Concept Cat: A two-armed cluster randomised controlled trial

Evaluation report

July 2025

Miguel Subosa, Fin Oades, Erin Dysart, Louise Tracey, Anna Brian, and Elena Rosa Speciani

Co-funded by:

Bright Futures North West Early Years Stronger Practice Hub

HEART – Midlands Early Years Stronger Practice Hub

Liverpool City Region and Beyond Early Years Stronger Practice Hub

The Education Endowment Foundation (EEF) 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

Department for Education

What Works Network

Contents

About the evaluator **3**

Executive summary **4**

Introduction **6**

Methods **15**

Impact evaluation **39**

Implementation and Process Evaluation **53**

Conclusion **92**

References **96**

Appendix A: EEF cost rating **99**

Appendix B: Security classification of trial findings **100**

Appendix C: Effect size estimation **102**

Appendix D: Residual plots from analysis models **107**

Appendix E: Missing data logistic regression output **116**

Appendix F: Subgroup interaction model results **118**

Further appendices **120**

About the evaluator

The project was independently evaluated by a team from RAND Europe and the University of Leeds.

Contact details:

Elena Rosa Speciani

RAND Europe,

Eastbrook House,

Shaftesbury Road,

Cambridge,

CB2 8DR

Tel: +44 1223 353 329

Email: erspeciani@randeurpope.org

Acknowledgements

We would like to thank a number of key people without whom the project would not have been successful. First, we would like to thank members of the delivery team for their collaboration and support throughout the project: Anna Branagan; Marie Gascoigne; Stephen Parsons; Victoria Riley; Michele Seidler; and Bibiana Wigley. Second, we thank Fin Oades and Lydia Lymperis (formerly RAND Europe) who supported the project in the beginning. Finally, many thanks to the Education Endowment Foundation team: Faizaan Sami and Thomas Mackay, for all their hard work.

Executive summary

The project

Concept Cat aims to improve early conceptual vocabulary in three- to four-year-olds. It teaches early verbal concepts linked to the mathematics and science curriculum, aiming to enhance Key Stage 1 outcomes. Spanning an academic year, it includes four components: whole-class introduction; meaningful play; parent-child tasks; and whole-class review. Additional support is provided to 'focus children' with below-average language skills. Lead early years practitioners (EYPs) attend a three-hour remote training; other early years (EY) staff attend a one-hour remote session. Settings also receive visits from Concept Cat coaches during the delivery to support effective implementation.

The project was a two-armed waitlisted cluster randomised controlled trial. A total of 1,040 children from 89 settings participated, with settings randomised to receive Concept Cat (treatment) or business as usual (control). 902 children were assessed at the end of the programme. The trial's primary outcome was early conceptual vocabulary. An implementation and process evaluation (IPE) incorporated practitioner and parent surveys, training observations, setting visits, and analysis of implementation monitoring data. Intervention delivery of the trial took place from September 2023 to June 2024.

As part of the Department for Education's Early Years Recovery Programme, the Education Endowment Foundation (EEF) is working with Stronger Practice Hubs (SPHs) across England to fund EY settings' access to evidence-informed programmes and study the programme's influence on practice and children's outcomes. This initiative aims to support education recovery following the pandemic, while also developing our understanding of effective professional development in the EY. The EEF has worked with HEART – Midlands Early Years SPH, Bright Futures North West Early Years SPH, and Liverpool City Region and Beyond Early Years SPH to fund settings' access to Concept Cat and evaluate the programme through an efficacy trial.

Table 1: Key conclusions

Key conclusions	
1.	Children in Concept Cat settings made, on average, two months' additional progress in understanding conceptual vocabulary, compared to children in other settings. This result has a moderate to high security rating.
2.	Among children with Early Years Pupil Premium (EYPP), those in Concept Cat settings made three months' additional progress in conceptual vocabulary compared to those in other settings. These results may have lower security than the overall findings because of the smaller number of EYPP children.
3.	Children in Concept Cat settings demonstrated, on average, two months' additional progress in their early numeracy development, compared to children in other settings.
4.	Compliance and fidelity to the programme design were moderate to high across settings. However, it was found that settings were not delivering additional activities for focus children, as set out in the design, indicating low fidelity in this regard.
5.	Concept Cat appears to have facilitated parents' involvement in the setting and understanding of their children's learning and development.

EEF security rating

These findings have a moderate to high security rating. This was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a number of schools. The trial was a well-designed and well-powered two-armed waitlisted cluster randomised controlled trial. However, the following factors reduced the security of the trial. Around 13.3% of the children who started the trial were not included in the final analysis because they were absent when testing data was collected. There were some differences in prior attainment between the children in settings, which delivered Concept Cat and those in the comparison settings. Given the presence of the ceiling effects with the primary outcome measure, the evaluation may underestimate the size of the impact on the pupils in the trial.

Additional findings

Children in Concept Cat settings made, on average two additional months' progress than those in the control group equivalent. This is our best estimate of impact, which has a moderate to high security rating. As with any study, there is always some uncertainty around the result: the possible impact of this programme ranges from one month's additional progress to four months' additional progress. Children in Concept Cat settings showed one additional month's progress on understanding complex sentences compared to the control group, suggesting the intervention also supports wider conceptual understanding.

Children receiving EYPP in the Concept Cat settings made three additional months' progress on average, compared to EYPP children in settings that did not deliver the Concept Cat programme. This possible impact ranges from zero month's progress and to a positive effect of up to five months' additional progress. These results may have a lower security than the overall findings because of the smaller number of children.

The compliance analysis suggests that delivering the intervention as intended was associated with up to four months' of additional progress in early conceptual vocabulary. However, large numbers of settings not sharing child attendance data at the beginning of the trial led to a high degree of missing information indicating low robustness in these estimates.

Overall, this evaluation provided strong support for the logic model. The impact evaluation has shown that Concept Cat has a positive impact on the intended outcomes of conceptual vocabulary and numeracy. The IPE further supports the notion of programme effectiveness, with respect to both children and their families. Moreover, the fact that all surveyed EYPs said they were going to continue implementing the programme after the trial, and that children clearly enjoy Concept Cat and this facilitated home implementation of the programme, shows the desirability of the programme overall.

Cost

The average cost of delivering Concept Cat was £1,206.80 per setting or £30.94 per pupil per year over three years. This assumes that all children aged three to four in a setting would participate in the intervention each year. The cost estimates include direct training costs, staff cover costs for training, physical resources, and Concept Cat coach visits.

Impact

Table 2: Summary of impact on Concept Cat

Outcome / group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No. of children	P-value	EEF cost rating
Conceptual vocabulary, all children	0.18 (0.05 – 0.30)	+2 months' progress	🔒🔒🔒🔒🔒	902	0.01	£ £ £ £ £
Conceptual vocabulary, EYPP-eligible children	0.25 (0.02 - 0.51)	+3 months' progress	N/A	171	0.03	N/A

N/A=not applicable.

Introduction

Background

There is evidence to suggest a link between conceptual language development and maths skills. Specifically, children who are exposed to more conceptual language in their early years (EY) tend to have stronger maths skills later in life (Lin *et al.*, 2021). One study found that children who were exposed to more maths-related language in their homes and preschools tended to have better maths skills in kindergarten, compared to their peers who had less exposure to maths-related language (LeFevre *et al.,* 2010). Another study found that children who were exposed to more spatial language (which is a type of conceptual language) in their preschool years had, on average, better spatial reasoning skills in elementary school (Verdine *et al.,* 2014).

Recent research further supports these findings. Purpura *et al.* (2017) highlight the importance of early numeracy and language skills as predictors of later mathematical achievement. They found that children's early numeracy skills, often supported by language development, are crucial for later success in mathematics. Additionally, Clements *et al.* (2020) emphasise the role of engaging children in mathematical discourse to enhance their conceptual understanding and problem-solving skills.

Evidence also suggests that children from disadvantaged backgrounds tend to have lower levels of conceptual vocabulary than their more advantaged peers, which in turn can have negative effects on their language, cognitive, and academic development. For example, Hart and Risley's (1995) landmark study found that, broadly speaking, children from low-income families heard significantly fewer words than children from higher-income families, and that the words heard by the former tended to be more limited in scope and complexity. This disparity in language exposure was found to have long-lasting effects on children's language and cognitive development, with children from low-income families having smaller vocabularies and weaker language skills, on average (Hart and Risley, 1995).

Further supporting these findings, Hoff (2013) confirms that socio-economic status can influence language development, with children from higher-income families typically experiencing richer language environments. This underscores the importance of addressing language disparities. More recent research suggests that interventions targeting language exposure can mitigate some of the negative effects associated with socio-economic disadvantages, promoting better language and cognitive outcomes for children from low-income families (Golinkoff *et al.* 2019).

Other studies have found similar patterns of disparities in language development between children from different socio-economic backgrounds. For example, a study by Fernald *et al.* (2013) found that children from low-income families had lower levels of exposure to both conversational and conceptual language than children from higher-income families, and that this disparity was evident as early as 18 months of age. In comparison, existing research strongly suggests that early language interventions have the potential to a major impact on shaping language skills (Fricke *et al.,* 2012; Scarborough, 2009) with a recent meta-analysis of EY language and communication-focused programmes demonstrating potential for high positive impact on children (EEF, 2023a). Therefore, there is a compelling case to develop and evaluate interventions that support conceptual language development, particularly for disadvantaged children. In the UK, however, few programmes have been evaluated with sufficiently robust methodologies, rendering robust evaluations of EY language development programmes in the UK an invaluable addition to the evidence base.

This evaluation forms part of the Department for Education (DfE) Stronger Practice Hubs (SPH) policy, which are designed to build evidence-informed practice in EY. The SPHs, launched in November 2022, form part of the DfE's Early Years Education Recovery Programme. A key aim of SPHs is to address the impact of the pandemic on young children by supporting EY settings to build local networks and share evidence-informed practices to ultimately improve the quality of education and care. The Education Endowment Foundation (EEF) is supporting the launch of the SPHs to build evidence around EY approaches. The EEF's work includes:

1. Selecting programmes from its open funding rounds to be part of the list of programmes that SPHs are able to make available as funded support in their region.
2. Providing funding, in addition to funding from SPHs, for programme providers to deliver their programme to EY settings as part of a research project.

3. Providing funding for independent evaluators to implement research of programmes on the list available to SPHs.

This evaluation was designed to help build evidence around what works in EY and was conducted on settings from three SPHs in Birmingham, Trafford, and Everton. Building this evidence is an important contribution, given the knowledge gap regarding language interventions in EY settings—and more specifically, language interventions in which parent-child interactions are a key component—that was identified by Law *et al.* (2017).

The efficacy trial for Concept Cat was carried out as a two-arm, waitlisted, cluster randomised controlled trial, with a 50:50 allocation of 89 EY settings to treatment and control groups. Baseline testing took place from September 2023 to October 2023, with randomisation in September 2023. The intervention itself was delivered from November 2023 until June 2024, corresponding to a total duration of 30 weeks.

Intervention

Concept Cat introduces children to concepts like ‘first’, ‘wide’, and ‘empty’ through a structured and engaging approach. It begins with explicit vocabulary instruction, followed by implicit learning through play-based activities. The process involves staff performing a scripted story with a toy cat, and minor adjustments are made to the free flow play environment to provide children with repeated exposure to the new vocabulary. For example, if the target word for the week is ‘empty’, sand and water trays would be available, and staff would incorporate the word ‘empty’ during playtime. Families are involved through simple home activities.

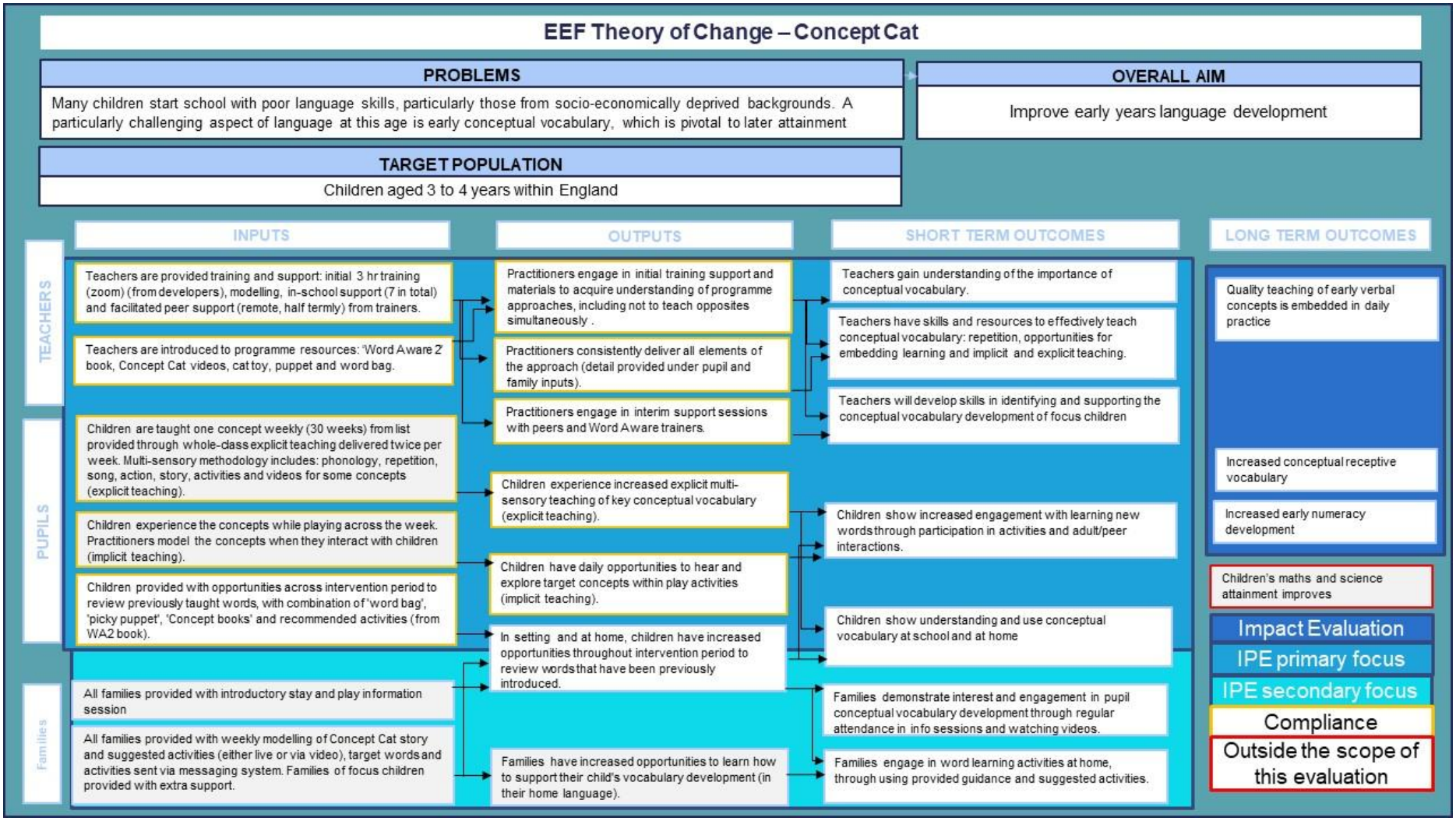
This sequence is designed to be accessible to a diverse range of children, including those with limited language skills. By focusing on one word per week, it allows for the development of a deep understanding. Instead of general vocabulary, Concept Cat specifically teaches early verbal concepts, such as ‘before’, ‘early’, and ‘through’, which are integral to the maths and science curriculum, ultimately aiming to enhance maths and science achievement at Key Stage 1.

Concept Cat was conceived by the founders of ‘Thinking Talking’ Stephen Parsons and Anna Branagan, and its teaching methodology is outlined in their book, ‘Word Aware 2’ (Parsons and Branagan, 2016). The programme is grounded in the STAR (Select, Teach, Activate, Review) approach to conceptual learning (Blachowicz and Fisher, 2015). Moreover, it seeks to interweave classroom practice with child-parent interactions, an area for which there is only a limited number of relevant previous evaluations.

Concept Cat is currently delivered, with approximately 300 EY practitioners (EYPs) trained in the Concept Cat teaching methodology each year. A quasi-experimental pilot impact evaluation of the Concept Cat programme showed promise with effect sizes of 0.42. However, there were some distinct limitations in the study, namely the lack of randomisation in the design and the limited sample size, with data having been collected from only two preschools. Furthermore, the programme has been adapted since the pilot (Hopkins *et al.,* 2022). In the pilot, Speech and Language Therapy (SALT) students supported staff in EY settings to deliver the intervention but in this study Concept Cat coaches supported setting staff to deliver the intervention. Furthermore, the training in this study was much more extensive compared to the pilot.

