.04 SettingID		
ICC	0.13			0.14		
N	70 SettingID			51 SettingID		
Observations	586			454		
Marginal R ² / conditional R ²	0.263 / 0.362			0.276 / 0.381		

Educator questioning skill was not a significant moderator of the LanguageScreen measure ($\beta = 0.21$, 95% CI: -4.04, 4.47; $t(40.77) = 0.099$; $p = .92$; Hedges $g = 0.02$, 95% CI: -0.36, 0.40). The unconditional ICC (95% CI) associated with setting for the primary outcome, storybook vocabulary measure, is 0.105 (0.040, 0.177) and the conditional ICC (95% CI) is 0.145 (0.058, 0.240).

Mediation analysis for number of books used during the intervention

Overall, session records indicated that on average children in the planned approach encountered 21.8 different books over the 36 sessions of the trial. This varied widely, with a minimum of five and a maximum of 37, and a median of 23 books. This was marginally less than the mean number of different books encountered in the child-led approach— $t(46) = 2.38$; $p = .021$.

Table 33: Session record data on number of books read

Approach	Mean total number of books read (SD)	Mean total number of different books read (SD)	Maximum different books read	Minimum different books read
Child-led	34.97 (8.44)	28.89 (8.81)	49	9
Planned	31.58 (6.50)	21.79 (11.26)	37	5

As the logic model had specified that the number of different books encountered could be a mechanism by which the approaches were more or less effective, it was felt appropriate to assess this variable as a potential mediator. As the total (direct and indirect) effects of approach were not significant in predicting the primary outcome variable, Storybook Vocabulary, this analysis was not pursued. However, the total (direct and indirect) effects relating to each approach were significant in predicting the secondary outcome (LanguageScreen score): the intervention effect significantly increases the outcome score by 4.33 units ($p = .024$). Because the 95% confidence interval does not cross zero (-8.40, -0.53) this is a statistically reliable result.

The ACME (Average Causal Mediation Effect) is not significant. There is no statistically reliable effect of the number of books read on the outcome. The estimate is -0.59 but the p -value (0.340) and the 95% CI (-2.04, 0.59) crosses zero.

The Average Direct Effect (ADE) directly increases the outcome by 3.74 units after taking account of the number of books read ($p = .076$). The possible size of this effect varies between a point advantage for the child-led approach to a 0.28 point advantage for the planned approach. Inclusion of number of books read accounts for around 12.7% of the effect of approach, but this is not a statistically reliable proportion of the effect ($p = .356$).

This mediation analysis is inconclusive: the effect of number of books used does not significantly affect LanguageScreen outcome, but inclusion of this factor does slightly reduce the effect of approach suggesting it may play a role in the outcome.

Implementation and process evaluation results

Response rates

For the pre-trial setting survey we received responses from 67 of the 82 settings that carried out pre-testing. For the pre-trial educator survey we received 100 responses from 64 different settings. For the post-trial setting survey we received responses from 62 different settings from the 78 settings remaining in the trial. This data therefore represents 78% to 82% of the total sample in each case.

Session records were completed and returned for analysis by 48 of the 78 settings (61.5%), indicating a moderate level of missing data and limiting the potential usefulness of this data. Session records were more likely to be returned by settings in the child-led condition than those in the planned condition, possibly indicating that record keeping was more onerous in the planned condition. Examples from completed session records are included in Appendix M. We received session records from 30 of the 38 settings in the child-led approach and 18 of the 38 settings in the planned approach. This high proportion of missing data limits our interpretation of the findings from session records in the planned approach.

Implementation (RQ7)

Child-led approach—fidelity to core principles

The child-led approach focused on three core principles: (1) children's agency in book selection, (2) following children's interests, and (3) spontaneous, child-initiated conversations. Evidence of fidelity is gathered from case study interviews and observations (four settings) and from the post-trial educator survey (31 settings). We also reviewed session records where available. We considered evidence for use of the core principles in each of these different sources of information. For example, within the post-trial survey we assessed how often the settings mentioned one of the core principles when describing what they liked or disliked about their approach. This is not an ideal measure of fidelity but it provides some useful information and avoids the potential response bias we might see in answers to direct questions about implementation (note that we used one direct question about implementation as a measure of compliance, above).

Open text comments in the post-trial survey indicated good adherence to the core principles across the settings. Over half of settings implementing the child-led approach (52%) explicitly mentioned children's choice or ownership when describing what they liked about the approach. Nearly one-third of the educators (29%) explicitly mentioned following children's interests, child initiation, or allowing children to share their interests. Similarly, session records for settings in this condition all reported a variety of children choosing the book each day. There was one setting whose session records indicated that they did not follow the child-led approach fully: they sometimes stated 'topic' for the 'child choosing', suggesting that the books were chosen because they aligned to the current topics. The remainder indicated that books were chosen by children.

Taken together, these findings suggest that most settings following the child-led approach were following this element of the approach.

In the interviews, educators described intentionally not pre-planning and instead following child-initiated topics for discussions. They acknowledged the challenge of occasional divergence from the story due to child-initiated topics and stated that it required experience and facilitation skills to redirect to the main topic without discouraging children's contributions.

'The question is, I can plan as much as I like to plan ... And the children will come at me with completely different answers or their own questions. So as practice ... I just go with what they want. How they want to take the session, how they want, what they want to ask' (CL1).

'If the children ask any questions during the story ... I'm happy to stop the story and just see make sure it's appropriate to the book ... as long as the question that's coming from the child is ... to do with the story. Yeah. Then we'll move forward with it' (CL2).

Settings varied in how they managed the book selection process in the child-led approach. Most developed systematic rotation processes ensuring every child had the opportunity to select books for the reading sessions. Session records indicated that one setting allowed group votes to choose the book, while another allowed children to choose a book of the week rather than for that particular session. The remainder of the settings had one child choosing each day.

Some settings let children choose from a small, curated selection of books, while others allowed wider choice. For those that offered a limited choice, it varied how often that choice was changed. Some settings retained a selection for several days or weeks to encourage children to get to know the books. Others changed the selection every day. Preparation involved physical set-up and book availability rather than advance planning of specific questions or vocabulary. Educators explicitly valued ‘going with’ children’s responses in the moment.

‘We have a whole range of books just up on the shelf ... and the staff, the practitioners, would swap them out. Sometimes, if we know that it’s a particular week—so like Dinosaur Week or Bee Week—we would put a couple of selections in to do with that ... Every Friday, we swap out the books that have already been used for that week’ (CL3).