Concept Cat is a whole-class intervention, targeting children aged between three and four, that seeks to facilitate the acquisition of key early verbal concepts. In turn, the acquisition of these early verbal concepts would support the attainment of learning competencies laid out in the Key Stage 1 core science and mathematics curricula. The Concept Cat methodology offers an alternative to the generally unstructured and less explicit way these core concepts are taught in standard practice. Moreover, the Concept Cat approach offers a combination of explicit and implicit teaching of concept words, which is embedded in daily practice. The programme is delivered over a full academic year (i.e. approximately 30 weeks), where each week coincides with the introduction of a new key word. Further details of the programme can be seen in the theory of change (Figure 1) below.

Figure 1: Theory of change for Concept Cat, developed with facilitation from the EEF theory of change workshop



The programme incorporates four core components in its implementation. These are a whole-class introduction to the word, meaningful play sessions, parent-child tasks, and a whole-class review.

The '**whole-class introduction**' establishes the key verbal concept for a given week, forming the 'explicit teaching' aspect of the programme, delivered through a multi-sensory methodology. The specific word is introduced by the EYP and is accompanied by both a unique visual symbol and a physical gesture. Phonology and repetition are also used, supporting the children's memory of the word's phonetic characteristics.

'**Meaningful play sessions**' provide an opportunity for more 'implicit teaching' to reinforce the learning that has occurred during the whole-class introductory phase, whereby the children come across the word within structured play situations. Children's families are also introduced to the word, and '**home-based tasks**' are suggested, such as 'on the way home, look out for roads that are narrow and not narrow. Look out for any narrow paths'. There is also a '**whole-class review**', which incorporates a range of activities using Concept Cat props such as the word bag and Picky Puppet, which will allow children to encounter the word during both the word's focus week, as well as subsequent weeks, concretising the knowledge the children have gained throughout the intervention's entire delivery period.

Children with the most limited language proficiency are typically identified by comparing their capabilities against the 'Early Learning Goals'; they are provided with additional support through extra modelling of the Concept Cat story and target words. For the purposes of this evaluation, children requiring additional support are referred to as 'focus children'. Guidance for identifying focus children was given in written form to settings and included questions to ask families and a flow chart to aid decision-making. This was then explained in the training session and followed up with discussion between the Concept Cat coach and the nursery lead practitioner. Setting staff made judgements based on their observations and in consultation with the Concept Cat coaches. No formal assessments were conducted.

The following criteria was provided (see Appendix P):

- aged three- to four-years old;
- able to sit and respond to an adult-led task for a few moments;[1]
- uses fewer words and shorter sentences than other children of the same age;
- not a reluctant speaker at nursery who speaks fluently at home;[2] and
- if the child speaks English as an Additional Language (EAL) then they must also have delayed language development in the home language(s).

The overall structure of the intervention is based on the STAR methodology, where a concept is 'selected', 'taught', then implicitly 'activated' through play and home-based activities, and subsequently 'reviewed' to encourage a deeper understanding of its meaning.

Delivery personnel and training

The programme is delivered by EYPs within the selected EY settings, with lead practitioners receiving a three-hour remote training session and other EY staff receiving a one-hour remote training session provided by Better Communication CIC. Each setting will also receive seven in-setting support visits from Concept Cat coaches. Coaches are speech and language therapists and teachers recruited as associates of Better Communication CIC and trained by the project leads, with full resources, including an intervention handbook, and regular group and individual supervision provided. During the second visit, setting staff under coach supervision carry out an initial assessment of six children, selected by practitioners (based on a range of language abilities) so the correct level of concept words is selected for

[1] This was informal guidance to enable EYPs to make a judgement at the start of the year, when most children were new to the setting and little was known about them. It is purposefully 'light touch' and not requiring further assessment, so as not to delay the commencement of the teaching.
[2] This criteria is designed to identify those who are most in need of additional support. Concept Cat is an inclusive approach that benefits a wide range of learners. Those with speech, language, and communication needs may need extra repetitions/exposures to make full use of the approach. Most other learners (including shy/selectively mute) will access the main teaching without additional support.

the setting. Coaches also model the 'Teach' element of STAR in the programme. Other sessions involve further modelling, observation, and discussion of 'Teach, Activate, and Review' elements of STAR and give further support on implementation of the programme. The sessions provide an opportunity for coaches to ensure practitioners are implementing the programme as it is intended and in a way that works for the setting. In addition to the coach support sessions, lead practitioners are also encouraged to attend six group support sessions across the academic year, allowing practitioners to share experiences and tips to improve delivery.

Duration and frequency

Delivery of the intervention commenced in early October, running for a total of 30 active weeks. As discussed above, the intervention is delivered using a weekly planner depending on the staffing situation of each setting (i.e. five days or three days) structure, where a new word is introduced at the start of each week for the programme's duration. The explicit teaching aspect occurs twice a week for approximately ten minutes, while the other implicit methodologies may be interwoven into additional less prescriptive activities throughout the week.

Materials

The key materials required for delivery are:

- the 'Word Aware 2' book (Parsons and Branagan, 2016) provides the overall teaching structure for each of the words;
- Concept Cat soft toy, word bag, and 'Picky Puppet' are used for introducing and reviewing the taught words;
- Lift Lessons animated videos of approximately 50% of the introductory stories; and
- in addition, each setting is provided with the printed materials for the word of the week for all of Level 1 and Level 2 words with the option to print locally any Level 3 words needed based on the Concept Cat Screen and as identified by the Concept Cat coach for each setting.

EY settings are required to find a few resources to incorporate into the introductory Concept Cat stories. These are simple everyday items such as socks or boxes or toys that settings are highly likely to already have.

Evaluation objectives

This evaluation had three main research questions, with each pertaining to analysis of the impact of the Concept Cat intervention on a distinct outcome. Within each of research questions 1 and 3, there were also a further three subsidiary research questions pertaining to subgroups of interest for this trial.

The primary research question of this project was:

> 1. What is the difference in early conceptual vocabulary development, measured by the 'Basic Concepts' subtest of the Clinical Evaluation of Language Fundamentals (CELF) Preschool-2 UK, of children in settings receiving Concept Cat in comparison to those children in control settings receiving business as usual?

The following sub-questions of primary research question 1 were also explored:

> 1a. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of Early Years Pupil Premium (EYPP)/Free Early Education Entitlement (FEEE)-eligible children, compared to non-EYPP/FEEE-eligible children?
>
> 1b. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of children with EAL, compared to non-EAL children?
>
> 1c. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of children with Special Educational Needs and Disabilities (SEND), compared to non-SEND children?

The secondary research questions of this project were:

> 2. What is the difference in early conceptual vocabulary development, measured by the 'Concepts and Following Directions' subtest of the CELF Preschool-2 UK, of children in settings receiving Concept Cat in comparison to those children in control settings receiving business as usual?
>
> 3. What is the difference in early numeracy development measured by the Early Numeracy Assessment (ENA) of the EY Toolbox of children in settings receiving Concept Cat in comparison to those children in control settings receiving business as usual?

The following sub-questions of the secondary research question 3 were also explored:

> 3a. What is the impact of the Concept Cat teaching methodology on the early numeracy development of EYPP/FEEE-eligible children, compared to non-EYPP/FEEE-eligible children?
>
> 3b. What is the impact of the Concept Cat teaching methodology on the early numeracy development of EAL children, compared to non-EAL children?
>
> 3c. What is the impact of the Concept Cat teaching methodology on the early numeracy development of SEND children, compared to non-SEND children?

The implementation and process evaluation (IPE) sought to answer the following questions:

> IPERQ1. How closely does the Concept Cat programme, as implemented in settings, follow the intended model (implementation fidelity), as outlined in the TIDieR (Template for Intervention Description and Replication) framework including extended implementation for focus children? What are the barriers and facilitators to implementation and how do these differ, if at all, between setting type (Private, Voluntary, and Independent [PVI]/Maintained schools)?
>
> IPERQ2. What, if any, adaptations have been made to the programme during implementation? Why were they made? What do they look like?
>
> IPERQ3. What is the nature of business as usual with regard to vocabulary instruction? How does this differ between control and intervention settings? What are the similarities/differences between setting type (PVI/Maintained)? How does programme delivery differ from business as usual?
>
> IPERQ4. Have practitioners attended mandatory training? To what extent have training and resources supported practitioners' ability to effectively teach Concept Cat? What is the quality of delivery (i.e. how well are different components of the intervention delivered; Education Endowment Foundation, 2004 p. 6)?
>
> IPERQ5a. To what extent have practitioners developed their knowledge about conceptual vocabulary and skills in identifying and supporting the conceptual vocabulary development of children with higher language needs (i.e. those identified as focus children)?
>
> IPERQ5b. To what extent are practitioners motivated to implement, and continue to implement, Concept Cat? Is this motivation different across setting type (PVI/ Maintained) and if so, why?
>
> IPERQ6. To what extent have settings engaged families with the programme and in what ways? Are there differences between setting type (PVI/Maintained) in the ways settings have engaged with families? How is this linked, if at all, to child outcomes?
>
> IPERQ7. What are the barriers and facilitators for families in home implementation of the programme, particularly for focus children, disadvantaged children, and those who are EAL? What, if any, are the wider impacts on the home learning environment (HLE)?
>
> IPERQ8. To what extent does Concept Cat result in positive or negative unintended consequences for settings, practitioners, children, families, and the HLE?

The Trial Protocol and Statistical Analysis Plan (SAP) are both available on the trial's page on the EEF website.

Ethics and trial registration

This evaluation is registered on the International Standard Randomised Controlled Trial Number (ISRCTN) Registry. The number is ISRCTN51286286.

The procedures described in this protocol are in line with the ethical standards of RAND Europe and the University of York.[3] They have been reviewed and approved by both RAND US' Human Subjects Protection Committee and the University of York's Education Ethics Committee. These ethical approval documents are found in Appendices I and J, respectively.

Consent to participate in the intervention and evaluation was obtained from parents or legal guardians, who act as decision-makers for individual children. In the interest of informed consent, settings provided parents and legal guardians with information sheets and withdrawal forms. If parents or legal guardians decided to withdraw their child from the intervention, evaluation, or both, they could do so by returning the withdrawal form. Parents or legal guardians were able to withdraw their children at any time throughout the intervention and evaluation. If a child's participation was withdrawn, the delivery and evaluation teams did not collect data from the child and any data previously collected on the child was deleted.

The University of York provided and collected consent forms from setting staff who participate in observations, interviews, or focus group discussions for the IPE. In addition, surveys with parents and practitioners included a privacy notice indicating to respondents that their participation was voluntary and that they could choose to withdraw at any time without penalty.

No member of the evaluation team has any conflict of interest with respect to the intervention or evaluation.

Data protection

Several teams are involved in controlling and processing data. RAND Europe and University of Leeds acted as joint data controllers, with Better Communication CIC and Elklan Training Ltd (from hereinafter Elkan) acting as data processors. Further details on this are outlined in the data flow diagram in Appendix K.

As part of the evaluation, Elklan collected information from settings about all the children that take part in Concept Cat. RAND Europe also asked Elklan to collect data on assessed child outcomes. Settings and Elklan provided this information using an Excel data collection form provided by the evaluation team. This form was shared via secure file transfer (i.e. Egress). The University of Leeds asked Better Communication CIC (the delivery team) to collect information from settings about key staff. Settings provided information on key staff through an Excel data collection form. This form was shared via secure file transfer. The University of Leeds used the data provided by settings on key staff to invite them to take part in a short survey and, where applicable, an interview. RAND Europe also used this data to contact settings to collect data on children's EYPP status.

At the end of the study, RAND Europe will submit the data in pseudo-anonymised format to the Office for National Statistics (ONS) Secure Research Service (SRS) for archiving in the EEF data archive. This data will only be identifiable to the DfE and may be matched to the National Pupil Database and other administrative data in subsequent research. The EEF and DfE will act as data controllers for the archive, along with contractors appointed to manage the archive.

The legal basis for RAND Europe is legitimate interests, as detailed in Article 6(1)(f) of the UK General Data Protection Regulation (GDPR 2016a). The legal basis for processing children's special category data[4] is because it is necessary for archiving purposes in the public interest, scientific, or historical research purposes, as detailed in Article 9(2)(j) of the UK GDPR (2016a). To ensure that all processing is fair and lawful, RAND Europe also completed a Legitimate Interest

[3] At the start of the trial Louise Tracey and Erin Dysart were based at the University of York. They subsequently moved to the University of Leeds and continued to work on the project from their new university. This is why the ethical clearance came from the University of York.
[4] 'Special category' data is personal data that needs more protection because it is sensitive, for example, health or ethnicity data.

Assessment and a Data Protection Impact Assessment. RAND Europe processed only what was required to meet these legal bases and has ensured security and safeguards were in place to protect the information.

The legal basis for the University of Leeds, the EEF, and DfE is where it is necessary for the performance of a task carried out in the public interest as set out in Article 6(1)(e) of the UK GDPR. The specific legislation, which allows this is Section 10 of the Education Act 1996. The legal basis for processing special category data is for reasons of substantial public interest as detailed in Article 9(2)(g) of the UK GDPR (GDPR, 2016b).

The evaluation team take information security extremely seriously and all team members have appropriate technical and organisational measures to protect personal data and special category data. Access to information is restricted on a need-to-know basis and security arrangements are regularly reviewed to ensure their continued suitability. The evaluation team collect and store all personal and special category data in accordance with the Data Protection Act (2018) and UK GDPR requirements. No personal information collected as part of this study was transferred outside of the European Economic Area.

All individually identifiable data held by RAND Europe will be destroyed one year after the end of the study (i.e. 2026). All individually identifiable data held by the University of Leeds will be destroyed five years after the end of the study (2030). Data in the EEF's archive in the ONS SRS will include data only individually identifiable to the DfE, the government department responsible for children's services and education, and is kept indefinitely for the purposes of future research. Anonymous data will be kept indefinitely by the University of Leeds.

These aspects were detailed in documents provided for all participants and/or parents/carers of children in the study, in the Memorandum of Understanding (MoU), information sheets, withdrawal forms, and privacy notices (see Appendices K, L, and M).

Project team

Table 3: Members of the delivery team

Delivery team			
Name	Institution	Role	Responsibilities
Marie Gascoigne	Better Communication CIC	Director	• Organise Data Sharing Agreements with settings • Conduct training and support activities for Concept Cat implementation in settings, including the delivery of one-on-one coaching sessions • Recruitment of settings and ongoing liaison • Recruitment of Concept Cat coaches and organisational support
Stephen Parsons		Associate, Concept Cat subject expert	
Anna Branagan		Associate, Concept Cat subject expert	
Victoria Riley		Associate, deputy for Marie Gascoigne	
Bibiana Wigley		Associate, lead for setting recruitment and liaison	
Michele Seidler		Associate, operational resources and people	

Table 4: Members of the evaluation team

Evaluation team			
Name	Institution	Role	Responsibilities
Elena Rosa Speciani	RAND Europe	Project lead	• Provide oversight and direction regarding evaluation design and methodology
Miguel Subosa	RAND Europe	Project manager	• Ensure timely delivery of evaluation activities and outputs, including coordination with the testing partner • Manage the drafting of documents required for data protection, ethical approval, and trial registration, including the study protocol • Manage the development of the final evaluation report and all interim outputs
James Merewood	RAND Europe	Analyst (statistician / economist)	• Develop the SAP • Conduct randomisation of settings to the treatment and control group • Oversee the statistical analysis of CELF Preschool-2 UK and ENA data for the impact evaluation
Fin Oades	RAND Europe	Analyst	• Develop data collection instruments for the impact evaluation • Conduct the statistical analysis including: cleaning up datasets; writing code for statistical analysis; and running statistical tests • Writing the final evaluation report
Louise Tracey	University of Leeds (formerly, University of York)	IPE project lead	• Conduct baseline and endline surveys with setting staff • Conduct setting visits to a selection of settings during the study • Develop data collection instruments for the IPE • Collect data required for the IPE • Analyse data collected for the IPE
Erin Dysart	University of Leeds (formerly, University of York)	IPE project manager	
Alex Hall	Elklan	Testing partner	• Collect data on participating children from settings • Coordinate administration of the CELF Preschool-2 UK at baseline and endline • Coordinate administration of the ENA at endline • Transmit results of baseline and endline testing to RAND Europe

Methods

Trial design

Table 5: Trial design

Trial design, including number of arms		Two-armed waitlisted, cluster randomised controlled trial
Unit of randomisation		EY settings
Stratification variable(s) (if applicable)		Setting type: PVI vs school-based settings Region: Northern West Midlands; Southern West Midlands; Trafford; and Everton
Primary outcome	Variable	Early conceptual vocabulary
	Measure (instrument, scale, source)	'Basic Concepts' subtest from CELF Preschool-2 UK
Secondary outcome(s)	Variable(s)	Early conceptual vocabulary Early numeracy
	Measure(s) (instrument, scale, source)	'Concepts and Following Directions' subtest from CELF Preschool-2 UK EY Toolbox Early Numeracy
Baseline for primary outcome	Variable	Early conceptual vocabulary
	Measure (instrument, scale, source)	'Basic Concepts' subtest from CELF Preschool-2 UK
Baseline for secondary outcome(s)	Variable	Early conceptual vocabulary
	Measure (instrument, scale, source)	'Concepts and Following Directions' subtest from CELF Preschool-2 UK

As detailed in Table 5 above, the trial was designed as a two-armed, waitlisted, cluster randomised controlled trial, which primarily assesses the impact of the Concept Cat teaching methodology on early conceptual vocabulary development among children aged three to four in EY education.

Given that the intervention has a whole-class focus, randomisation occurred at the setting level, with each setting being allocated to either a group that receives the Concept Cat intervention (treatment group), or a group that receives business as usual (control group). Randomisation was stratified according to region so that each region was proportionately represented across both trial arms, while also ensuring that the delivery team had an appropriate number of settings per trainer in each region.

The four regional areas used for stratification were: Northern West Midlands; Southern West Midlands; Trafford; and Everton. All settings situated in the West Midlands are part of the 'HEART – Midlands Early Years' SPH, all settings situated in Trafford are part of the 'Bright Futures North West Early Years' SPH, and all settings situated in Everton are part of the 'Liverpool City Region and Beyond Early Years' SPH. Stratification was also based on setting type (i.e. PVI or school-based setting) to ensure a similar representation of each type of setting in each region. Having a balance of both setting types in the treatment and control groups ensured that findings from the trial were applicable to all setting types.

The primary outcome being measured was early conceptual vocabulary, as operationalised by the 'Basic Concepts' subtest from CELF Preschool-2 UK. This trial also sought to evaluate the impact of receiving the intervention on two secondary outcomes: an alternative measure of early conceptual vocabulary, operationalised by 'Concepts and Following Directions' subtest from CELF Preschool-2 UK; and early numeracy, operationalised by the EY Toolbox ENA. For the primary outcome and early numeracy secondary outcome, scores from the baseline 'Basic Concepts' subtest from CELF Preschool-2 UK were used as the baseline measure, while for the secondary measure of early conceptual vocabulary, scores from the baseline 'Concepts and Following Directions' subtest from CELF Preschool-2 UK were used as a baseline measure.

Participant selection

Settings

As detailed in the trial protocol, settings were recruited into the study with support from Better Communication CIC, with selection of settings from three EY SPHs based on input from the EEF and the DfE[5]:

- West Midlands: 'HEART – Midlands Early Years' SPH;
- Trafford: 'Bright Futures North West Early Years' SPH; and
- Everton: 'Liverpool City Region and Beyond Early Years' SPH.

These three EY SPHs span across three regional areas in England: the West Midlands; Trafford; and Everton. However, the stratification process used for randomisation split settings in the West Midlands into north and south.

Settings were eligible to apply to participate in the trial if they met the following criteria:

- settings expected to have a minimum of 15 children aged three to four (in Foundation 1) enrolled to attend for at least 15 hours a week in the academic year 2023/2024;
- settings completed all baseline measures (surveys and child consent/data) and facilitated assessments within their setting prior to randomisation;
- settings agreed to participate fully in the evaluation, including completing the programme , if selected to be in the intervention group and completing all evaluation requirements (both control and intervention) in the academic year 2023/2024;
- settings had not implemented Concept Cat within the last two years, or had not participated in Wave 1 of the Concept Cat pilot;
- settings did not have any staff employed who have attended Word Aware EY training within the last three years; and
- settings had not accessed Concept Cat resources through Lift Lessons (https://liftlessons.co).

Outcome measures

Baseline measures

Baseline assessment for children in EY settings participating in the trial consisted of two measures, both provided by Pearson Clinical Assessments[6]:

[5] A total of 18 SPHs were provided with a curated menu of programmes available for them to select for their area. In consultation with the EEF, these three SPHs selected Concept Cat a programme of interest.
[6] Wiig, Semel, and Secord (2017), see: CELF-5 UK - Clinical Evaluation of Language Fundamentals - Fifth Edition | Pearson Clinical Assessment UK.

1. **Early conceptual vocabulary.** Operationalised using the 'Basic Concepts' subtest from CELF Preschool-2 UK. This was the baseline measure used when modelling both the primary outcome, and the early numeracy secondary outcome. This is in line with the EEF guidance, which states that prior attainment should be controlled for using a regression model when the outcome is attainment, whereas when the outcome is not attainment, a parallel prior measure should be used (EEF, 2022). As explained below, the use of this measure as a baseline for the early numeracy secondary outcome was expected to reduce the burden to schools participating in the trial and that there would be sufficient overlap in the two types of ability that this baseline measure would be at least partially predictive.
2. **Early conceptual vocabulary.** Operationalised using the 'Concepts and Following Directions' subtest from CELF Preschool-2 UK. This was the baseline measure used when modelling the early conceptual vocabulary secondary outcome.

For the primary outcome, baseline and endline testing were conducted using the same measure, maximising the precision of the impact estimate for the primary outcome and accounting for imbalance at baseline. The 'Basic Concepts' subtest is one of seven subtests within the CELF Preschool-2 UK, a standardised, individually administered assessment of expressive and receptive linguistic ability specifically designed for children aged three to six. It is widely used in EY outcome assessments (for more detail, see 'Primary outcome' section below).

It must be noted that while the use of the CELF Preschool-2 UK 'Basic Concepts' subtest as baseline for early numeracy (see secondary outcomes) will likely reduce the pre- and post-test correlations for this outcome, it was deemed that there would be sufficient overlap between the latent outcomes being measured by each test for it to be a suitable baseline when modelling all outcome measures. The decision not to collect baseline scores for the EY Toolbox ENA was taken to both avoid over-burdening settings and ensure that baseline testing could be completed within five weeks to facilitate full delivery of the intervention. At analysis, ENA endline scores were well correlated with baseline 'Basic Concept' scores, affirming the appropriateness of the use of the primary outcome as a baseline for early numeracy in the secondary analysis (r=0.61).

For the alternative measure of early conceptual vocabulary used as a secondary outcome, baseline testing was similarly conducted using the same measure as endline. The 'Concepts and Following Directions' subtest is an additional subtest from the same broader CELF Preschool-2 UK assessment, which more assesses understanding of concepts within a complex grammatical context, rather than their understanding of more underlying concepts. Further details for this subtest are provided in the 'Secondary outcomes' section below.

Settings were asked to provide the evaluation team with a list of all eligible children. Elklan were provided with these child lists and a protocol for random ordering.[7] On the day of testing, the testers used the randomly ordered number lists and the child list to test children in a specific order. This ensured that there was no selection bias in who was tested on the day. Testers used this approach until a maximum of 15 children were selected for testing. Selection for testing was informed by child-level eligibility criteria, namely that the child would be expected to be in attendance for the full 30-week delivery period of the intervention.

Baseline assessment testing was carried out by Elklan in September 2023, prior to randomisation. This testing was performed, on behalf of Elklan, by independent test administrators—all of whom were qualified speech and language therapists. Test administrators were trained in the use and administration of CELF Preschool-2 UK, including how to conduct practice sessions. Data was initially collected using paper-based sheets as per standard delivery of CELF Preschool-2 UK before being uploaded by the test administrators to a secure portal to facilitate continual quality assurance by both Elklan and the evaluation team. Results in the last row of Table 19 indicate that at baseline there was little difference in the CELF Preschool-2 UK scores between the treatment and control groups.

[7] This was a series of numbered lists based on setting size, with numbers being presented in random order for the given number of eligible children in a setting.

Primary outcome

Due to this study being an efficacy trial, the impact of the intervention was only explored in relation to one primary outcome, as requested by the EEF (EEF, 2022). As discussed above, the primary outcome of interest, in the assessment of the efficacy of the Concept Cat intervention, is the same as the baseline measure:

1. **Early conceptual vocabulary.** Operationalised using the 'Basic Concepts' subtest from CELF Preschool-2 UK.

Post-test assessment of participating children's early conceptual vocabulary was conducted after the full delivery of the programme over June 2024. The primary outcome being assessed following implementation is the same subtest from the CELF Preschool-2 UK. This is the same primary outcome used in the pilot study (Hopkins *et al.,* 2022). The CELF Preschool-2 UK is a standardised, individually administered assessment of expressive and receptive linguistic ability specifically designed for children aged three to six. It is widely used in EY outcome assessments. The CELF Preschool-2 UK consists of seven subtests, including 'Basic Concepts'. The CELF Preschool-2 UK was judged fit for purpose since: i) it is designed to be brief (taking around five to seven minutes to administer per subtest); ii) it directly measures receptive vocabulary development; iii) it is UK norm-referenced; and iv) it has strong psychometric properties, with test-retest reliability ranging from 0.77 to 0.96 (for ages three to 11) and from 0.74 to 0.95 (for ages four to 11) (EEF, 2023).

The CELF Preschool-2 UK 'Basic Concepts' subtest measures a child's knowledge of four fundamental concepts: dimension and size; direction, location, and position; number and quantity; and quantitative equality. The testing process requires the child to respond to a description of a concept provided by selecting a picture from a set of options that they believe best corresponds with or exemplifies a given concept. The score the child receives at the end of the test is equal to the number of correctly identified concepts, and the test is terminated after five consecutive incorrect responses. The primary analysis utilised the raw scores of this subtest (scored from 0 to 18).

The pilot found slight ceiling effects in the 'Basic Concepts' subtest at baseline and understand that this may be attributable to design factors, such as the sampling strategy used (e.g. reception aged pupils, aged four to five, lower-than-average number of children with EAL) (Hopkins *et al.,* 2022). A significant negative skew (-1.76) was encountered in our baseline scores for the primary outcome, which led to concerns this could become a ceiling effect at endline, with this having significant implications for the analysis and potential underestimation of the treatment effect.

Following the identification of a negative skew baseline, the evaluation team employed three approaches at endline to assess whether the outcome distribution had become a 'ceiling effect'. First, the study team visually inspected the distribution of endline scores to identify any clustering at the upper end of the scale. This was then complemented with a calculation of Pearson's coefficient of skewness, providing insight into the extent of deviation from a normal distribution. Finally, we calculated the proportion of children achieving scores within 1 standard deviation (SD) of the maximum score. A study by Uttl (2005), which explored the effects of low measurement ceilings on analysis of verbal learning tests, found that ceiling effects may significantly increase the likelihood of Type II errors when 25% of outcome scores fall within 1 SD of the maximum value. While the study explored this in relation to tests consisting of nine to 15 items (slightly lower than the 18-item test used for this evaluation), this cut-off was employed when assessing the existence of a ceiling effect in our endline data. By integrating these three approaches, it was possible to gauge the likelihood of there being a ceiling effect in the primary outcome. The results of these checks are presented in the 'Outcomes and analysis' section below and reveal a substantial ceiling effect, requiring us to carry out further analysis on the primary outcome, which aims to account somewhat for the 'censoring' in the model outcome.

Secondary outcomes

The impact of the Concept Cat intervention on two secondary outcomes is also being captured in this trial:

1. **Early conceptual vocabulary.** Alternatively operationalised by the 'Concepts and Following Directions' subtest from CELF Preschool-2 UK.
2. **Early numeracy.** As measured using the EY Toolbox ENA.

These are outlined further below.

Early conceptual vocabulary

The first secondary outcome evaluated in this efficacy trial offers an alternative measure for early conceptual vocabulary, as operationalised by the 'Concepts and Following Directions' subtest from CELF Preschool-2 UK. This subtest measures receptive language (like 'Basics Concepts') but within an instructional sentence, with increasingly longer sentences involving more complex grammar, working memory, and executive functioning. For example, 'Point to the first monkey, second bear, and then the last fish. Go.'

This subtest measures conceptual vocabulary by evaluating a child's ability to: i) understand spoken directions containing concepts that require logical operations; ii) remember names, orders, and characteristics of items mentioned; and iii) identify the target from among several choices. A key distinction between this subtest and the 'Basic Concepts' subtest is that it explicitly assesses understanding of concepts within a complex grammatical context. Given Concept Cats specific focus on vocabulary, the latent constructs being measured by this subtest were deemed considered adequately proximal to those targeted by the intervention. The secondary analysis utilised the raw scores of this subtest (scored from 0 to 22).

Early numeracy

As outlined in the programmes theory of change (see Figure 1), Concept Cat specifically concentrates on instructing fundamental verbal concepts crucial to the mathematics and science curriculum. In doing so, it seeks to improve future attainment in Key Stage 1 mathematics and science. Recognising the central role of mathematics and science in the intervention's desired outcomes, early numeracy was incorporated as a secondary measure to assess Concept Cat's efficacy. This was measured using the EY Toolbox ENA, a set of eight short game-like tasks covering five distinct skill domains using a maximum of 85 items:

- **number sense**, which pertains to early numerical concepts and language (12 items) and rapid quantitative comparison (six items);
- **cardinality and counting**, which refers to counting a subset of items (six items), identifying digits and quantities (six items), matching digits and quantities (six items), completing number sequences (six items), discerning the relative position of digits based on their quantity (six items), and identifying the ordinal position of an object with respect to other objects in a line (six items);
- **numerical operations**, which measures a child's ability to derive information from a basic, verbal, mathematical problem (six items) and solving basic numerical equations (six items);
- **spatial and measurement constructs**, which assesses a child's ability to understand spatial and measurement concepts, such as length, size, and geospatial relations (13 items); and
- **patterning**, which refers to children's ability to discern and complete increasingly complex patterns (six items).

These five skill domains seek to measure fundamental numeric abilities found to have an important role in shaping social, emotional, cognitive, and life outcomes (Dawson *et al.,* 2020). The ENA consists of a set of iPad-based assessment tools suitable for use with children in EY settings by EYPs, with all eight tasks taking approximately five minutes to administer. Using skip and stop rules, the ENA adjusts its difficulty based on the child's age and questions answered correctly or incorrectly. It has been found to demonstrate good psychometric properties, yielding a test-retest reliability of 0.89 (Howard *et al.,* 2022).

The overall ENA score is the sum of a child's correct responses to each item, giving a maximum score of 85. While the ENA allows for discrete categorisation of scores across different skill domains, this evaluation will use the overall score in aggregate.

Sample size

Power calculations and minimum detectable effect size (MDES) calculations were performed using the *PowerUp!* tool (Dong and Maynard, 2013). The assumed desired power of 0.8, alpha of 0.05, and a high pre-/post-test correlation of 0.75. The CELF Preschool-2 UK has a published test-retest correlation of 0.95 for receptive vocabulary (Eadie *et al.,* 2014), but the evaluation team made the estimate more conservative for two reasons. First, the retest in this particular

instance was administered between two days and 24 days after the initial assessment. Furthermore, the Nuffield Early Language Intervention (NELI) effectiveness trial found pre-/post-test correlations of 0.75 using another version of the CELF (Dimova *et al.,* 2020).

The initial power calculations documented at protocol stage were based on both information provided in the EEF's Invitation to Tender, as well as on subsequent meetings with the EEF and the Concept Cat delivery team. As can be seen in Table 17, the MDES at the protocol stage was 0.249 for all children and 0.312 for EYPP children. This was calculated using a setting-level random assignment, based on the assumption of equal allocation of settings to intervention and control groups. These calculations also assumed that the average number of children per setting was 15, and that the average number of EYPP-eligible children per setting was three.

Upon collection of pupil baseline data, these MDES calculations were updated to reflect the actual trial sample, with an average total setting size of 12 children (minimum of four; maximum of 15), and an average of two EYPP-eligible children per setting. The results of these power calculations on our analytical sample at randomisation, generated an MDES of 0.253 for the overall sample, and an MDES of 0.346 for the EYPP-eligible subsample.

The MDES at analysis stage was 0.177 for the overall sample, and 0.363 for the EYPP subgroup.

Randomisation

Settings allocated to the treatment group received Concept Cat training and were expected to deliver the Concept Cat programme during the academic year 2023/2024, while those allocated to the control group were expected to carry on with business as usual until the following academic year (2024/2025) when they then received training and support to deliver Concept Cat. All settings (i.e. regardless of assignment to the treatment or control groups) were provided incentives in two tranches: £200 on completion of baseline assessments; and a further £200 on completion of all endline assessments. These funds are to be used at the discretion of the setting and could be used to buy an intervention programme of their choice once the trial ends. Randomisation was stratified on region and setting type (PVI/school-based setting) to reduce potential bias that might have been introduced through imbalance in these characteristics across treatment and control arms.