Overall, session records indicated that, on average, children in the child-led approach read 28.8 different books over the 36 sessions of the trial. This varied widely, with a minimum of nine, a maximum of 48, and a median of 29.5 books.

Planned approach—fidelity to core principles

The planned approach focused on advance planning, targeted vocabulary instruction, pre-planned questioning, and structured implementation. Evidence of fidelity is gathered from case study interviews and observations (four settings), session records, and the post-trial educator survey (31 settings).

All of the session records indicated planning for almost all of the sessions. One setting allowed children to choose books for a few of their sessions. However, the low rate of return for session records (18/38) in the planned approach compared to the child-led approach (30/38) suggests that there may be some gaps in fidelity in some cases.

Nearly half of survey respondents (45%) in the post-trial survey mentioned planning or preparation, using phrases like ‘pre-thought about what we were doing’ and ‘planning questions beforehand meant I ensured they were well thought out’ (P1). Interview data showed systematic implementation: P1 selected curriculum-aligned books and identified unfamiliar vocabulary like ‘spade’, ‘spray’, and ‘hose’ for pre-teaching, while WRN743 brought a Viking helmet to pre-teach context. The majority of the settings stated that the approach fitted well with their routine.

Nearly a quarter of educators (23%) mentioned targeted vocabulary teaching as demonstrated in comments such as ‘vocabulary was chosen and discussed before session’ and ‘I liked being able to target vocabulary’. P2 noted: ‘We would have looked through it and took out any keywords ... any words that maybe they don’t know.’ Questioning was mentioned by 45% of educators.

Most of the session records indicated key vocabulary that the educator planned to focus on in the story and almost all highlighted some planned questions.

While educators did not directly comment about adhering to the structure of the approach, they commented on how it supported consistent implementation: ‘it was well structured for other adults to deliver’ and ‘allowed staff to be confident in asking questions’.

The observations and interviews indicated that educators took varied approaches to planning the questions. One setting used a particular framework, ‘Talk for Writing’, to plan the questions used each session. Another focused on using open questions. P3 explained: ‘I’ve tried to plan questions ... I’ve tried to do more open questions that will get them to give some response rather than yes or no.’ Case study observation in one setting showed evidence of an educator using the questions to scaffold conversations. Critically, educators used plans as scaffolds rather than scripts.

'The who, what, where, and why questions, that's what was the focus ... the questions flow depending on what the children are giving us ... Sometimes I've forgotten some questions because the children asked their own' (P4).

Session records indicated that a significant number of settings in the planned approach repeated the same book over two or three sessions, with different questions asked each day. This is a contrast to the child-led settings where it was rare for the same book to be read two days in a row.

Adaptations—child-led approach

The majority of educators reported making no change to the child-led approach stating that it naturally fitted within their existing storybook reading routines. In some cases this meant that sessions were held with the whole class while in others smaller groups were used.

The most common adaptation involved practical adjustment in the book selection process. One setting reported using a book voting system to ensure equal participation in the choice process, giving the whole class a choice of three books from the book corner to vote on. This helped to maintain children's choice while negotiating the practical challenge of multiple children wanting to select.

One setting reported sometimes repeating the same book for consecutive days when it was observed that children had not fully understood during the first reading.

In the case studies, one setting provided a choice between the same five different books consistently over a few weeks. The books were on display in the book corner and there were linked activities to these books within the class. This meant that children still had the opportunity to choose a book while also allowing educators to select appropriate books and plan linked activities.

Adaptations—planned approach

As with the child-led approach, most settings reported that the planned approach fitted within their existing routine and that they made no particular changes. Most of the adaptations mentioned involved changing the time of the sessions or similar minor changes that were anticipated within the guidance. As with the child-led approach, some settings carried out the storybook reading with the whole group while others used smaller groups.

A few settings mentioned changes to accommodate children's interests or preferences, either by re-reading books from previous sessions or reading books brought from home if children requested it. They also mentioned responding to children's spontaneous questions and comments during the sessions.

Case study settings did not show significant changes made to the approach. As mentioned above, one setting used the 'Talk for Writing' framework to plan questions and activities.

One setting emphasised that it had been easy to fit the sessions into the normal routine:

'In terms of the preschool room, it fitted in that they do story times regularly, so they normally have one in the morning, they'll have one before lunch, they have one in the afternoon. They normally read at least three stories, I would say, in a day. So I was able to slot in generally in one of those' (P3).

Implementation—summary

Case studies, session records, and survey data indicated a good level of fidelity and indicated clear differentiation between the two approaches. There was also significant variation within each approach in line with the philosophy of the Teacher Choices approach. Analysis of the session records indicated some potentially important differences between the approaches:

- settings in the planned approach were less likely to return their session records, perhaps indicating they had not been completed;
- on average, there were more different books encountered in the child-led condition than the planned condition; and
- educators in the planned approach tended to plan to reread the same book two or three times in a row, whereas this rarely happened in the child-led approach.

Dosage (RQ8)

We have two separate estimates of dosage: from the session records and from the post-trial survey. We had intended to use session records to record dosage at a child level, however, as there was a large amount of missing data from the session records, we also report findings from the post-trial survey.

Session records

Session records were used to record how many sessions each child attended. The average number of sessions attended was very similar for the two conditions (child-led, $m = 25.39$; planned, $m = 25.84$).

The number of sessions attended was not significantly correlated with post-test vocabulary score for either condition (child-led: $r(193) = 0.00$; $p = .99$; planned: $r(98) = -0.02$; $p = .87$). This association remained nonsignificant after controlling for pre-test vocabulary score (child-led: $r(190) = 0.08$; $p = .28$; planned: $r(95) = 0.05$; $p = .62$). There was therefore no evidence of a relationship between number of sessions attended and vocabulary growth, however, the large amount of missing data makes it difficult to draw conclusions from this.

Post-trial survey

In the post-trial survey, educators were also asked to estimate how many sessions were provided, though this information was at the level of the setting rather than individual child. Responses indicated that educators in the child-led approach delivered slightly more sessions, on average, than those allocated to the planned approach—37.2 versus 34.8—though there was a high level of similarity between the two conditions.

Table 34: Reported dosage

Approach	Average weeks	Minimum weeks	Maximum weeks	Average weekly sessions	Minimum weekly sessions	Maximum weekly sessions
Child-led (23)	12	10	14	3.1	3	5
Planned (25)	11.7	7	15	3	2	3

Perceived value and impact of the sessions (RQ9)

Perceived strengths of the child-led approach

In the post-trial educator survey, educators were asked, ‘What did you like about the storybook reading approach you followed?’ Almost all in the child-led condition (26/31 or 84%) commented that they valued the approach for the fact that it allowed children to take the lead in the storybook session. Respondents did not always explicitly articulate why that was valuable but several mentioned engagement (9/31, 29.0%), enjoyment (4/31, 12.9%), and a feeling of ownership (5/31, 16.1%) for the children. For example:

‘Children were more engaged as it was their choice and their listening skills are more engaged.’