Randomisation was conducted by the RAND evaluation team on 20 September 2023. While the EEF guidance suggested collecting all baseline measures before randomisation (EEF, 2022), because of the delivery team's need to book settings for training, some settings were randomised before baseline measures were collected. To mitigate any risks to the evaluation the following measures were taken:

- To mitigate against potential attrition, only settings that had booked a date for baseline testing, had shared child data, and had signed and returned a Data Sharing Agreement were eligible for randomisation.
- To avoid post-allocation demoralisation, the results of the randomisation were shared with the delivery team so they could organise staff allocation, but settings were not informed of their allocation until they completed testing. Settings were not informed of their allocation until they had completed baseline testing (to avoid post-allocation demoralisation).

The randomisation was blinded and conducted using STATA software (StataCorp LLC, College Station, TX, USA), with the code used for this randomisation provided in Appendix G.

Statistical analysis

Primary outcome analysis

The following section is informed by the EEF's analysis guidance for efficacy trials (EEF, 2022). As detailed in the 'Introduction' section above, this efficacy trial has one primary research question:

1. What is the difference in early conceptual vocabulary development, measured by the 'Basic Concepts' subtest of the CELF Preschool-2 UK, of children in settings receiving Concept Cat in comparison to those children in control settings receiving business as usual?

To address the primary research question, an intention-to-treat (ITT) analysis was undertaken using multilevel modelling with fixed effects and random intercepts, as per the EEF guidance (EEF, 2022). This assumes a consistent average treatment effect across settings, in line with the whole-class delivery of the intervention while simultaneously accounting for setting-level variation in the mean outcome pre- and post-intervention. As the analysis of Concept Cat's efficacy was based on an ITT principle, data is analysed according to the group as randomised, regardless of whether the treatment was received as intended, and irrespective of withdrawal from the intervention post-randomisation, or deviations in programme implementation (Torgerson and Torgerson, 2008). This principle is key in ensuring an unbiased analysis of 'real-world' intervention effects and is in line with the EEF's guidance (EEF, 2022).

The EEF recommend clustering using the unit of randomisation (EEF, 2022). To account for the nested nature of the data, a hierarchical linear model with two levels (setting, and child) was fitted, controlling for pre-intervention scores at the child level. This allowed for the potential setting-level heterogeneity in the expected impact of the intervention. Due to the stratified randomisation process, the model also included terms identifying the regions and PVI status (strata).

Impact was estimated by fitting the model in **Equation 1**.

As discussed above, Equation 1, as follows, is known as a 'random intercepts' model because $\beta_{0j}=\beta_0+u_j$ is interpreted as the setting-specific intercept for setting j and $\beta_{0j}\sim i.i.d\ N(\beta_0,\sigma_{u2})$ is random (it is a number that can take any value).

Equation 1

$$Y_{ij} = \beta_0 + \beta_1 \mathrm{CONCEPTCAT}_j + Z_j\beta_2 + X_{ij}\beta_3 + u_j + e_{ij}$$

Where:

- Y_{ij}= 'Basic Concepts' subtest scores from CELF Preschool-2 UK for child *i* in setting j at endline;
- β_0 = the cluster-level coefficient for the slope of a predictor on early conceptual vocabulary;
- $\mathrm{CONCEPTCAT}_j$ = a binary indicator of the setting assignment to intervention [1] or control [0];
- β_1 = the individual-level coefficient denoting the estimated effect of assignment to treatment on the primary outcome;
- Z_j = setting-level characteristics (i.e. the stratifying variables of geographical location and PVI status, as used for randomisation);
- X_{ij} = child-level characteristics for child *i* in setting *j*, including the 'Basic Concepts' subtest from CELF Preschool-2 UK score at baseline to reduce bias in estimates (EEF, 2022);
- u_j = setting-level residuals; and
- e_{ij}= individual-level residuals.

The coefficient β_1 in Equation 1 above represents the parameter of interest, with respect to the primary outcome measure. Equation 1 was also replicated for each of the two secondary outcome measures (see 'Secondary outcome analysis' section below). Since it is plausible for early conceptual vocabulary scores to be age-correlated, bias may be introduced into the measurement of β_1. As such, a primary regression model, where age (in months) has been included as a covariate in X_{ij}, was also conducted as an additional sensitivity analysis.

Including baseline scores for the CELF Preschool-2 UK 'Basic Concepts' subtest (X_{ij}) in the regression model controls for prior attainment, hence, providing a more conservative and comparable method that will be usable across potential future trials by the EFF (EEF, 2022).

The effect size (Hedges' g) was calculated for β_1 (see 'Estimation of effect sizes' section below). The effect size is standardised using unconditional variance in the denominator and confidence intervals (CIs) will be reported to communicate statistical uncertainty as 95% CIs, in line with the EEF guidance (EEF, 2022). This tells us the average effect of the intervention on child outcomes in treatment settings compared to those in control settings.

Secondary outcome analysis

As detailed in the 'Introduction' section, this efficacy trial has two secondary research questions:

2. What is the difference in early conceptual vocabulary development, measured by the 'Concepts and Following Directions' subtest of the CELF Preschool-2 UK, of children in settings receiving Concept Cat in comparison to those children in control settings receiving business as usual?
3. What is the difference in early numeracy development measured by the ENA of the EY Toolbox of children in settings receiving Concept Cat in comparison to those children in control settings receiving business as usual?

As detailed in the 'Outcome measures' subsection below, the evaluation team attempted to answer these research questions by exploring the impact of the Concept Cat intervention on two secondary outcome measures: early conceptual vocabulary (measured by the 'Concepts and Following Directions' subtest of CELF Preschool-2 UK) and early numeracy (measured by the EY Toolbox ENA).

For both secondary outcomes, the secondary analysis followed the same procedures from the primary analysis, using Equation 1 to estimate each respective secondary outcome model, substituting the primary outcome variable in turn with each of the secondary outcome variables.

For the secondary outcome analysis for early numeracy, the X_{ij} vector for child *i* in Setting *j* was represented by the baseline scores for the 'Concepts and Following Directions' subtest of CELF Preschool-2 UK. On the other hand, the alternative measure of early vocabulary used baseline scores for the 'Concepts and Following Directions' subtest. While neither early conceptual vocabulary nor early numeracy are age-standardised, the analysis did not include age in the child-level characteristics, X_{ij}, since the trial was within one year group, somewhat minimising concerns over age effects.

To reduce the likelihood of Type I errors arising from the use of multiple secondary outcomes, we also conducted a testing correction for our secondary analysis. The evaluation team intended to account for this using a Romano-Wolf correction, given that it accounts for potentially correlated multiple outcomes, and the dependence structure of test statistics, through bootstrap (or permutation) resampling from the original data (Clarke *et al.,* 2019). While this method is more analytically and computationally intensive than the Bonferroni correction, it has been found to provide stronger control against the family-wise error rate, especially when the multiple outcomes are correlated (Clarke *et al.,* 2019). However, upon collection of endline data, missingness encountered in scores for both secondary outcomes was found to be unequal, with certain children missing endline data for one but not the other. As such, different datasets were effectively used for the analyses. Given that the correction outlined above is to be used in instances where multiple hypotheses are tested from the same dataset, it was not applied in this instance.

Analysis in the presence of non-compliance

As previously discussed, while the ITT approach used in this trial provides a more conservative estimate of intervention efficacy—aiming to more accurately capture impact in 'real-world' settings—it may also underestimate the benefit of Concept Cat, specifically among those who receive the intervention. Given that Concept Cat is a continuous whole-class intervention taking place across an extended 30-week period, incomplete participation in the intervention is inevitable (e.g. with children being absent from settings).

While the primary analysis outlined above assesses the impact of 'offering' the intervention, the average treatment effect of the treated (ATT) was ascertained by exploring its estimated impact in the presence of non-compliance, capturing the effect of the intervention specifically on those who received it. For this analysis, the evaluation team utilised the EEF's definition of compliance as 'the extent to which the critical ingredients of the intervention are delivered to and/or received by the target participants' (EEF, 2022).

After discussions with the delivery team, both Concept Cat session attendance and the number of words taught during these sessions were chosen to inform a measure of compliance. Information on words learned is collected by the delivery team at the setting level by recording the weekly focus word, providing a measure of week-on-week fidelity to the Concept Cat intervention.

Measuring child attendance is more complicated. Not all settings systematically collect attendance data, making it difficult to have accurate records of whether children were in settings on certain weeks. Therefore, as a proxy for actual child attendance, the evaluation team used attendance patterns reported by settings (i.e. the number of hours, days, or

sessions a child is expected to attend a setting) at the start of the intervention in Autumn Term 2023. It is important to note that attendance patterns do not give granular data on whether a child was actually present in the setting (e.g. if a child was absent due to illness or a holiday). However, given the unavailability of actual attendance data, attendance patterns were considered a useful proxy.

These two metrics were selected, after consultation with the delivery team, because they each capture specific aspects of compliance. The proposed attendance measure informs us how often children were expected to be present to receive the intervention. However, relying solely on attendance might incorrectly classify settings where children were present, but the intervention was not delivered by staff, as 'compliant'. Likewise, exclusively focusing on 'words learned' might mistakenly categorise absent children as 'compliant'.

To generate a single measure of compliance, these two metrics were combined in a way that enabled meaningful and straightforward interpretation: a single binary measure of compliance was generated, based on specific pre-set cut-offs. Cut-offs were based on discussions between the EEF, Concept Cat, and the evaluation team and reflect the delivery team's views of what constitutes suitable dosage. This binary variable took on the value of 1 (indicating 'compliance') if both of the following conditions were satisfied:

- the child meets the eligibility criteria of having attended at least 15 hours every week over the 30-week delivery period;[8] and
- the setting has taught 30 words over the 30-week delivery period.

The compliance thresholds of: i) 15 hours per week; and ii) a total of 30 words taught over the 30-week delivery period were discussed and agreed with the delivery team as constituting compliance. While the authors recognise that using a proxy binary variable drastically limits the range of the attendance data collected—that is, compared to a continuous variable that directly counts the number of hours attended or the number of words taught by the setting—this was deemed a necessary compromise, due to the fact that not all settings systematically collect child attendance data (as discussed above).

Details of this compliance measure are presented in Table 6 below. While the ultimate compliance indicator for the Complier Average Causal Effect (CACE) analysis is a binary measure with specific thresholds for each aspect of compliance, a descriptive overview of both child attendance and setting delivery is also provided, which combined provides a more complete indication of intervention dosage received by children, in the absence of formal attendance records.

Table 6: Child-level compliance measures

Compliance criterion	Data source	Compliance indicator
Setting-level compliance	Log of number of words taught by settings recorded by delivery team	Binary compliance indicator, which takes on a value of 1 for each child who met the eligibility criteria for attendance, and who's setting taught 30 words over the full delivery period
Child-level compliance	Data from settings indicating whether a child attended 15 hours or more per week during Autumn Term 2023	

However, given the binary nature of this proposed compliance indicator, it is not possible to use it to understand how dosage (i.e. the number of weeks during which children were present for an intervention session) impacts the primary outcome.

For each aspect of the analysis, the CACE was calculated through a two-stage least squares instrumental variable approach. The first stage of this approach involved regressing the binary compliance indicator on allocation to treatment, providing an estimate of how assignment of children and settings to receive Concept Cat encourages uptake of the intervention. This effectively provides an overall 'compliance rate'. The second stage modelled the primary outcome (as

[8] The original model was five half-days a week in a Maintained setting. However, in order to make sure that children attending PVIs had enough dosage 15 hours over three days was seen as the minimum in order to make good progress.

presented in Equation 1) but included this predicted compliance in place of treatment assignment. The Hedges' g derived from this model was similarly calculated to provide an estimate of the CACE. The results of each of these two models allowed us to discern the degree to which compliance with the Concept Cat intervention improves outcomes for children. This analysis was conducted for the primary outcome only.

Missing data analysis

Missing data can arise from a variety of sources and at multiple stages of a trial, including but not limited to participant attrition at the setting or child levels, errors by test administrators, and errors in data collection. While the ITT basis of the investigation required inclusion of all available data from settings as randomised, the robustness of estimated effect sizes from this analysis may be impeded if in instances where endline data is incomplete for all randomised settings. As reported in the 'Impact evaluation results' section below, missingness encountered for the primary outcome was beyond 5%, and therefore, a non-random pattern of missingness could potentially bias the results (Schafer, 1999). In accordance with the EEF analysis guidance, the evaluation team therefore, further investigated the nature of, and potential systematic patterns to, this missingness (EEF, 2022).

As an initial step in the missing data analysis, the evaluation team examined attrition across trial arms to evaluate bias, through creating cross-tabulations of the proportions of missing values on characteristics available at baseline, at both child and setting levels. This provides preliminary insight into missingness patterns across treatment and control groups. Furthermore, to explore whether missing data in the primary outcome was systematically driven by observable characteristics and therefore, 'missing at random' (MAR), missingness was then modelled at follow-up (defined as children with missing primary outcome data at endline) as a function of the baseline covariates included in the aforementioned cross-tabulations. This consisted of running logistic regression models, using a binary outcome variable denoting missingness at endline (where 1 'missing' and 0 'complete') and mirroring the multilevel structure of the models used in the main analysis, with children nested within settings. As further detailed in the 'Impact evaluation results' section below, while the results of these analyses did suggest that the data was potentially MAR, the only variable found to be driving this missingness was setting type. Given that this variable was already included in the primary analysis model, no further analysis account for data MAR was required. It should be noted that missingness systematically driven by unobservable characteristics not captured in the data, and therefore 'missing not at random' (MNAR), may exist whenever missingness is encountered. While it is impossible to statistically test for the presence of MNAR, further sensitivity analyses can be run in cases where MNAR is suspected. However, given the moderate degree of missingness overall (13%), and the predominant mechanisms of missingness being known (i.e. children leaving settings and random child absence on testing days), it was deemed that MNAR was highly unlikely and therefore, no such additional sensitivity analyses were conducted.

Subgroup analyses

As detailed in the trial protocol, this study asks additional research questions pertaining to Concept Cat's impact on three subgroups:

> 1a. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of EYPP/FEEE-eligible children, compared to non-EYPP/FEEE-eligible children?
>
> 1b. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of children with EAL, compared to non-EAL children?
>
> 1c. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of children with SEND, compared to non-SEND children?

Two main approaches were employed to subgroup analysis. The first approach incorporated rerunning the model used in the primary analysis, but on a restricted sample of children identified as being a member of the relevant subgroup. In the case of EYPP-eligible children (research question 1a), a key target group for the EEF, the study team conducted subgroup analysis for children eligible for these funding provisions, using the same model used for the primary analysis (Equation 1), but where the analytical sample only contained children from this subgroup. Information on child eligibility was obtained from settings themselves in the baseline data collection phase, and analysis was undertaken using a binary EYPP variable, where EYPP-eligible = 1; non-EYPP/FEEE-eligible = 0. This subsample model allows for exploration of whether the intervention is effective within the EYPP subgroup specifically.

Further sensitivity analysis was also conducted by introducing an interaction term to the primary outcome model between the binary EYPP-eligibility indicator and treatment assignment. This allowed us to explore the potential *differential* effects of Concept Cat on EYPP/FEEE-eligible children, while retaining the whole analytical sample in the model.

The mediating impact of EYPP-eligibility was estimated by fitting the model in **Equation 2**, found below, which is similarly a 'random intercepts' model as used for the main primary and secondary analysis models.

Equation 2

$$Y_{ij} = \beta_0 + \beta_1 \mathrm{CONCEPTCAT}_j + \beta_2 \mathrm{EYPP}_{ij} + \beta_3 \left(\mathrm{CONCEPTCAT}_j * \mathrm{EYPP}_{ij}\right) + \beta_4 Z_j + \beta_5 X_{ij} + u_j + e_{ij}$$

Where:

- Y_{ij}= 'Basic Concepts' subtest scores from CELF Preschool-2 UK for child *i* in setting *j* at endline;
- β_0 = the cluster-level coefficient for the slope of a predictor on early conceptual vocabulary;
- $\mathrm{CONCEPTCAT}_j$ = a binary indicator of the setting assignment to intervention [1] or control [0];
- β_1 = the individual-level coefficient denoting the estimated effect of assignment to treatment on the primary outcome;
- $\mathrm{CONCEPTCAT}_j * \mathrm{EYPP}_{ij}$ = the interaction term between assignment to treatment and EYPP-eligibility;
- β_2 = the individual level coefficient denoting the effect of EYPP-eligibility on the primary outcome.
- β_3 = the individual-level coefficient denoting the mediating effect of EYPP-eligibility on the impact of assignment to treatment on the primary outcome;
- Z_j = setting-level characteristics i.e. the stratifying variables of geographical location and PVI status (as used for randomisation);
- X_{ij} = child-level characteristics for child *i* in setting *j*, including the 'Basic Concepts' subtest from CELF Preschool-2 UK score at baseline to reduce bias in estimates (EEF, 2022);
- u_j = setting-level residuals; and
- e_{ij}= individual-level residuals.

Identical analyses were conducted with subgroups of children with EAL (research question 1b) and SEND (research question 1c) to identify any differential impact. For all subgroup analyses, effect sizes and statistical uncertainty were calculated from the coefficient for the interaction term and communicated as per the primary analysis (also see subsection on 'Estimation of effect sizes').

Additional analyses and robustness checks

There is significant variability in setting sizes, with the smallest setting included in the trial only having four children tested at baseline, and the largest having 15 children tested, with an overall mean of 12 children tested per setting. As set out in the trial protocol, the intended number of children for each setting included in the trial was 15. After baseline testing was complete, 28 settings (15 settings in the treatment group, 13 settings in the control group) out of 89 settings have met this threshold. While the evaluation team initially planned to conduct a subgroup analysis of settings meeting this 15-child threshold, only two settings yielded complete endline data for 15 children, and therefore, the analysis was not conducted.

As detailed in the 'Impact evaluation results' section, a significant ceiling effect was observed in the endline scores for the primary outcome. Ceiling effects can increase bias in model estimates due to insufficient variability at the upper limit of a scale, effectively 'censoring' the 'true' measurement of the latent characteristic being operationalised and leading

to significant uncertainty around estimated treatment effects. In this case, it could lead to an underestimation of the true treatment effect.

Given this, the impact of the Concept Cat intervention on early vocabulary development was re-estimated using a Tobit model. Tobit models are designed to re-estimate linear models while accounting for censoring in the outcome, such as ceiling effects, allowing for more accurate measurement of associations between an independent variable and a latent outcome (McBee, 2010). They combine the use of a probabilistic density function to estimate the relationship between an independent variable and the outcome below the threshold, with a cumulative density function to model the relationship above the threshold (McBee, 2010). The structure of and other covariates used in this model were otherwise identical to the primary analysis model.

Estimation of effect sizes

The effect sizes calculated for our analysis used the formula for cluster randomised controlled trials given in the EEF evaluator guidance (EEF, 2022), as adapted from Hedges (2007)[9]:

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\frac{(n_1 - 1)sd_1{}^2 + (n_2 - 1)sd_2{}^2}{n_1 + n_2 - 2}}}$$

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between the intervention and control group adjusted for baseline characteristics and $\sqrt{\frac{(n_1-1)sd_1{}^2+(n_2-1)sd_2{}^2}{n_1+n_2-2}}$ is an estimate of the pooled unconditional population SD.

From the primary outcome model, we took each group's adjusted mean and variance to calculate the effect size. This variance is the total variance (across both child and setting levels, without any covariates, as emerging from a 'null' or 'empty' multilevel model with no predictors). The ES therefore, represents the proportion of the population SD attributable to the intervention (Hutchinson and Styles, 2010). A 95% CI for the ES, which takes into account the clustering of children in settings, is also reported. Effect sizes, in terms of the treatment effect, were calculated for each of the models estimated.

Estimation of intracluster correlation coefficient (ICC)

Since the nesting of children within settings is an important analytical tenet of this efficacy trial, ICC is an essential metric for disentangling the degree of variance in the outcomes that is explained by between- and within-cluster variation. For the final analysis, ICCs have been reported as they were at three stages: during the protocol stage; at randomisation; and during analysis (see Table 17). The ICC at analysis stage was based on the primary outcome measure at both baseline and endline; and was calculated using: i) the same model as Equation 1; and ii) a model similar to that documented in Equation 1 but with no covariates, hence accounting for the clustering of children in settings (i.e. the 'empty model').

[9] Where:

$(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the adjusted difference in means in the outcome between treatment and control group, accounting for the multilevel structure of the data.

$sd_1{}^2$ is the variance of the treatment group; and equally defined for the control group.

n_1 is the number of individuals in the treatment group, equally defined for the control group.

IPE[10]

Research methods

The IPE used a mixed methods approach to collect data for the efficacy trial, incorporating the following addressed in the subsections below.

Baseline practitioner survey

An online baseline practitioner survey was distributed to all lead practitioners in intervention and control settings to collect information about existing 'usual practice', such as: i) existing vocabulary teaching practice (especially in terms of settings using Word Aware or STAR approaches); ii) existing practice in identifying children with higher language needs; and iii) other potential moderators (such as practitioner qualifications and motivation and setting type and class size).

Eighty-four responses were received from across 79 settings; 43 responses from EYPs in settings allocated to the control group (31 Maintained, 12 PVIs), 41 responses from lead practitioners in settings allocated to the intervention group (28 Maintained, 13 PVIs). Response rate at the setting level was 90% (or 79 out of an expected 88 settings). On the other hand, response rate at the EYP level was 95% (or 84 out of 88 expected respondents), but this includes five additional responses from EYPs from settings where two surveys were submitted. This was because settings were asked that the EYP receiving/who would receive[11] the three-hour training should complete the survey and, in some instances, this was more than one EYP. Table 7 shows the responses received.

Most settings completing the survey had more than 30 children in their cohort (see Table 8). Furthermore, the percentage of EAL, EYPP, and children with speech and language needs was spread across control and intervention groups and setting types (see Table 9). The survey was designed by the evaluators and was distributed, via Qualtrics, to practitioners through email during September 2023 prior to randomisation. The survey was closed following randomisation in October 2023 one week after the commencement of training. The baseline survey was designed to address IPERQ3 and IPERQ5b with regard to business as usual and motivation to implement Concept Cat. The survey also helped the evaluation team to purposely select settings to visit for observations.

Table 7: Baseline practitioner survey responses received by respondent role

	Intervention			Control		
	Maintained	PVI	Total	Maintained	PVI	Total
Setting managers	10	10	20	8	9	17
EYP	18	3	21	23	3	26
Total	28	13	41	3	12	43

[10] See **IPE guidance** for further details.
[11] In control settings, the lead EYP was the person who would be receiving the three-hour Concept Cat training the following academic year. In intervention settings, the lead EYP was the person who would be receiving the three-hour Concept Cat training as part of the evaluation research.

Table 8: Cohort size by allocation (intervention or control) and setting type (Maintained or PVI) at baseline

Cohort size	Intervention			Control		
	Maintained	PVI	Total	Maintained	PVI	Total
10 or less	0	0	0	0	0	0
11 to 15	0	0	0	0	1	1
16 to 20	2	0	2	4	2	6
21 to 25	4	3	7	3	1	4
26 to 30	0	0	0	4	0	4
More than 30	22	10	32	20	8	28
Total	28	13	41	31	12	43

Table 9: Percentage of EAL, EYPP, and children needing speech and language support by allocation (intervention or control) and setting type (Maintained or PVI) at baseline

	% of children	Intervention			Control		
		Maintained	PVI	Total	Maintained	PVI	Total
EAL	0–20%	13	12	25	14	8	22
	21–40%	5	0	5	7	0	7
	41–60%	1	0	1	2	0	2
	61–80%	2	0	2	3	0	3
	81–100%	5	0	5	2	1	3
	Total	26	12	38	28	9	37
EYPP	0–20%	11	12	23	13	9	22
	21–40%	5	0	5	10	1	11
	41–60%	4	0	4	2	0	2
	61–80%	1	0	1	0	0	0
	81–100%	0	0	0	0	1	1
	Total	21	12	33	25	11	36
Speech and language support	0–5%	8	11	19	14	9	23
	6–10%	3	1	4	7	2	9
	11–20%	6	2	8	3	1	4
	21–30%	4	0	4	4	0	4
	31%+	0	0	0	1	0	1
	Total	21	14	35	29	12	41

Training observations

Training observations (n=2) were completed by two members of the evaluation team for quality assurance purposes. An observation schedule for the three-hour training (delivered to lead practitioners) was developed by the evaluation team with support from the delivery team and using information gained from the pre-trial. A separate observation schedule was developed by the evaluation team, with support from the delivery team, to collect data on the one-hour staff training. Both schedules were designed to ensure the key aspects of the programme, as outlined in the TIDieR model. In addition, they supported the development of a quality and fidelity framework, which was used for the development of surveys, observation schedules, and interview schedules, as well as to support interpretation of the data at analysis. The delivery team shared a summary of this information with the evaluation team. The training observations took place between September 2023 and October 2023. This aspect addressed IPERQ1 with regard to implementation fidelity.

Baseline parent survey

An online baseline parent survey was created by the evaluation team on Qualtrics and distributed to settings via email. The settings were sent the links to the surveys in the communications documents sent by the delivery team; settings were responsible for distributing the survey link via email to parents. The survey was to be completed by the main parent/carer for each child in both intervention and control settings. Where surveys are not completed by parents, this was followed up by the evaluation team and the delivery team, via the settings.

In total, 221 parents (147 intervention, 74 control) completed the baseline survey, representing a 25% response rate (221 out of a potential 880, assuming an average of ten children aged three to four per setting). The survey was designed to establish, at baseline, current home literacy practices and relevant aspects of the HLE. The Home Learning Environment Index (HLEI; Melhuish *et al.,* 2013), was embedded into the parent baseline (and endline) surveys. The HLEI consists of eight items designed to measure the HLE (e.g. 'Does anyone at home ever read with your child?'). Parents select a response from a set of pre-determined options (e.g. 'occasionally, or less than once a week', 'one or two days a week', 'three times a week'). This aspect addressed IPERQ6 with regard to monitoring home literacy practices and the HLE (when linked with endline survey data). It also sought to understand any differences for children who are disadvantaged, children with EAL, and children with higher language needs. The parent baseline survey was distributed during September 2023 to October 2023.

Concept Cat monitoring data

Concept Cat monitoring data (n=45, 32 Maintained, 13 PVI settings; intervention only) consisted of three separate elements:

- **Coach visit to settings logs.** These were developed by the Concept Cat developers and were used to gather data on inputs as outlined in the programme theory of change (e.g. the words taught within the setting), implementation quality and fidelity (e.g. the quality of teaching and inclusion of reviewing elements of the programme), and perceived impact. The logs were completed by Concept Cat coaches during their setting visits during the implementation year. Data was used to monitor implementation fidelity and addressed IPERQ1. Data for all settings was provided to the evaluation team in June 2024.

- **Good practice network logs.** These were developed by the Concept Cat developers and were used to gather data on attendance (as outlined below), contributions made by practitioners within the supervisions, and the sharing of resources and good practice. The logs were completed by Concept Cat coaches. Data was used to monitor compliance and implementation fidelity addressing IPERQ1 with regard to implementation fidelity and IPERQ4 with regard to attendance. This data was shared with the evaluation team in June 2024.

- **Training logs/attendance data.** These were used to capture data in the following ways:

    - Training attendance for lead practitioners/practitioners who had received the three/one-hour training (respectively). Data was used to monitor compliance as outlined above and addressed IPERQ1 and IPERQ4. Data was collected in October 2023.

    - Good practice network attendance for lead practitioners was used to capture the number of group sessions attended by lead practitioners. Data was used to monitor compliance as outlined above and addressed IPERQ1 and IPERQ4. Data was collected in June 2024.

    - Class lists/attendance/pupil turnover data (n=382) (72%, 382 of a potential 527) was collected to ensure children were eligible for the intervention (i.e. attended the setting for at least 15 hours per week) at baseline and to monitor how much dosage of the intervention was received by the intervention children at endline. This data was used to inform impact. Baseline child attendance data was gathered following recruitment in September 2023 and endline data was gathered in June 2024.

Setting visit observations

Setting visit observations were conducted with a subsample of intervention settings (n=4) and control settings (n=4). Settings were purposely sampled depending on setting type and pre-selected responses to the baseline practitioner survey including, whether the setting explicitly taught concepts and whether the setting had previously used the STAR

approach or the Word Aware approach. It was hoped that the observations would be carried out evenly across PVI and Maintained settings. However, one PVI setting dropped out of the observation at short notice and was hence, replaced with another setting; the only setting that could accommodate this was a Maintained setting.

The evaluation team deemed it better to have an uneven number of Maintained and PVI settings rather than fewer visits due to the small sample size. Of the eight visits that took place, three settings were PVI settings (two intervention and one control), and five were Maintained settings (two intervention and three control). Each visit was scheduled for half a day. The aim of these visits was to establish what Concept Cat looked like in settings compared to 'usual practice' (especially in terms of settings using Word Aware or STAR approaches and teaching concepts).

Observation schedules were developed by the evaluation team with input from the developers and using the implementation fidelity and quality framework. The schedules themselves were designed to understand whether similar approaches to Concept Cat were being used in control settings. The observations covered: i) IPERQ2, to assess any adaptations intervention settings had made to the implementation of the programme; ii) IPERQ3, which assessed the nature of business as usual in terms of normal conceptual vocabulary instruction and normal practice in identifying children with higher language needs (which has been triangulated with survey data); and iii) IPERQ4 to address the quality of delivery in intervention settings. Setting visits took place over the period of February 2024 to May 2024.

Embedded setting visit observations

Embedded setting visit observations were also conducted in intervention settings (n=4; two Maintained settings and two PVI settings) who had not taken part in the general setting observations. Settings were purposely sampled based on geographical spread, setting type (an equal mix of PVIs and Maintained settings), and answers given within the survey, such as setting size and numbers of EYP and EAL children. Two settings were classed as large settings (more than 30 children), and two were classed as small settings (less than 30 children). In addition, they covered both high (above 80%) and low (20% or below) numbers of EAL and/or EYPP children.

The purpose of these observations, which took place over three consecutive days (two hours per day), was to monitor implementation fidelity to understand, on a more in-depth basis, how the different elements of the intervention (i.e. whole-class implementation, implicit play, whole-class review, and family engagement) were implemented over a number of days. Observation schedules were developed by the evaluation team with support from the developers and were informed by the fidelity and quality implementation framework (discussed above). Data collected as part of the observation visits was designed to address IPERQ1, IPERQ2, and IPERQ4 (i.e. implementation fidelity, adaptations to the programme, quality of delivery, and the extent to which training and resources supported practitioners' ability to teach Concept Cat). The embedded setting observations were conducted by two researchers from the evaluation team and took place during the period of April 2024 to June 2024, so settings had, had time to fully embed the programme following training.

Practitioner interviews

Practitioner interviews were conducted with practitioners from intervention settings (n=7) and control settings (n=4) following the setting observations and embedded setting observations. It was originally planned for eight intervention setting interviews to take place, but one setting was not able to do the interview during the observation visit, and despite repeated requests, was unable to participate in a follow-up interview.

The interview schedules were developed by the evaluation team to understand business as usual practice (control settings), implementation fidelity, particularly focusing on teachers' skills, resources, and knowledge and how this is embedded in teaching practice, and how practitioners support those with higher language needs. The interviews also covered barriers and facilitators to delivery as intended and adaptations and the reasons why adaptations, if any, had been made. They therefore, addressed IPERQ1, IPERQ2, IPERQ4, IPERQ5a and IPERQ5b. The interviews took place over the period of February 2024 to June 2024.

Practitioner endline surveys

Online practitioner endline surveys were developed by the evaluation team and distributed, via Qualtrics links, to all lead practitioners (who had received [intervention group]/who would receive [control group] the three-hour training), setting managers, and one EYP (who had received [intervention]/who would receive [control] the one-hour training). As with the baseline survey, in some settings, more than one EYP received the three-hour training. In 17 settings, managers

had also received the three-hour training (15 settings), the one-hour training (one setting), or both (one setting). A total of 157 responses were received from across 78 settings, (a response rate of 89% at the setting level, 78 out of a potential 88 settings), 78 control (62 Maintained, 16 PVI[12]), 79 intervention (55 Maintained, 24 PVI). The response rate for setting managers was 68% (60 out of a potential 88 respondents), for EYPs receiving the three-hour training the response rate was 67% (59 out of a potential 88 respondents), and at the EYP level for those receiving the one-hour training it was 43% (38 out of a potential 88 respondents).

Table 10 shows the responses received by setting role. The majority of settings completing the survey had more than 30 children in their cohort (see Table 11) and the percentage of EAL, EYPP, and children with speech and language needs was relatively balanced across control and intervention groups, although EYPs from Maintained settings reported having slightly higher numbers of EYPP, EAL, and children with higher language needs compared to PVI settings (Table 12).

The purpose of the survey was to collect data pertinent to different research questions (i.e. IPERQ1, IPERQ2, IPERQ4, IPERQ5a, IPERQ5b, IPERQ6, and IPERQ8). The survey sought to uncover any changes in: practitioner knowledge, understanding, and motivation; perceived changes in parental engagement; programme adaptations; perceived impact on child attainment (especially for disadvantaged children, children with higher language needs and EAL children); and potential wider impacts and unintended consequences. The survey with setting managers also included questions with regard to the costs of the programme and unintended consequences. The survey was distributed in May 2024 to allow practitioners enough time to complete the survey prior to the end of the intervention period.