'The children got to choose a book and felt important and listened to for doing so.'

'The children engaged lots and asked questions around the story more. Knowing that the books that were chosen, the children were interested in.'

Some educators (4/31 or 13%) observed that children often shared their reasoning for choosing particular books leading to meaningful conversation and connection to the stories. Educators observed that children were more interested and enthusiastic when reading stories. As one case study educator commented:

'You have, if you've chosen that book, your engagement is a lot more focused because you're excited ... even some of you children who do struggle a bit with their listening and attention ... their focus and their concentration is for longer periods of time' (CL3).

The approach also supported children with English as an additional language, with educators noting that some EAL children showed particular interest in books, which they chose repeatedly. One respondent noted that what they liked was:

'That the children had the opportunity to choose whichever story they wanted, even if we had read it the day before. Children also became really familiar with certain stories, were joining in and repeating phrases from the stories. We also had some EAL children who showed a particular interest in one or two books' (CL4).

Some educators (6/31) explicitly talked about how they felt that the child-led sessions led to more expressive language use by the children:

'Encouraged children to be more confident in speaking about the books. We could see children's language develop as the programme ran.'

Some educators also thought that it helped to broaden children's exposure to diverse texts and vocabulary while simultaneously allowing them to familiarize themselves with stories at their own pace. Additionally, the opportunities for questioning and discussion were particularly valued. Educators felt that the opportunity to choose books for reading improved children's engagement with the stories resulting in better acquisition of language and enthusiasm in sharing their own ideas and thoughts.

'They were able to recall some of the facts from the story a lot more easily because they're focused and engaged with it ... they're using keywords from the book ... engagement levels, the language that's coming back, the recall from the story, you know, is really good because they are focused more, they're more excited to share their ideas and things' (CL3).

Limitations and challenges—child-led approach

Around a third of the sample of educators felt there were no particular limitations with the child-led approach (10/31 or 32.2%). The most commonly cited issue, mentioned by 22.5% of respondents, was that children tended to choose the same books repeatedly rather than exploring new texts. Our analysis of the session records, however, showed that this perception did not seem to be accurate: children in the child-led approach were exposed to a wider variety of books, on average, than those in the planned approach. It may have been that educators were aware of this risk and managed child choice by rotating books regularly. One educator noted:

'No limitations as long as we had new books regularly rotated in the reading area (each half term), which kept the children interested and exposed them to new texts.'

Managing the book-choice process itself presented difficulties in some settings. Multiple children often wanted to be the ones to choose the story, which initially created disappointment, although children eventually understood that everyone would get turns. Some settings struggled with ensuring all children had equal participation in choosing books, especially in larger groups or where children attended part-time on different days. A few settings let all children listen to the stories but only children enrolled in the trial select the books, which caused some unrest (and had not been the intention of the researchers). However, in most settings, once the routine for choosing a storybook was established, children engaged well:

‘Children all wanted to be able to choose a story, so in the beginning this was difficult. As the weeks went on [and] the children realised they all would be able to choose a book if not multiple books, then they understood’ (CL5).

The responsive nature of the approach was also challenging for some educators who valued planning and structure. It was commented that when they did not know which book would be chosen it was difficult to pre-plan props, sensory items, and activities to support learning. It was also noted that the approach didn’t allow educators to use the storybook reading sessions strategically to support current themes or learning focuses or to plan specific learning outcomes in advance. The tension between responsiveness and planned teaching emerged as a concern for some practitioners who felt that ‘children sometimes need planned activities to extend knowledge and understanding’.

Notably, as described above, most of the settings continued with other storybook activities alongside Story Choices and 70.4% of the child-led settings continued to include acting out stories or using story sacks in their routine.

One observation demonstrated a potential limitation with the child-led approach. During one of the observed reading sessions with a less experienced educator, a child selected a book that the educator did not know (*Handa’s Surprise*). The book contains some unusual vocabulary, (for example, ‘guava’) and builds towards a surprise that is shown in the pictures but not the writing. As the educator was not aware of this in advance, they were not able to direct the children’s attention to the pictures to help build their understanding. This showed how important it is that educators are knowledgeable about the storybooks offered in the child-led sessions to allow them to lead learning appropriately.

Strengths—planned approach

When asked about what they liked about their approach, educators implementing the planned approach often mentioned its structure, the opportunity to plan ahead, and the intentionality to incorporate vocabulary instruction. Some (6/31, 19.4%) mentioned that they found it helpful to be able to pre-select key vocabulary to teach before or during the storybook reading session. This element of the approach helped them to choose books with targeted vocabulary that children might not select independently; moreover, it helped educators to link vocabulary and themes across different contexts and to select books related to current topics. This provided the children with repeated exposure to target vocabulary and improved understanding.

‘It allowed me to introduce new vocabulary to the children. I could pick books with challenging vocabulary and read books the children probably wouldn’t pick.’

Educators liked the elements of selecting learning objectives, questions, and anticipating discussion points. This preparation helped them to develop clarity on what they wanted children to gain from each book. The structured questioning framework was particularly appreciated with educators (8/31, 25.8%) noting that having pre-planned questions helped to make them feel confident about their storybook reading. They felt that the structure encouraged the use of different question types (completion, recall, open-ended, ‘wh’-questions, distancing) rather than spontaneous questions that might be narrower in scope. As one educator noted: the approach ‘made me think more carefully about the questions I asked’.

Additionally, the planned approach was thought to be effective in supporting less experienced staff with interactive storybook reading sessions. The pre-planned questions and structure helped to ensure that children experienced rich interactions regardless of which practitioner led the session. Moreover, compatibility with existing practices emerged as another key strength. Many educators noted that the planned approach aligned well with their established routines, topics, and teaching frameworks, particularly for those already using structured literacy approaches. They valued being able to repeat key texts to improve children’s familiarity with books. One educator stated that the approach also supported transition preparation and self-regulation for ‘rules and routines’.

A few educators (3/31, 9.7%) talked about engagement as a benefit of the planned approach but this was less commonly mentioned than in the child-led approach.

Educators highlighted the value of the routine around the storybook sessions:

'But actually, the more I've done it, the more engaged they've become. And every time I walk over now, one of the little boys is, like, "Oh, have you come to read a book?"' (P3).