Table 10: Responses received for endline practitioner survey

Completion	Intervention			Control		
	Maintained	PVI	Total	Maintained	PVI	Total
Setting managers	11	8	19	19	5	24
Setting manager + three-hour training	6	1	7	6	2	8
Setting manager + one-hour training	0	0	0	1	0	1
Setting manager + three-hour AND one-hour training	1	0	1	0	0	0
EYP three-hour training	17	9	31	16	5	21
EYP three-hour training + one-hour training	1	6	7	1	4	5
EYP one-hour training	19	0	19	19	0	19
Total	55	24	79	62	16	78

Table 11: Cohort size by allocation (intervention or control) and setting type (Maintained or PVI) at endline

Cohort size	Intervention			Control		
	Maintained	PVI	Total	Maintained	PVI	Total
10 or less	0	5	5	0	0	0
11 to 15	1	1	2	2	3	5
16 to 20	4	4	8	4	1	5
21 to 25	6	2	8	8	2	10
26 to 30	11	4	15	16	2	18
More than 30	33	8	41	32	8	40
Total	55	24	79	62	16	78

Table 12: Percentage of EAL, EYPP, and children needing speech and language support over allocation (intervention or control) and setting type (Maintained or PVI) at endline

% of children		Intervention			Control		
		Maintained	PVI	Total	Maintained	PVI	Total
EAL	0–20%	29	16	50	23	13	36
	21–40%	8	3	11	11	0	11
	41–60%	7	1	8	10	2	12
	61–80%	4	0	4	6	0	6
	81–100%	3	0	3	4	0	4

[12] One school who classed themselves as a special school was added to the PVI setting so as not to identify the setting within the analysis.

	Total	51	20	71	54	15	69
EYPP	0–20%	19	15	34	18	12	30
	21–40%	13	3	16	15	1	16
	41–60%	14	0	14	15	0	15
	61–80%	2	0	2	4	0	4
	81–100%	0	0	0	0	0	0
	Total	48	18	78	52	13	65
Speech and language support	0–5%	5	7	12	3	4	7
	6–10%	10	7	17	9	6	15
	11–20%	21	4	25	14	2	16
	21–30%	10	2	12	20	2	22
	31%+	4	0	4	12	0	12
	Total	50	20	70	58	14	72

Endline parent surveys

Online endline parent surveys were developed by the evaluation team and distributed via a Qualtrics link to settings via delivery team communications. Settings were responsible for distributing the link to parents. Where low levels of parental completion of the survey were noted (at the setting level), this was followed up by the evaluation team and the development team via the settings. While the evaluation team strived to get as many responses as possible it should be noted that parental focus was a secondary outcome. At endline, the survey was completed by 246 respondents (134 intervention, 112 control), which represents a 28% response rate (246 respondents out of a potential 880 respondents assuming an average of ten three- to four-year-old children per setting). This is a similar response rate to that observed with parents who completed the parent baseline survey (25% response rate; 221 respondents out of a potential 880 respondents; see Table 13). In addition, the response rate is similar across control and intervention groups at baseline and endline. The evaluation team are confident that the responses came from a similar pool of respondents. However, few parents completed the survey at both timepoints (n=19 responses; intervention only).

The endline survey sought to identify any changes in home literacy practices and the HLE, through the embedded HLEI (Melhuish *et al.,* 2013; see Table 14) and to determine for families in intervention settings: i) how settings provided information to parents relating to Concept Cat (e.g. words taught); ii) additional support given to children with higher language needs; iii) barriers and facilitators to implementing the programme in the home environment; iv) perceived gains in children's conceptual knowledge and wider language; and v) any unintended consequences of the programme.

Thus, the endline parent surveys sought to address: i) IPERQ6, to establish engagement in the programme and whether this may be linked to child outcomes; ii) IPERQ7, to understand barriers and facilitators of home implementation and any wider impacts on the HLE; and iii) IPERQ8, to understand any unintended consequences of the programme for the families. Endline parent surveys were distributed in May 2024 to allow parents enough time to complete prior to the end of the programme.

Table 13: Demographic questions from the parent survey[a]

		Intervention			Control			
		Baseline	Endline	Total	Baseline	Endline	Total	Overall total
Other children	No	50	32	85	27	25	56	141
	Older and younger	13	10	25	9	5	15	40
	Older	49	50	105	22	45	74	179
	Younger	5	42	80	16	37	55	135
Total		**117**	**134**	**295**	**74**	**112**	**200**	**495**
Language spoken	English	142	109	261	85	86	171	432
	Bengali	1	0	1	0	1	1	2
	Pashto	0	3	3	0	0	0	3
	Polish	1	1	2	0	3	3	5
	Punjabi	2	6	8	1	3	4	12
	Urdu	0	5	5	0	6	6	11
	Other	1	11	12	5	15	20	32
	Total	**147**	**135**	**292**	**91**	**114**	**205**	**497**
Developmental concerns	No	136	116	252	69	97	166	418
	Yes	23	20	43	20	16	36	79
	Social emotional	1	4	5	7	3	10	15
	Language and Comms	22	12	34	12	8	20	54
	Numbers	0	0	0	1	0	1	1

Total	**159**	**136**	**295**	**89**	**113**	**202**	**497**

[a] Total in the final column refers to 'No' and 'yes' answers only.

Table 14: Responses received for the HLEI at baseline and endline by child's setting allocation

Allocation	Baseline	Endline
Intervention	143	108
Control	74	97

Analysis

The IPE was designed to test the workings of the logic model to check whether the intervention was operating as hypothesised. Table 15 shows how the methods described above answered the research questions and how they link to the implementation dimensions. In addition, Appendix O shows how the findings were used to support or counter the logic model and its constituent elements.

Table 15: IPE methods overview

Focus	Research questions	Data collection											Implementation dimensions							
	Covered in Wave 1 and trial	Monitoring data			Observations of training delivery	Setting observations	Embedded setting observations	Practitioner interviews	Baseline practitioner survey and quiz	Endline practitioner survey and quiz and manager survey	Baseline parent survey	Endline parent survey	Fidelity/adherence	Dosage	Quality	Reach	Responsiveness	Programme differentiation	Monitoring of control condition	Adaptation
		Coach visits to settings	Group supervision logs	Training logs / attendance data																
Implementation	**IPERQ1.** How closely does the Concept Cat programme, as implemented in settings, follow the intended model (implementation fidelity), as outlined in the TIDieR framework including extended implementation for focus children? What are the barriers and facilitators to implementation and how do these differ, if at all, between setting type (PVI/Maintained schools)?																			
	IPERQ2. What, if any, adaptations have been made to the programme during implementation? Why were they made? What do they look like?																			
	IPERQ3. What is the nature of business as usual with regard to vocabulary instruction? How does this differ between control and intervention settings? What are the similarities/differences between setting type (PVI/Maintained)? How does programme delivery differ from business as usual?																			

Impact on teaching, knowledge, and practice	**IPERQ4.** Have practitioners attended mandatory training? To what extent have training and resources supported practitioners' ability to effectively teach Concept Cat? What is the quality of delivery (i.e. how well are different components of the intervention delivered; Education Endowment Foundation, 2024., p. 6)?	
	IPERQ5a. To what extent have practitioners developed their knowledge about conceptual vocabulary and skills in identifying and supporting the conceptual vocabulary development of children with higher language needs (i.e. those identified as focus children)?	
	IPERQ5b. To what extent are practitioners motivated to implement, and continue to implement, Concept Cat? Is this motivation different across setting type (PVI/Maintained) and if so, why?	
Impact in the home	**IPERQ6.** To what extent have settings engaged families with the programme and in what ways? Are there differences between setting type (PVI/Maintained) in the ways settings have engaged with families? How is this linked, if at all, to child outcomes?	
	IPERQ7. What are the barriers and facilitators for families in home implementation of the programme, particularly for focus children, disadvantaged children, and those who are EAL? What, if any, are the wider impacts on the HLE?	
Unintended Consequences	**IPERQ8.** To what extent does Concept Cat result in positive or negative unintended consequences for settings, practitioners, children, families, and the HLE?	

NB: orange cells = implementation dimensions; purple cells = data collection, pink cells = also covered in Wave 1 and trial.

Quantitative data

Surveys

Quantitative data was gathered through a variety of closed questions (e.g. Likert scales) in the EYP and baseline and endline parent/carer surveys to test the outputs, short-term outcomes, and long-term outcomes of the logic model, as well as the causal assumptions. Raw data from the surveys was transformed into percentages to allow for comparative analysis (i.e. across control and intervention groups, across baseline and endline, and across setting types). Where appropriate, baseline and endline survey data were analysed ('Usual practice' and 'Parental engagement' sections in 'IPE results' section) to show changes over the duration of the evaluation period.

For the most part, responses from all participants answering the survey were analysed. However, at baseline, only one EYP from each setting was asked to complete the survey while, at endline, up to three practitioners per setting were asked to complete the survey. This was because, at baseline, the evaluation team wanted to gain an overview of practice prior to the intervention, and one practitioner was deemed sufficient. At endline, the evaluation team wanted to gain the perspectives of more than one practitioner, particularly in intervention settings. Therefore, for the analysis conducted on the types of interventions used across control and intervention settings, and across setting types, it was deemed appropriate to only conduct the analysis at the setting level (see section on 'Usual practice' in 'IPE results' section). Where possible, responses from EYPs who completed both the baseline and endline survey was prioritised.

HLEI

Data from the HLEI (which was embedded within the baseline and endline parent/carer surveys) was transformed into an overall score for each parent/carer who completed the survey at each timepoint. Independent t-tests were performed on baseline and endline data where the participant sample was different at baseline and endline. A separate analysis was conducted on those who completed the HLEI at both timepoints (n=19 respondents). Dependent t-tests were performed on the data provided by the same participants. The data was used to understand whether the Concept Cat programme influenced the HLE.

Observations

The embedded observation schedules were designed to capture data on implementation of the key elements of the programme and measure implementation fidelity by understanding whether practitioners were able to continuously deliver all elements of the Concept Cat approach. The observation schedules were developed to allow the evaluators to tally the number of times a particular Concept Cat element was used within the Teach, Activate, and Review of the STAR approach, Focus children, and Environment elements of the programme. The tallies were then summed to provide an overall score within each area.

The control/intervention midpoint observation schedules were designed to capture data on the key strategies being used across control and intervention settings to understand usual practice. The observation schedules were designed so the evaluators could identify whether a strategy was used within the settings observed; this was then summed to provide a score for control and intervention settings to allow them to be compared. The analysis meant that differences and similarities could be identified between control and intervention groups on the frequency in which strategies were used.

Monitoring data

Monitoring data was gathered from the Concept Cat coach logs to monitor compliance (attendance at training, supervisions, and group supervisions), implementation fidelity, and quality. For each of the key elements of the coach observations (Teach, Activate, Review elements of the STAR approach, families, focus children, and children attending the setting on a part-time basis), there were separate categories that were listed on the log that were collected as part of the monitoring data. For example, the heading of 'Teach' included separate sections on whether parts of the 'Teach' elements were observed, an observation on delivery of this element, and whether all EYPs were involved in implementation.

Under each element, the researcher rated separate sections for either being quality of implementation (i.e. delivery of observation, eight categories) or fidelity (e.g. all staff involved in implementation, 11 categories). Where Concept Cat coaches had indicated implementation as green in the logs, the evaluation team coded a value of '1'. Where Concept Cat coaches indicated amber, it was coded as a value of '0.5', and where it was indicated red, it was coded as a value

of ‘0’. The logs were completed by Concept Cat coaches for each visit so where settings had all six visits there was a maximum score of 6 for each of the categories per setting. In some of the logs, the category was left blank since it was not a focus of the visit and therefore, not listed as being observed. In this instance, all settings were given a score of 1 as we could not assume this element was not being implemented. This allowed the evaluation team to give a quality and fidelity rating by totalling the score across all settings (broken down by setting type), which was a maximum score of 189 for Maintained settings and 74 for PVI settings. These scores were then calculated as percentages: high (85 to 100%), medium (75 to 84%), or low (<74%) for quality and fidelity for each of the key elements of the programme.

Qualitative data

Surveys

Qualitative data was gathered through open-ended questions within the EYP endline survey to provide additional information where certain answers had been given to quantitative questions. As such, this data was hand-coded deductively to support the quantitative data analysis. Examples from this qualitative data are provided in the text to provide illustrations of the main codes. Such extracts are labelled according to the respondent’s setting type to provide further context.

Observations

Researcher notes from the embedded observations and the control/intervention midpoint observations were coded deductively to draw out common themes around programme implementation. This approach was taken to support the quantitative data analysis.

Interviews

Interview data from EYP interviews was transcribed and imported into NVivo 14 qualitative analysis software. One researcher familiarised themselves with the data and created a coding frame using deductive and inductive techniques. The coding frame reflected the research questions and elements of the theory of change while being sufficiently flexible to allow the data to generate new themes previously unidentified within the IPE. This coding frame was then sense-checked and modified following feedback from the two researchers who conducted the observation visits and interviews.

Extracts from the interviews are provided in the findings to illustrate the main themes emerging from the analysis and, as such, should be seen as indicative of the larger corpus. Interviewees are identified by code number and type of setting to provide additional context. When a viewpoint was expressed by a limited number of interviewees, this is indicated in the main text. However, this data is primarily qualitative and should not be interpreted as an attempt to quantify practitioners’ perspectives and experiences.

Attribution

Where qualitative data is used in this report, they are labelled to indicate the source of the data, the reference number and allocation, and setting type (i.e. source, setting reference and allocation, setting type, e.g. Researcher notes, 271C, Maintained).

Costs

Data on implementation costs was collected through endline surveys with practitioners in both control and treatment settings and through a brief questionnaire with the delivery team.

Cost data from control settings was compared against the cost of implementing Concept Cat against business as usual, which could correspond to the costs associated with implementing other programmes that are similar in scope to Concept Cat. Performing this comparison also aligns with EEF’s 2023 cost evaluation guidance (EEF, 2023b).

The cost evaluation considered both direct and indirect costs incurred by implementing Concept Cat and those incurred by implementing similar programmes. These direct and indirect costs include but are not limited to: i) time away from teaching due to participation in training and other programme activities; ii) staff cover for teaching staff participating in out-of-setting programme-related activities; iii) prices of instructional materials; and iv) additional staff workload required to run the programme.

Timeline

Table 16: Timeline

Dates	Activities	Responsible parties
October 2022 – November 2022	• IDEA (Intervention Delivery and Evaluation Analysis) workshop and set-up meetings	RAND Europe, University of Leeds, Better Communication CIC, the EEF
December 2022 – February 2023	• Development and finalisation of data protection documents (Data Protection Impact Assessment, Legitimate Interests Assessment, and Data Sharing Agreement) • Ethical approval	RAND Europe, University of Leeds
December 2022 – March 2023	• Development and finalisation of recruitment documents (MoU, information sheets, privacy notices, withdrawal forms)	RAND Europe
March 2023 – September 2023 September 2023	• Development of the trial protocol • Publication of the trial protocol	RAND Europe The EEF
May 2023 – July 2023	• Development of baseline surveys	University of Leeds
June 2023 – September 2023	• Recruitment of settings into the trial	Better Communication CIC
September 2023	• Collection of child-level demographic data (e.g. EYPP status, SEND status, and EAL status)	RAND Europe
September 2023 – October 2023	• Baseline test administration (CELF Preschool-2 UK) • Randomisation and informing schools of their allocation	Elklan RAND Europe
September 2023 – October 2023	• Baseline practitioner survey administration • Baseline parent survey administration • Collection of attendance data (baseline) • Training observations	University of Leeds
October 2023 – November 2023	• Development of quality and fidelity framework	University of Leeds
November 2023 – December 2023	• Analysis of baseline surveys	University of Leeds
November 2023	• Official registration of trial protocol on ISRCTN	RAND Europe
November 2023 – January 2024	• Development of observation and interview schedules	University of Leeds
February 2024 – June 2024 April 2024 – June 2024	• Midpoint setting visit observations • Practitioner interviews • Embedded setting observations	University of Leeds
March 2024 – May 2024	• Development of endline surveys	University of Leeds
April 2024 – July 2024 August 2024	• Development of SAP • Publication of SAP	RAND Europe The EEF
May 2024 – July 2024	• Analysis of observations	University of Leeds
June 2024	• Coach visit to settings logs • Good practice network logs • Collection of attendance data (endline)	University of Leeds
June 2024 – July 2024	• Collection of compliance data from settings	RAND Europe
June 2024 – July 2024	• Endline test administration (CELF Preschool-2 UK, EY Toolbox ENA)	Elklan
May 2024 – July 2024	• Endline practitioner survey administration • Endline parent survey administration	University of Leeds
July 2024 – October 2024	• Analysis of impact evaluation data (baseline vs endline tests) • Analysis of compliance data • Analysis of monitoring data • Analysis of practitioner interviews • Analysis of endline survey data	RAND Europe, University of Leeds
October 2024 – November 2024	• Writing the final report	RAND Europe, University of Leeds

Impact evaluation

Participant flow including losses and exclusions

Figure 2: Participant flow diagram (two arms)

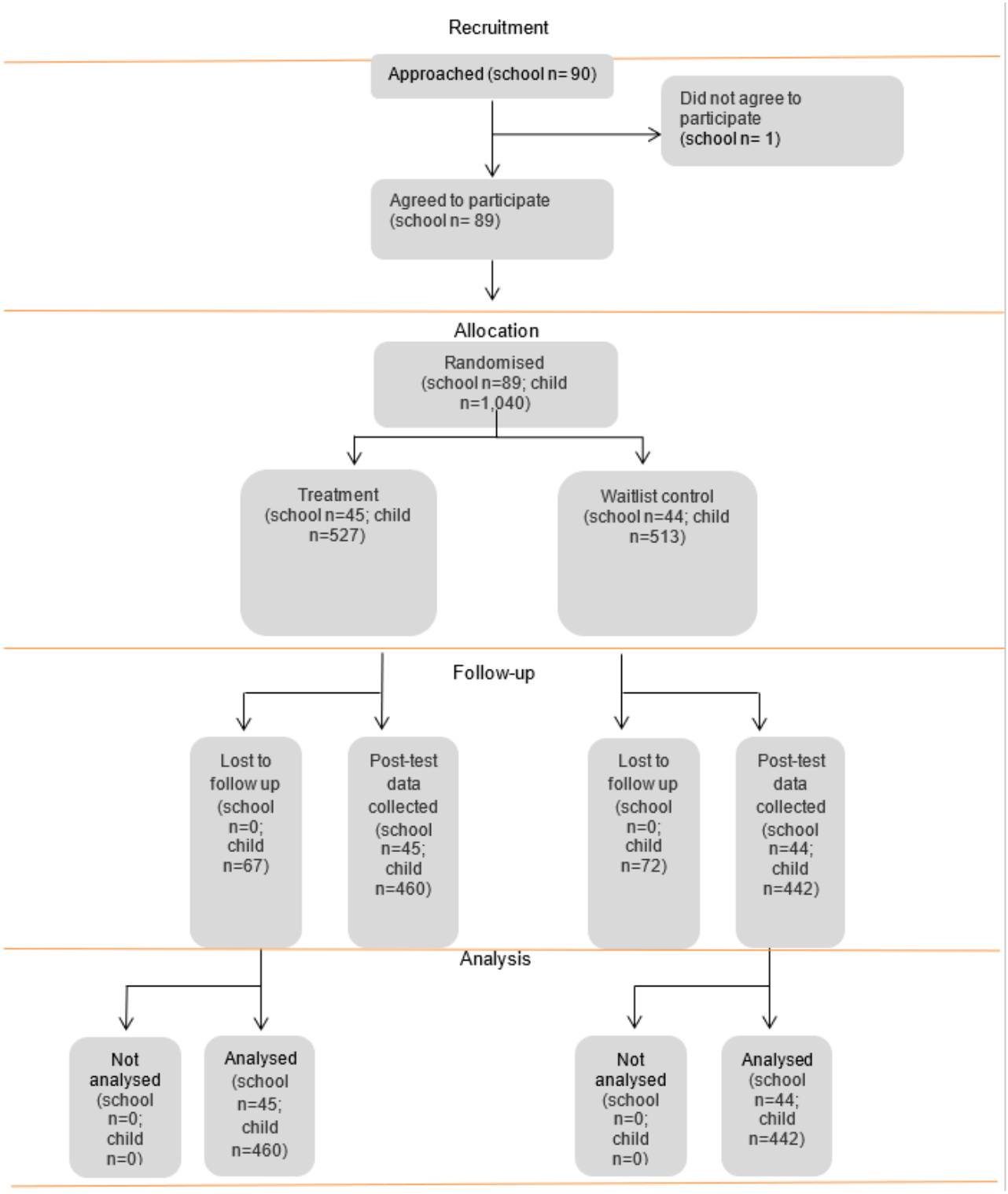


Table 17: MDES at different stages

		Protocol		Randomisation		Analysis	
		Overall	EYPP	Overall	EYPP	Overall	EYPP
MDES		0.249	0.312	0.253	0.346	0.177	0.363
Pre-/post-test correlations	Level 1 (child)	0.75	0.75	0.75	0.75	0.60	0.56
	Level 2 (setting)	0.15	0.15	0.15	0.15	0.88	0.30[a]
ICCs	Level 2 (setting)	0.15	0.15	0.15	0.15	0.14	0.04
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2	2	2
Average cluster size		15	3	11.7	1.8	10.1	1.9
No. of settings	Intervention	45	45	45	45	45	45
	Control	45	45	44	44	44	44
	Total:	90	90	89	89	89	89
No. of children[b]	Intervention	675	68	527	78	460	85
	Control	675	68	513	78	442	86
	Total:	1,350	136	1,040	156	902	171

[a] It is important to note that in the subgroup analysis, the pre-/post-test correlations use the covariance and SDs calculated just for each group, giving subgroup specific correlations rather than pooled ones. When you have a very small subgroup, like EYPP that, behave very differently to the others, then the subgroup pre-/post-test correlations can be very different from the pooled ones.
[b] Based on the number of children that could be visited over the course of two days of testing.

As presented in Table 17 above, the MDES at analysis was found to be 0.177 for the overall sample, a significant improvement to the MDES estimated at both protocol (0.249) and randomisation (0.253) stages. This increase is largely driven by an improvement in the pre-/post-test correlations at Level 2 (0.88), compared to what was assumed at both protocol and randomisation stages (0.15). Our original estimations for Level 2 pre-/post-test correlations were particularly conservative given the paucity of pre-/post-test correlations in the EY space when the protocol was written. Subsequently there has been increasing research to suggest that there is a strong relationship between pre-/post-test correlations at the pupil and school levels, at least at primary and secondary school (Singh ***et al.,*** 2023). The ICC at analysis stage was found to be 0.14, with an average cluster size of 10.1, falling from 11.7 at randomisation. For a Level-1 and Level-2 R^2 values for both secondary outcomes, please see Appendix F.

The trial was insufficiently powered to detect a moderate effect among the EYPP subgroup, with a MDES of 0.363. Although a very low ICC of 0.04 was observed for the subgroup, the number of children in the EYPP subgroup was lower than expected at protocol stage, averaging only two individuals per setting. Furthermore, the pre-/post-test correlations are lower than for the overall sample.

Attrition

Figure 2 above presents the extent of attrition, at both the child and setting level, encountered at each stage of the trial. There was no setting-level attrition, with only one setting withdrawing prior to randomisation and was as such excluded from the trial altogether. As depicted in Table 18 below, the trial saw a total child attrition rate of 13.3% between the randomisation and analysis stages, with the control group experiencing a marginally higher attrition rate of 14.0%, compared to 12.7% within the treatment group. Thirty-eight percent (N=52) of this attrition was a result of children leaving

settings, and therefore were not able to be tested at endline regardless of attempts at follow-up. The remaining 62% (n=86) of this missingness arose from children being absent for endline testing at all attempts of follow-up, which is unsurprising given the EY context of the trial.

Table 18: Child-level attrition from the trial (primary outcome)

		Intervention	Control	Total
Number of children	Randomised	527	513	1,040
	Analysed	460	442	902
Child attrition (from randomisation to analysis)	Number	67	72	138
	Percentage	12.7%	14.0%	13.3%

Child and setting characteristics

Setting-level characteristics

The balance of setting type, region, and setting-level proportions of females, EYPP-eligible, EAL, and SEND children across treatment and control arms at randomisation is presented in Table 19 below. The make-up of PVI settings is well balanced across the two trial arms, at 27% and 29% in treatment and control groups, respectively. The make-up of settings from each region exhibits a similarly high degree of consistency across trial arms, with settings from all regions forming between 23% and 27% across treatment and control groups. The balance of these two categorical setting-level variables is expected given their inclusion as stratification variables in the randomisation process. In terms of the average make-up of subgroup children in settings, the proportion of females, EYPP-eligible, and EAL children exhibit particularly good balance across treatment and control arms (53.3% vs 53.4% average proportion of female children; 18.7% vs 20.0% average proportion of EYPP-eligible children; 26.1% vs 25.3% average proportion of EAL children). However, the setting-level proportional make-up of SEND children, while only 2 percentage points higher in control (7.8%) than treatment (5.9%), represents a slightly larger magnitude of difference on account of the very small number of SEND children in the sample overall.

Child-level characteristics

For child-level characteristics, baseline balance in terms of child gender, EYPP-eligible children, EAL children, SEND children, child age, and CELF Preschool-2 UK ‘Basic Concept’ scores is similarly presented in Table 19 below. Child gender exhibits very high balance across treatment and control groups, with female children constituting 53.9% and 53.8% of each trial arm, respectively. The proportion of EYPP-eligible children overall in treatment and control is well balanced, with this subgroup making up 19,4% and 20.1% each group, respectively. EAL children also make up a similar proportion of each group, consisting of 24.3% of the treatment group, and 24.0% of the control group. The overall proportion of SEND children, as was the case at the setting level, presents a higher degree of difference (2 percentage points at 5.6% treatment and 7.6% control), relative to the size of SEND subsample.

Baseline scores in the primary outcome at randomisation were marginally higher in the control group than in intervention (effect size = 0.11), with this difference holding low statistical certainty, as indicated by the CIs on the effect size (-0.07–0.27).

Table 19: Baseline characteristics of groups as randomised

Setting level (categorical)	Treatment group		Control group		
	n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Setting type:					
School-based	32/44 (0)	73%	32/45 (0)	71%	
PVI	12/44 (0)	27%	13/45 (0)	29%	
Region:					
Trafford (Bright Futures SPH)	11/44 (0)	25%	11/45 (0)	24%	
Everton	11/44 (0)	25%	12/45 (0)	27%	
West Midlands (HEART North)	12/44 (0)	27%	11/45 (0)	24%	
West Midlands (HEART South)	10/44 (0)	23%	11/45 (0)	24%	
Setting level (continuous)	n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Prop. Female	45/45 (0)	53.3%	44/44 (0)	53.4%	
Prop. EYPP-eligible	45/45 (1)	18.7%	43/44 (1)	20.0%	
Prop. EAL	45/45 (0)	26.1%	45/45 (0)	25.3%	
Prop. SEND	44/45 (1)	5.9%	44/44 (0)	7.8%	
Child level (categorical)	n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Gender:					
Female	284/527 (0)	53.9%	276/513 (0)	53.8%	
Male	243/527 (0)	46.1%	237/513 (0)	46.2%	
EYPP:					
EYPP-eligible	102/527 (0)	19.4%	100/513 (16)	20.1%	
Not EYPP-eligible	425/527 (0)	80.6%	397/513 (16)	79.9%	
EAL:					
EAL	128/527 (0)	24.3%	123/513 (0)	24.0%	
Not EAL	399/527 (0)	75.7%	390/513 (0)	76.0%	
SEND:					
SEND	29/527 (13)	5.6%	39/513 (0)	7.6%	
Not SEND	485/527 (13)	94.4%	474/513 (0)	92.4%	
Child level (continuous)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size
Age in months	513/513	43.41 (3.53)	527/527	43.54 (3.51)	
CELF Preschool-2 UK 'Basic Concepts'[a]	513/513	11.12 (4.39)	527/527	11.60 (4.19)	0.11 (-0.07–0.27)

[a] *Based on all available data at baseline.*

Outcomes and analysis

Primary analysis

CELF Preschool-2 UK 'Basic Concepts' subtest

As outlined in the 'Methods' section above, the primary outcome used for the trial was the 'Basic Concepts' subtest from the CELF Preschool-2 UK assessment. At endline, scores for this outcome had a mean of 15.26, and an SD of 2.85 across the whole sample. The distribution of endline scores is presented in Figure 3: Distribution of CELF Preschool-2 UK 'Basic Concepts' scores at endline and point towards a significant ceiling effect, which was expected given the distribution of scores encountered at baseline. While this has significant implications for the robustness of the primary analysis, the evaluation team proceeded with analysis as planned, with an additional sensitivity check to account for the ceiling effect in the outcome using a Tobit model (see 'Additional analyses and robustness checks' section below).

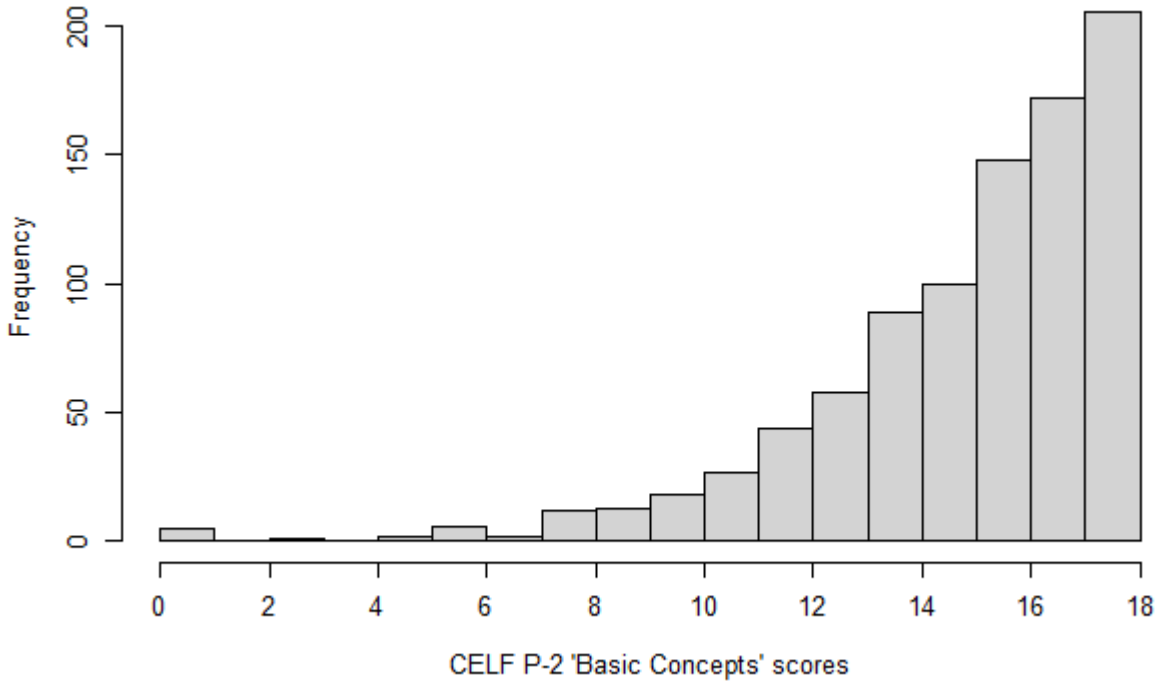


Figure 3: Distribution of CELF Preschool-2 UK 'Basic Concepts' scores at endline

The impact of Concept Cat on the primary outcome was explored using a linear random intercept model, accounting for the multilevel structure of data, with children nested within settings, and controls added for baseline scores and stratification variables (see the 'Methods' section above for more details). A complete case analysis was used, generating an estimate of the treatment effect on an ITT basis. Initial inspection of the residuals from the model showed a non-normal distribution (see Appendix F); as such, the displayed CIs and associated p-value were re-estimated using bootstrapping.

The results of the primary analysis are displayed in Table 20 below, with the information used to calculate the effect size presented in Table 21. The results point towards a positive association between assignment to receive the Concept Cat intervention, and conceptual vocabulary, as measured by the 'Basic Concepts' subtest. The effect size of 0.18 is indicative of approximately two months' additional progress among the treatment group, with the relatively narrow CIs (0.05–0.30) and associated p-value (0.01) allowing for a firm rejection of the null hypothesis. In relation to the intervention's associated theory of change, therefore, these results provide evidence to suggest that Concept Cat may improve children's development of early conceptual receptive vocabulary. However, given the significant ceiling effects encountered in endline scores, the results should be interpreted with caution, due to the inadequacy of the measure to effectively differentiate between the early vocabulary skills of children who scored the highest on the test. As discussed below, the higher magnitude of effect observed within the EYPP subgroup may suggest that this ceiling effect is contributing to a slight underestimation of the 'true' effect size in the primary analysis.