Some educators emphasised the value of the approach for building language skills:

'I definitely think it's been useful because like I said, in terms of what we've seen in the Wellcome [language] screening is that ... we've had some children that maybe at the start of the year were on red and now ... are on green. Again, I had one child at the start of the year that didn't talk. So he started with us in January. Hadn't been talking, and sort of over the last six months, just an explosive amount of language that's just come out and now he's talking every day and like I said, that's really beneficial for us really' (P4).

Limitations and challenges—planned approach

Similar to the child-led approach, 22.5% of educators (7/31) felt that the planned approach had no particular limitations. The most consistent limitation, mentioned by 25.8% (8/31) of respondents, was the inability to respond to child interests or preferences as effectively as they would have liked. The second most consistent limitation, also mentioned by 22.5% (7/31) of respondents, was that the children sometimes lacked engagement during sessions. This was a notable contrast with the child-led condition, where lack of engagement was not mentioned as an issue.

Educators noted that 'not reading a child's first choice of book' meant some children were less interested in stories. It was commented that children sometimes appeared disengaged or didn't want to engage on certain days when the chosen book didn't align with their interests.

Furthermore, educators felt that they 'couldn't ask those in-the-moment questions or focus on questions from children that slightly deviated from those we had planned for'. The requirement to follow pre-planned questions felt restrictive to some practitioners who valued following children's spontaneous interests and 'going off on a tangent as to where they take the book'. One educator noted that 'the planned-for stories took away from the responsive element, for example, reading a book linked to the weather that day', highlighting how advance planning could prevent timely, contextually relevant teaching moments.

The third most commonly mentioned limitation was that the planned approach involved a significantly increased workload. This was noted by 5/31 or 16.1% of the respondents.

Multiple educators noted that planning questions and selecting appropriate books was time consuming and contributed to bigger workload compared to spontaneously choosing a book for reading. The record-keeping required for the trial was an added burden but educators acknowledged this was necessary for research purposes. Some educators found it 'difficult to tell before reading the story what words/parts of the story may need clarifying', meaning careful preparation didn't always match children's actual interest in the moment.

There was some suggestion that the planned approach might not facilitate the engagement of all the children in the session, with the more confident children engaging more with the questions. One educator noted:

'But on the whole, they've seemed to have listened, enjoyed. It's been the same sort of ones that have answered the questions. I've found the more confident ones, but they all have been looking, listening' (P4).

However, she emphasized that all the group have enjoyed the stories, including those with additional needs:

'But I have to say, they've all been engaged and happy to listen, on the whole, unless they've, you know, had a bad day and they're completely dysregulated and can't focus anyway. They've done well, yeah' (P4).

Examination of the results for children eligible for EYPP is relevant for considering this issue. Similar to the larger sample, there was no reliable evidence that one approach was more effective than the other in improving storybook expressive vocabulary in children eligible for EYPP and there was no significant interaction between approach and EYPP membership. The effect size of the difference between approaches is in the same direction, but somewhat larger in EYPP children ($g = -$

0.60 versus $g = -0.19$), giving some indication that the effects are similar for EYPP children as they are for the group as a whole.

Sustainability and continuation

In the post-trial educator survey, 93.5% of those in the child-led condition and 87.1% of those in the planned condition said that they would continue to use the approach in the future. When asked if they would make changes using it in the future, 56.5% said they would make no changes—67% of the child-led approach and 45% of the planned approach.

One educator in the child-led approach stated:

'The approach really complements and can strengthen existing practice. Children choosing books gives autonomy and ownerships and provides joy and opportunities for learning that may have been missed and locked out of following a planned-only approach.'

The most commonly suggested change was to employ a hybrid approach, having some planned sessions and some child-led sessions. This was mentioned by 19.4% of the respondents and was more commonly mentioned by those in the planned condition (29.0%, compared to 9.7%).

One educator in the child-led approach said:

'I prefer to do a mix of planned storytime sessions with books and props of my choosing (linked with the current theme/season/interest) which the children can access afterwards, combined with ad-hoc stories requested by children—one to one or in small groups, as well as larger group storytimes with books the children choose.'

Other suggested changes differed by approach. The most common planned change in the child-led approach was to consider how changing the options of books available could help guide children's choices (mentioned by 12.9% of respondents). Suggestions included swapping book choices more frequently and adding more book options for children to choose from. It was also suggested to use the same book on consecutive days to improve understanding and link activities to learning, rather than having random repetition based on individual children's spontaneous selections.

A similar proportion of those in the planned approach said that they would reduce the amount of written planning they would do in the future. For example, one educator in the planned approach stated that they would—

'not have it as restrictive, but just have prompts that could be used if a practitioner wished to.'

This issue was echoed by the low return rate of the session records from the settings using the planned approach, possibly indicating they had not completed the planning as requested.

The overall pattern suggests that while planned approach educators recognised the value of structured storybook reading with planned literacy instruction, they also felt the approach needed flexibility to allow children's choice and responsiveness to emerging questions and discussion points.

Cost and resources

As a Teacher Choices Trial, a cost evaluation was not undertaken. The costs of the choice approach are minimal, as educators are encouraged to change existing practices rather than add further ones. In particular, settings were encouraged to use their own books for the storybook sessions, though we did not ask whether any settings bought new books in order to take part.

Summary—perceived value and impact of the sessions

In general, the educators were largely positive about their allocated approach, whichever one they were using. They felt that both approaches were beneficial for language development, however, there were some differences that arose: educators in the child-led approach consistently talked about how the sessions facilitated engagement, enjoyment, and confidence

around books, while educators in the planned sessions seemed to suggest that engagement from the children was a little more variable. There were also more comments about the workload involved in the planned approach.

Importantly, educators saw the value of both approaches. Some comments indicated that the educators in the child-led sessions would like to include books that they selected or planned, while educators in the planned sessions would like to sometimes include books chosen by the children.

Moderators and contextual factors (RQ11)

Setting context

Details on the variation across setting contexts is presented in Appendix L. Overall, this demonstrates that the contexts vary widely across different settings but that there are no systematic differences between settings in the two different conditions, except that there are qualified teachers in a higher proportion of settings in the child-led approach.

Moderation analyses assessing the impact of two different factors are presented in the impact evaluation and show no significant effect of the skills or experience of the educators.

The impact of contextual factors—quantitative analysis

In order to assess the impact of contextual factors we carried out a series of multiple regressions to predict the primary and secondary outcomes (storybook vocabulary and LanguageScreen total score). In the first step of the regression we controlled for pre-test vocabulary score and in the second step we controlled for intervention condition. We then added the contextual variables individually on the third step. These analyses should be interpreted with caution because we have not accounted for the clustered nature of the data.

Whether the setting was maintained or PVI made a marginal difference to the LanguageScreen outcome. This is in line with the findings from the impact evaluation, where setting type influenced LanguageScreen but not vocabulary.

Guidance and materials

In the case study interviews, educators across both child-led and planned conditions described the guidance materials as 'clear', 'simple', 'self-explanatory', and 'really useful'. The structured format, including do's and don'ts sections and FAQ components, was effective in reducing implementation anxiety. Settings explicitly stated they faced no major challenges with the guidance. However, despite overall satisfaction, educators made some suggestions to further improve the guidance specifically for the benefit of less experienced practitioners. As suggested by one setting, a video example of the child-led approach would be beneficial for practitioners to learn about effectively facilitating the approach, especially in relation to questioning and maintaining conversations.

Conclusion

Table 35: Key conclusions

Key conclusions	
1.	Children in both groups made significant progress over time on the bespoke storybook expressive vocabulary measure. The mean score in the child-led group (32.27) was 1.22 points higher, on average, compared to the mean score in the planned group (31.05). The result has a moderate to high security rating.
2.	Children eligible for EYPP in the child-led group achieved a score (31.25) that was 1.81 points higher in storybook expressive vocabulary, on average, compared to children in the planned group (29.44). These results have a lower security rating than the overall findings as they were based on a smaller sample of children.
3.	Children in the child-led group achieved a score in wider oral language skills (100.07) that was, on average, 3.98 points higher than that in the planned group (96.10).
4.	There was no evidence of a difference in attitudes to literacy between the children in the planned storybook settings, on average, and those in the child-led settings.
5.	Many of the educators in the child-led group felt that the child-led approach led to high child engagement and empowerment around storybook reading, while some of those in the planned approach felt that children were sometimes not very engaged. In addition, children in the child-led condition were, on average, exposed to a wider variety of books, perhaps due to letting the children choose. Educators in the child-led group also felt that that approach led to more opportunities for expressive language use by the children.

Impact evaluation and IPE integration

Evidence to support the logic model

The IPE indicated that educators generally found the guidance clear and feasible and that it showed good fidelity to the principles of each approach. Comments about the limitations of each approach mentioned how the allocated approach reduced opportunities for the other approach, which was part of the intention in order to compare two different conditions to storybook reading with minimal contamination between conditions.

This evidence gives some confidence that the two approaches were generally implemented as intended, however, we were limited in our assessment of compliance as we did not review practice directly except in eight case study settings. We asked settings if they had followed the guidance given in the post-trial survey but only 61 replied of which 60 stated that they had delivered the approach as planned. Nonetheless, there was little evidence—from session records, case studies, and the post-trial survey—of settings not following the approach.

Responses from the educators indicated that the children in the child-led condition seemed to be engaged and enthusiastic about their sessions, as predicted by the logic model, though in contrast to the logic model, this did not translate into more positive attitudes towards books, at least on our measure.

In contrast, the planned condition allowed educators to think carefully about what books to introduce and how to support language development. This also aligns with the logic model. The logic model theorised that the planned approach would lead to greater gains in vocabulary and oral language development, however, contrary to these predictions, the child-led approach was associated with higher scores in storybook expressive vocabulary, albeit with statistical uncertainty, and in broader oral language skills. A revised logic model detailing the potential causal chain is shown in Figure 8. Based on insights from the IPE, there are a few possible reasons for the possible impact of the child-led approach, each suggested by comments from educators taking part. First, the child-led approach was associated with higher engagement in the sessions. Second, the child-led approach encouraged opportunities for children to verbally express their views about books. Finally, the child-led approach allowed children to revisit favourite books several times and become familiar with the language. This

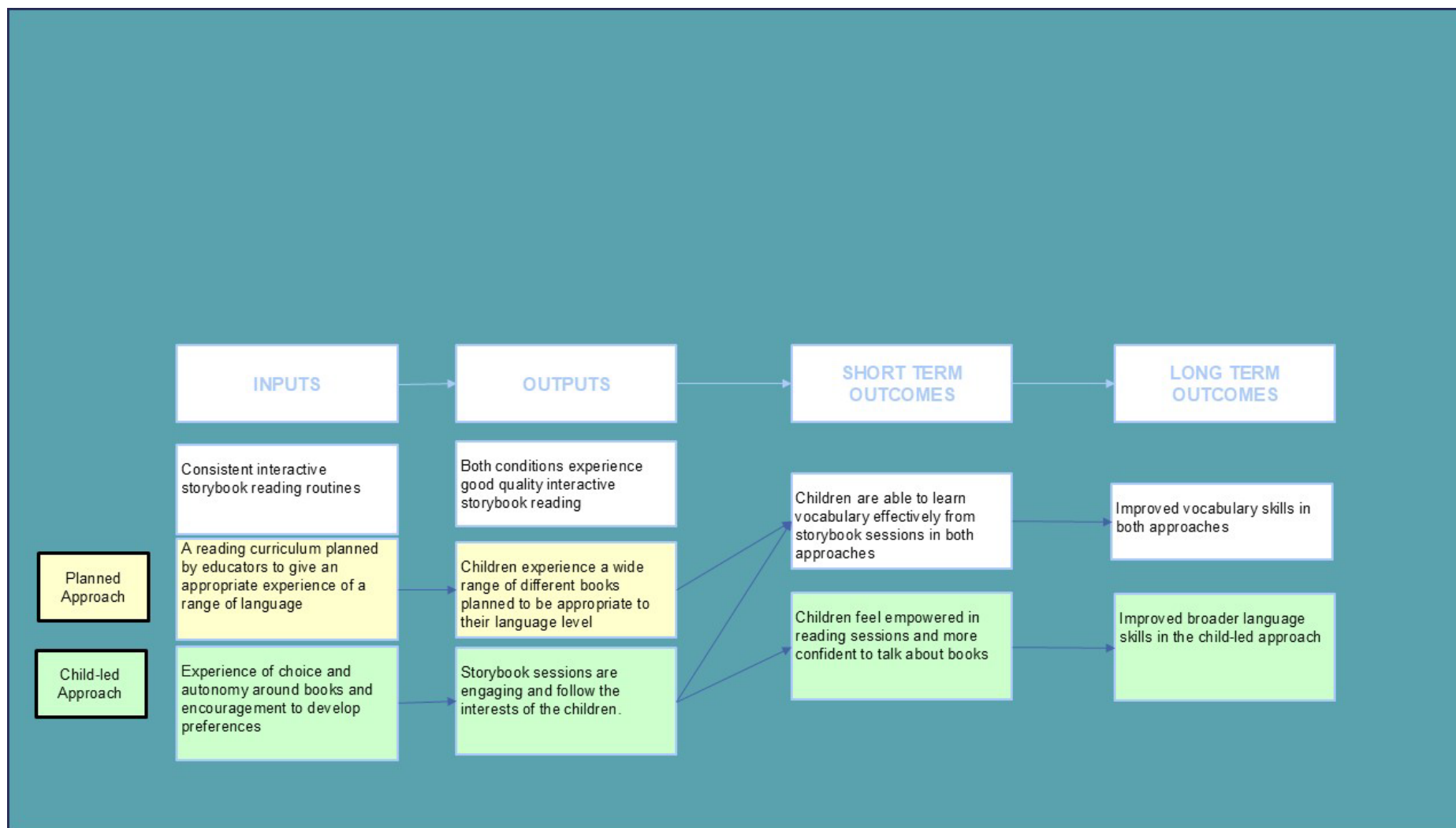
may have been particularly useful for children with lower levels of language or EAL. Further research is needed to distinguish between these alternative explanations.

We intended to measure child attitudes to literacy activities with the ATLA measure. However, this measure showed a lack of sensitivity, with many children scoring at ceiling (in other words, many children felt very positive about books and reading both before and after the trial. Previous uses of this measure have focused on 4-6 year old children, rather than 3-4 year olds, and it is likely that children as young as 3 have difficulty reflecting on their attitudes to literacy reliably. It is therefore arguably more appropriate to use the educator reports of child engagement in the storybook sessions themselves, rather than children's views of storybook reading in general, to assess engagement and enjoyment. Of course, this has the important limitation that educator views are subjective and do not involve comparison of the two approaches, as they have only delivered one of the two approaches.

The settings taking part in the trial were very varied, with total numbers of children attending the setting ranging from eight to 132 encompassing both small, targeted provisions and large settings catering for children of different age groups. They also varied in their approach, context, and in the socioeconomic demographic areas in which they were situated. In general, results were consistent across a wide variety of contexts. However, there were some indications that children showed better progress in maintained settings on average. There are many reasons why this might be that go beyond the scope of this trial. The EPPE study argues that preschool settings that are effective in improving language outcomes are likely to have highly qualified staff and a focus on education rather than purely childcare (Sylva et al., 2004). Our analysis suggested that the sessions were led by a qualified teacher in 55.3% of maintained settings whereas in the PVI settings only 35% were led by a qualified teacher, which is in line with this hypothesis, though a smaller difference than might have been anticipated.

The advantage for the child-led reading approach was mirrored in the EYPP sample, with a Hedges *g* effect size of -0.60, however, due to wide variability in the sample this did not reach significance. Nonetheless, it is an indication that the approach was potentially effective for children receiving EYPP.

Figure 8: Revised logic model to illustrate findings



Interpretation

Overall, these findings have implications for storybook reading in preschool classrooms. It is well established that interactive storybook reading is a useful tool to help develop early language, and much early years pedagogy is storybook based. However, the dilemma of whether it is more effective to choose and plan books for the children or whether to let the children choose for themselves has not previously been addressed in the literature. This trial indicates that children in both groups made significant progress over time on the bespoke storybook expressive vocabulary knowledge measure. The mean score in the child-led group (32.27) was 1.22 points higher, on average, than that in the planned group (31.05) but the wide range of scores means that the evidence is inconclusive. The possible difference in impact between the two groups ranged from a 2.72 point advantage in the child-led group to 0.26 point advantage in the planned group. However, there are some signs of greater improvement in broader language skills when children are given agency to choose storybooks for themselves. Children in the child-led group achieved a score in wider oral language skills (100.07) that was, on average, 3.98 points higher than that in the planned group (96.10). The possible difference in impact between the two groups ranged from a 0.77 point to a 6.26 point advantage for the child-led group. We suggest this may be a result of higher levels of engagement in the child-led storybook reading sessions. We believe that there is, however, a place for both approaches to storybook reading, and many of the educator responses indicated that in the future they plan to have both planned sessions and child-led sessions in their weekly routine.

A key issue for educators in the child-led choice approach to address was how to manage children choosing a book in a group reading situation. They often used a rota system to ensure that each child had a turn in choosing. Settings also varied in whether they allowed the child free choice or to choose from some pre-selected options. Preselecting a few options may allow both a degree of planning and a degree of responsiveness.

McGeown et al. (2024) review child agency and book choices in primary school reading and they argue that children have to be taught how to choose a book—how information on the cover can help them decide what the book is about and whether they are likely to enjoy it. This work is aimed at older children reading independently but some of the principles are likely to be relevant to preschool storybook reading. It may well be that helping children to choose a book to read helps them to think about their own reactions and preferences to different types of book, which will help to build reader self-efficacy in the future.

We had some concerns that only the child choosing the book would have increased engagement, while the remainder of the group were unaffected. However, that did not seem to be the case, perhaps because children often like similar books or have group favourites, or perhaps just because the increased agency in itself makes the sessions more engaging. We would therefore encourage preschool settings to involve children in choosing the books for group storybook reading on a regular basis.

One of the concerns voiced by educators about the child-led storybook sessions is that children tended to pick a book that they knew well, rather than a new one, thus limiting their breadth of exposure. In fact, our analysis suggested that the child-led approach actually led to reading a greater breadth of books. Many educators in the planned condition planned to read the same book two or three days in a row, which was unusual in the child-led condition. It is possible that the additional workload of planning book-reading interactions inhibited the variety of books chosen. The benefits of repeating books versus choosing a wide variety is something that should be examined further, as described in the Future Research section.

Limitations and lessons learned

Power estimations were broadly accurate. The effect size was smaller than the MDES for the primary outcome measure, though it was larger than the MDES for the secondary outcome measure. LanguageScreen measures a wider range of language skills, including receptive and expressive word-level and sentence-level skills, while the Storybook Vocabulary measure focuses on expressive vocabulary only. It is possible that the child-led approach helped to build these broader level language skills rather than specifically expressive vocabulary. Alternatively, it is possible that this is due to higher levels

of sensitivity and reliability in the LanguageScreen measure, which has been more extensively piloted and standardised than the Storybook Vocabulary measure.

Setting recruitment was difficult, particularly for PVI settings, however, once recruited, we had less attrition than had been predicted based on previous early years trials. Settings that took part were clearly committed to the research and welcomed the opportunity to reflect on elements of their practice. We feel that further research working with PVI settings is vital and hope that more settings become open to taking part in research with time.

We saw excellent storybook reading practice in many varied settings and it was difficult to fully capture the range of good practice using the observations and survey. In addition, time constraints meant that we were not able to select settings with the best progress for case studies but instead took a cross section of different types of setting and condition. This allows us to reflect on different ways in which the sessions were implemented but it does not allow us to give guidance on the best approach to implementation.

The comparison of two different storybook reading conditions meant that we did not have a ‘business as usual’ control condition. It was not felt to be ecologically valid in this context. This has positive and negative effects. First, use of untreated control groups can over-estimate the size of an intervention effect, as educators are aware they are delivering the ‘test’ intervention and modify their behaviour to place more focus on the teaching subject in question than those in the control group. Alternatively, settings allocated to the control group may actively choose an alternative approach as a result of being allocated to the control group. Having two parallel conditions with equal emphasis on each avoids these potential issues.

Nonetheless, the study design does cause some complications in analysis and interpretation. It is important to highlight that we do not show that planned storybook reading is ineffective, by any means. On the contrary, we felt able to use an RCT without a no-treatment control condition precisely because it is so well established that interactive storybook reading is effective in multiple different contexts. We have merely indicated that including an element of child choice and agency is likely to lead to somewhat stronger language outcomes.

It is also important to highlight that many responses in the IPE indicated that a hybrid approach—including some planned sessions and some child-led sessions in the weekly routine—is perceived by educators to be most effective at promoting engagement and supporting breadth and depth of experience.

Future research and publications

Future research should consider how important variation or repetition in storybook choices is in driving outcomes. We anticipate that this is likely to be a complex issue, variable according to child characteristics.

During the scoping phase of this project, many educators spoke about the importance of child-led learning in preschool settings and about a skilled educator being able to respond to a child’s interest ‘in the moment’. In fact, this was the basis of RQ2 in the scoping phase. At that point, we found it difficult to conceptualise how a randomised controlled trial could be used to assess spontaneous interactions, however, we believe that this trial has potentially examined just these types of interactions, albeit indirectly. Further research could investigate the extent to which language improvements are driven by the books used in the sessions versus by the language interactions around those books.

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), pp. 431–442. doi: 10.1080/01621459.1995.10476535
- Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015) 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software*, 67 (1), pp. 1–48. doi: 10.18637/jss.v067.i01
- 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. Doi: 10.1111/1467-9817.12255
- 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. Doi: 10.2307/30035543
- Christensen, R. (2023) 'ordinal-Regression Models for Ordinal Data_', R package version 2023.12-4.1: <https://CRAN.R-project.org/package=ordinal>
- 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), pp. 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. Doi: 10.1111/cdev.13225
- Fox, J., Kleiberg, 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>
- Gaure, S. (2013) 'OLS with Multiple High Dimensional Category Variables', *Computational Statistics & Data Analysis*, 66: 8–1. Doi: [10.1016/j.csda.2013.03.024](https://doi.org/10.1016/j.csda.2013.03.024)
- 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. Doi: 10.1002/rrq.130
- Harrell, F. E. (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, New York: Springer-Verlag. <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>
- Holm, S. (1979). 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian Journal of Statistics*, 6 (2), pp. 65–70. <https://www.jstor.org/stable/4615733>
- 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*, 345, e5840. <https://doi.org/10.1136/bmj.e5840>
- Kraemer, H. C., Kiernan, M., Essex, M. and Kupfer, D. J. (2008) 'How and Why Criteria Defining Moderators and Mediators Differ Between the Baron & Kenny and Macarthur Approaches', *Health Psychology*, 27 (2S):S101-8. doi: 10.1037/0278-6133.27.2(Suppl.).S101. PMID: 18377151; PMCID: PMC3376898
- Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. (2017) 'lmerTest Package: Tests in Linear Mixed Effects Models', *Journal of Statistical Software*, 82 (13), pp. 1–26. doi: 10.18637/jss.v082.i13

- McGeown, S., Ricketts, J. and Shapiro, L. (2024) 'Love to Read: A Co-Designed Intervention to Motivate and Engage Child Readers', Nuffield Report. <https://www.nuffieldfoundation.org/project/love-to-read-co-designed-intervention-motivate-engage-child-readers>
- Nation, K., Dawson, N. J. and Hsiao, Y. (2022) 'Book Language and Its Implications for Children's Language, Literacy, and Development', *Current Directions in Psychological Science*, 31 (4), pp. 375–380. Doi: 10.1177/09637214221103264
- 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. DOI: <https://doi.org/10.1017/S0305000996003005>
- 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. Doi: <https://doi.org/10.1111/cdev.13730>
- Snow, G. (2020) 'blockrand: Randomization for Block Random Clinical Trials', R package version 1.5: <https://CRAN.R-project.org/package=blockrand>
- Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I. and Taggart, B. (2004) 'The Effective Provision of Pre-School Education (EPPE) Project, Technical Paper 12, *The Final Report: Effective Pre-School Education*'. <https://discovery.ucl.ac.uk/id/eprint/10005308/1/eppe12sylva2004effective.pdf>
- 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, pp. 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, pp. 1–17. Doi: <https://doi.org/10.1016/j.ecresq.2018.01.010>
- 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), pp. 681–694. [https://doi.org/10.1002/\(sici\)1097-0258\(19990330\)18:6<681::aid-sim71>3.0.co;2-r](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), pp. 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), pp. 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), pp. 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. Doi: <https://doi.org/10.1037/0012-1649.24.4.552>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... and Yutani, H. (2019) 'Welcome to the Tidyverse', *Journal of Open Source Software*, 4 (43), 1686. doi: 10.21105/joss.01686

Appendix A: Security classification of trial findings

OUTCOME: Storybook Vocabulary Measure

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	MDES	Attrition					
5	Randomised design	≤ 0.2	0-10%					
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%					
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%	3		Adjustment for threats to internal validity 0		3
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%					
0	No comparator	≥ 0.6	>50%					

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Low	Statistician conducting randomisation was blinded to identify of settings; no baseline imbalances exceeding .05 SD.
Threat 2: Concurrent Interventions	No information available	No reported evidence of this.
Threat 3: Experimental effects	Low	There is no 'control' group as such – two active arms are being compared to each other. Participants in both arms talked about using a hybrid approach in the future – hard to tell if this had started to creep in during the trial period.
Threat 4: Implementation fidelity	Moderate	There appears to be moderate compliance and fidelity in the approaches according to endline surveys, observations. Given low returned rates of session records, there is some uncertainty about the extent to which pupils were consistently exposed to the approach.

Threat 5: Missing Data	Moderate	Missing data is high (30%). Analysis accounting for missing data is comparable with the primary analysis result but with a difference in the setting type parameter value.
Threat 6: Measurement of Outcomes	Low/Moderate	Bespoke outcome measure used in the trial was validated pre-trial and appropriate for the study. Apart from 22 out of 956 children where teachers administered the baseline assessment, the measure was administered by independent researchers blind to allocation.
Threat 7: Selective reporting	Low	Trial is pre-registered and analysis conducted and reported in line with published study plan.

Initial padlocks: 3 padlocks. The design is a two-arm cluster randomised controlled trial with setting-level randomisation. The MDES at randomisation stage was .3. However, there was 29.8% child attrition from randomisation to analysis.

Reason for adjustment for threats to validity: 0

Final padlocks: 3 padlocks

Appendix B: Effect size estimation

Appendix table 2: Effect size estimation

Outcome			Child-led		Planned			
	Un-adjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome (SD^2)	n (missing)	Variance of outcome (SD^2)	Pooled variance (SD^2)	Population variance (if applicable)
Post-test <u>Storybook</u> vocabulary test score	1.4	1.22	412 (14%)	63.8	390 (19%)	63.68	64.16	-
Post-test <u>LanguageScreen</u> – Total Score	3.8	3.5	341 (29%)	207.36	330 (31%)	158.76	187.69	-
Post-test <u>LanguageScreen</u> -Expressive Vocabulary	0.3	2.9	341 (29%)	176.89	330 (31%)	158.76	171.61	-
Post-test <u>LanguageScreen</u> – Listening comprehension	3.0	3.4	341 (29%)	171.61	330 (31%)	169	174.2	-
Post-test <u>LanguageScreen</u> – Receptive vocabulary	3.2	2.9	341 (29%)	210.25	330 (31%)	193.21	204.49	-
Post-test <u>LanguageScreen</u> – Sentence Completion	2.0	2.0	341 (29%)	228.01	341 (29%)	158.76	196	-
Post-Test <u>ATLA</u> measure	0.12	0.25	380 (20.1%)	22.18	341 (29%)	22.56	22.37	-

Further appendices

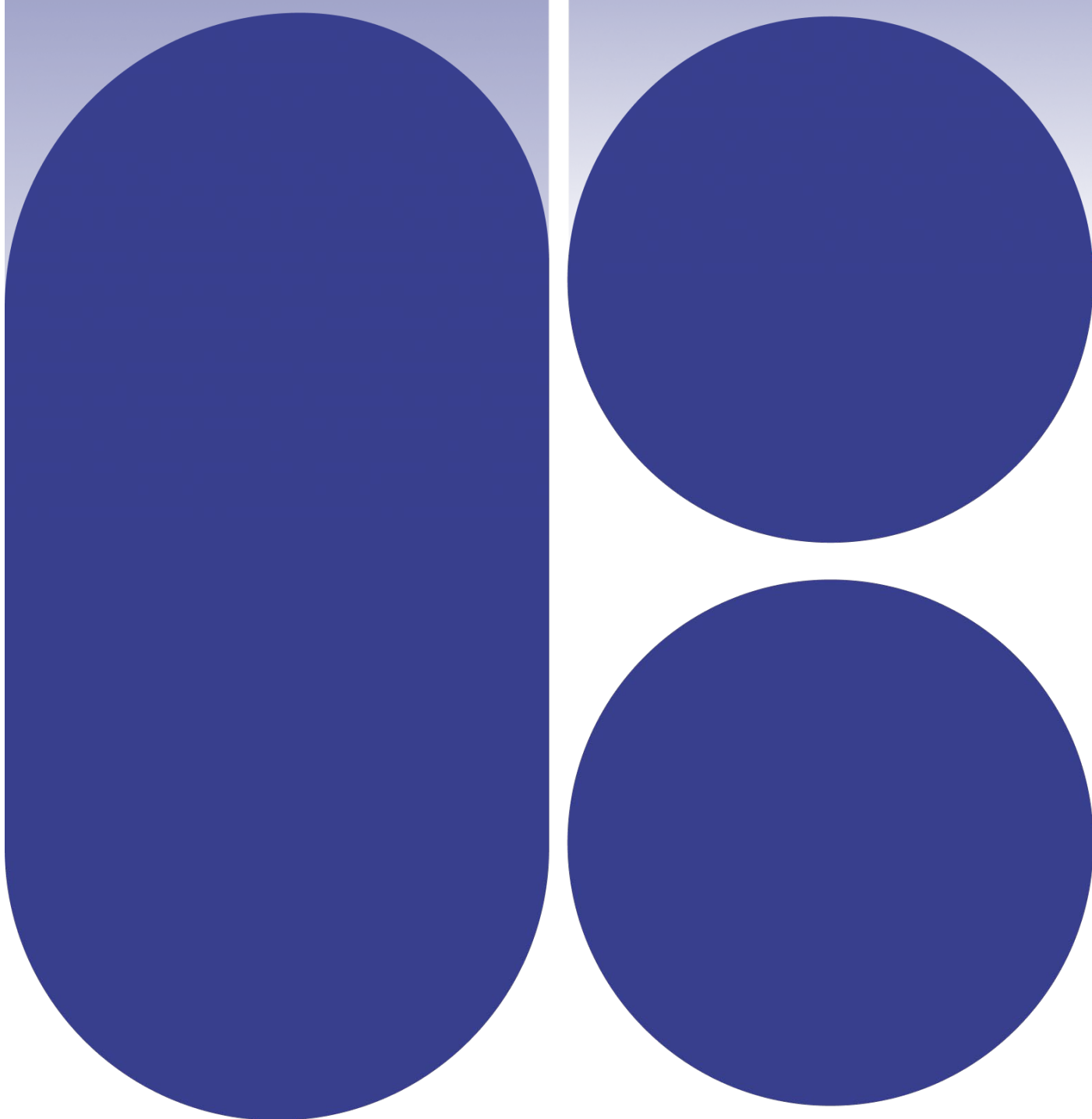
Appendices C – M can be found published in a separate document named '*further appendices*' published on the project page.

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at <https://educationendowmentfoundation.org.uk>



Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London

<https://educationendowmentfoundation.org.uk>

 [@EducEndowFoundn](https://twitter.com/EducEndowFoundn)

 Facebook.com/EducEndowFoundn