Table 20: Primary analysis

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	460 (67)	15.57 (15.32–15.82)	442 (72)	14.94 (14.67–15.21)	902 (460; 442)	0.18 (0.05–0.30)	0.01

Table 21: Primary analysis effect size estimation

Outcome	Unadjusted differences in means	Adjusted differences in means	Intervention group		Control group		Pooled variance
			n (missing)	Variance of outcome	n (missing)	Variance of outcome	
CELF Preschool-2 UK 'Basic Concepts' subtest	0.63	0.5	460 (67)	7.7	442 (72)	8.4	8.07

Secondary analysis

CELF Preschool-2 UK 'Concepts and Following Directions' subtest

As detailed in the 'Methods' section above, the evaluation also explored the impact of the intervention on an alternative measure of early conceptual vocabulary, using CELF Preschool-2 UK 'Concepts and Following Directions' scores. The endline scores for this outcome had a mean of 9.44, and an SD of 4.04. The distribution of endline scores is depicted in Figure 4 below, and points towards a broadly normal distribution.

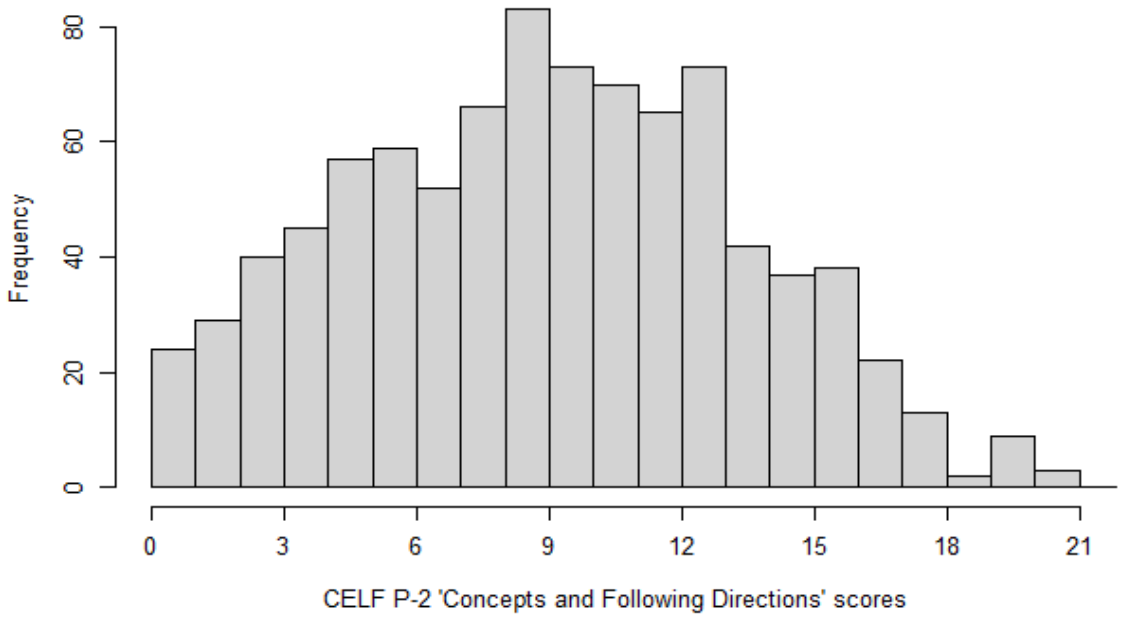


Figure 4: Distribution of CELF Preschool-2 UK 'Concepts and Following Directions' scores at endline

As with the primary analysis, the impact of the Concept Cat intervention on this alternative measure of early conceptual vocabulary was explored using a random intercept linear regression model, with an identical structure to the primary analysis model except for using baseline scores for the 'Concepts and Following Directions' subtest to control for prior attainment. After running this model, the CIs and p-values for the estimated effect size were again re-calculated using bootstrapping, on account of residual non-normality (see Appendix E).

As depicted in Table 22 below, our analysis provides no evidence that the Concept Cat intervention improves conceptual vocabulary, as alternatively operationalised by the 'Concepts and Following Directions' subtest. The small effect size of 0.05, and wide CIs, are indicative of an additional one month's progress for this outcome. As outlined in the 'Outcome measures' subsection, this secondary measure assesses understanding of concepts within a complex grammatical context, requiring children to identify target words based on instructions from test administrators. The positive effect detected in the primary analysis, combined with the observed lesser effect detected on this secondary outcome, suggests that Concept Cat may have more impact on early concept comprehension, than it does on understanding concepts within a complex grammatical context. Therefore, in terms of the short-term outcomes proposed in the theory of change, while Concept Cat seemingly helps to benefit children's *understanding* of concepts, improvements to the *understanding* of these concepts within complex grammatical contexts may not be as immediately evident.

Table 22: Secondary analysis: CELF Preschool-2 UK 'Concepts and Following Directions' subtest

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Concepts and Following Directions' subtest	460 (06)	9.66 (9.25–10.08)	442 (72)	9.21 (8.8–9.62)	902 (460; 442)	0.05 (-0.07–0.17)	0.42

EY Toolbox ENA

As detailed in the programme's theory of change (see Figure 1), a proposed longer term outcome of the intervention is increased early numeracy development, with initial improvements to early conceptual vocabulary facilitating an improved understanding of key concepts to science, technology, engineering, and mathematics (STEM) subjects at Key Stage 1. Therefore, the impact of the Concept Cat intervention on early numeracy was explored, as measured by the EY Toolbox ENA. The mean for this outcome at endline was 32.41, with an SD of 13.10. The distribution of scores at endline across the overall sample are presented in Figure 5 **Error! Reference source not found.**and reflect a broadly normal distribution.

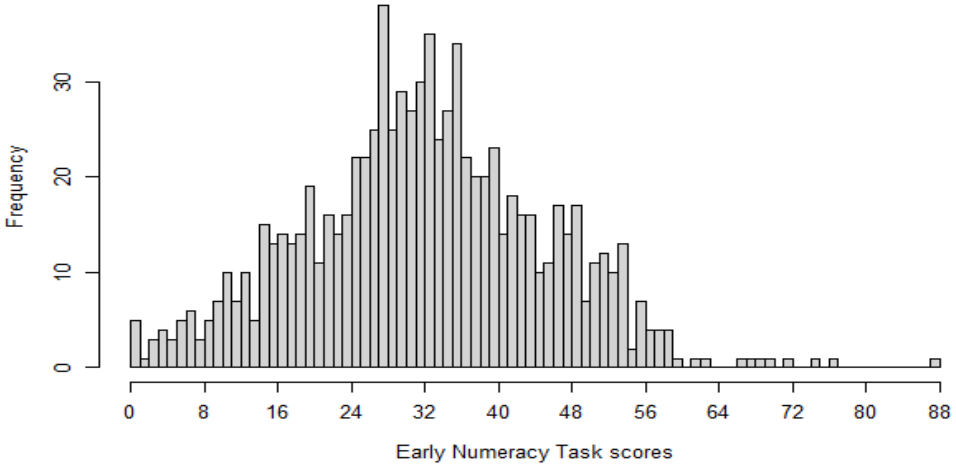


Figure 5: Distribution of EY Toolbox ENA scores at endline

The impact of the Concept Cat intervention on this early numeracy outcome was explored using an identical model structure and covariates used in the primary analysis model. Again, non-normality was observed in the model residuals (see Appendix E); therefore, CIs and their associated p-value were re-estimated using bootstrapping. It should be noted that the sample encountered a slightly higher degree of missingness for the EY Toolbox ENA outcome (17.6%), when

compared to the other two outcomes. However, given the limited size of this difference, it was not deemed necessary to warrant further sensitivity checks, and the analysis was performed as planned. The results from this model are presented in Table 23 below and provide evidence that the Concept Cat intervention leads to improvements in early numeracy. The effect size of 0.13 is indicative of approximately two months' progress, with moderately narrow bootstrapped CIs (0.01–0.25) allowing us to reject the null hypothesis in this instance.

While the key element of Concept Cat's theory of change is its aim to ultimately improve STEM attainment through enhancing comprehension of concepts fundamental to the Key Stage 1 curricula for these subjects, this is a longer term aim of the intervention. In this way, any improvements to early numeracy would be expected to be mediated through initial improvements to early conceptual vocabulary, given that the intervention specifically targets language development. The presence of an effect size for the secondary outcome, alongside the detected impact on the primary outcome, supports this proposed mechanism in the theory of change—that is, improvements in conceptual vocabulary are associated with improvements in early numeracy.

Table 23: Secondary analysis

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (Boot. 95% CI)	Boot. p-value
EY Toolbox ENA	444 (83)	33.81 (32.59–35.03)	416 (97)	30.92 (29.67–32.17)	860 (444; 416)	0.13 (0.01–0.25)	0.04

Analysis in the presence of non-compliance

The primary analysis detailed above, employed a complete case analysis, calculating a treatment effect on an ITT basis. In this way, while it captures the effect of being assigned to receive the intervention, it does not account for potential non-compliance to the intervention in terms of both child attendance and setting-level delivery. As outlined in the 'Methods' section above, the degree to which actual amount of participation in the intervention impacted CELF Preschool-2 UK 'Basic Concepts' subtest scores at endline was also explored.

Compliance to the intervention was documented in two ways: i) child attendance patterns; and ii) the number of words taught by the setting over the delivery period. While our overall measure of compliance took the form of a binary indicator (i.e. '1' if both child-level and setting-level requirements were met and '0', otherwise), Table 24 and Table 25 provide a disaggregated breakdown of child- and setting-level compliance. As can be seen in the tables below, compliance was particularly high at the child level, with all but eight children for whom data was collected meeting the minimum attendance requirement at the start of Autumn Term 2023. Setting-level compliance was more moderate, with 31 out of 45 settings teaching 30 words before endline testing.

Table 24: Child-level compliance

	Number of children
Attended 15 hours or more at start of Autumn Term 2023	381
Did not attend 15 hours or more at start of Autumn Term 2023	8
Missing	138

Table 25: Setting-level compliance[a]

Number of words taught	Number of settings
30 words	31
<30 words	13
Missing	1

[a] NB: Data on words taught at settings was not continuous. The delivery team provided the dates by which 30 words were delivered. The 30-word compliance threshold was established through comparing the date of delivery of 30 words to the date of endline testing; if settings had delivered 30 words before endline testing, they were deemed 'compliant'.

The overall combined child- and setting-level compliance is also presented in Table 26 below. Of the 393 children for whom we have compliance data, 307 (78%) were deemed 'compliant'. It is important to note however, that due to some settings failing to provide child attendance patterns for the start of Autumn Term 2023, we encountered a significant degree of missing compliance data (27%), which resulted in these children being excluded from the subsequent CACE analysis. Given this high level of missing child-level compliance information, it is impossible to rule out that this attendance data is MNAR, and therefore that there may be unobserved variables affecting both the likelihood of inclusion in the compliance analysis, and primary outcome scores.

Table 26: Overall compliance

Compliance	Number of children
Compliant	364
Non-compliant	29
Missing	144

Estimation of CACE

The results of the CACE analysis are presented in Table 27 below. They show a strong association between intervention receipt and early conceptual vocabulary, as measured by the CELF Preschool-2 UK 'Basic Concepts' subtest. The effect size of 0.27 suggests that full compliance with the Concept Cat intervention may be associated with up to four months' additional progress in the outcome, supported by narrow bootstrapped CIs (0.12–0.41). In relation to the programme's associated theory of change, therefore, this may suggest that children who receive complete delivery of the intervention experience significant gains to early conceptual vocabulary compared to children who received either no or partial delivery. While this may indicate greater intervention efficacy when delivered as intended, these findings must be interpreted with caution due to the substantial missingness in child-level compliance data. An examination of endline primary outcome scores reveals systematic differences between the compliant group and the treatment group with missing compliance data. The higher effect size observed in the CACE analysis compared to the primary analysis may be driven by a lower mean and higher variance in endline scores among settings that did not provide child-level compliance data, compared to those that did, potentially biasing the effect size estimate.

Table 27: CACE analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest (CACE)	338 (189)	15.8 (15.52–16.07)	442 (71)	14.94 (14.67–15.21)	780 (338; 442)	0.27 (0.12–0.41)	0.00

Further sensitivity analysis to account for contamination

As discussed in the 'IPE results' section below, one of the settings in the control group implemented a more limited version of the Concept Cat intervention, using the video and story materials used as part of delivery. While the analysis for this trial has been conducted on an ITT basis, we conducted further sensitivity analysis, rerunning the primary analysis on a restricted sample that excludes this setting, to account for potential contamination across the two trial arms. The results of this analysis on the reduced sample are presented in Table 28 below. This analysis produced a negligible difference to the observed effect size and bootstrapped CIs, suggesting that the contamination encountered in the trial had very little impact on the observed effect size in the primary analysis.

Table 28: Primary analysis sensitivity check (without contamination setting)

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	460 (67)	15.57 (15.32–15.82)	432 (71)	14.96 (14.68–15.23)	892 (460; 432)	0.18 (0.04–0.31)	0.01

Missing data analysis

As presented earlier in Table 18: in the 'Attrition' subsection above, the total missingness encountered in the trial totalled 13.3% across both trial arms, and therefore, the study team conducted further analysis to explore potential patterns to this missingness, in line with the proposed approach set out in the 'Methods' section.

Descriptive exploration of missingness

Table 29: Balance of characteristics across treatment and control arms, at randomisation and analysis

Child level (categorical)	**At randomisation**				As analysed			
	Intervention		Control		Intervention		Control	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Female	284/527 (0)	54%	276/513 (0)	54%	244/460 (0)	53%	235/442 (0)	53%
Male	243/527 (0)	46%	237/513 (0)	46%	216/460 (0)	47%	207/442 (0)	47%
EYPP-eligible	102/527 (0)	19%	100/513 (16)	20%	85/460 (0)	18%	86/442 (11)	20%
Not EYPP-eligible	425/527 (0)	81%	397/513 (16)	80%	375/460 (0)	82%	345/442 (11)	80%
EAL	399/527 (0)	24%	123/513 (0)	24%	109/460 (0)	24%	103/442 (0)	23%
Not EAL	128/527 (0)	76%	390/513 (0)	76%	351/460 (0)	76%	339/442 (0)	77%
SEND	29/527 (13)	6%	39/513 (0)	8%	28/460 (0)	6%	32/442 (0)	7%
Not SEND	485/527 (13)	94%	474/513 (0)	92%	419/460 (0)	94%	410/442 (0)	93%
Child level (continuous)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean
Age in months (at end of Summer Term)	527/527 (0)	53.5	513/513 (0)	53.4	460/460 (0)	53.4	442/442 (0)	53.5

As a basic first step, the potential underlying patterns in missingness were explored by comparing the proportional balance of particular characteristics between treatment and control arms, at randomisation and analysis. This balance of characteristics is presented in Table 29 above.

As displayed in Table 29 above, balance of key characteristics between treatment and control groups remains fairly consistent between randomisation and analysis stages, with no child-level characteristics demonstrating more than a single percentage point change in either treatment or control groups. This initial descriptive exploration of missingness suggests that there is no discernible pattern to missing endline data in the primary outcome.

Further exploration of missingness

Despite this initial exploration providing no clear indication of endline data being MAR, we conducted further confirmation using logistic regression models. As outlined in the 'Methods' section, this involved creating a binary variable, which took on the value of '1' if the child had missing data, and then modelling this dummy variable on the covariates included in the primary analysis model, still accounting for the multilevel structure of the data. The characteristics presented in Table 29 above were also included, allowing for the exploration of potential observable driving factors to missing 'Basic Concepts' data at endline.

The raw results of these multilevel logistic regression models are presented in Appendix F and provide some indication that missing endline data for the primary outcome is related to setting type, with children attending PVI settings being less likely to have missingness at endline (with a p-value of 0.02 in the second and third logistic models). While this points towards some degree of data MAR, setting type was already controlled for in the preceding analytical models. Therefore, no further sensitivity checks were required to account for this. Furthermore, the pseudo R^2 accompanying all three of these logit models is low, with this reaching a maximum of 0.06 for the third model, indicating that the covariates included in these models only explain a very small fraction of the missingness encountered at endline, suggesting that the degree to which missingness is correlated with observables is limited.

It is important to note when missingness is high, and no discernible pattern in this missingness can be attributed to observable characteristics beyond setting type, it is difficult to rule out the possibility that data is MNAR and could therefore, be contingent on unobserved variables not captured in our dataset. However, given the relatively low incidence of missingness, and the predominant mechanism of missingness being known (i.e. children leaving settings and random absence on testing days), missingness in endline data, beyond what was explained by the models outlined above, was assumed to be missing completely at random (MCAR). Under this assumption, further sensitivity analysis to explore this possibility was not deemed to be required.

Subgroup analyses

As per the 'Methods' section, three sets of subgroup analyses were conducted to explore the impact of the Concept Cat intervention on children who are EYPP-eligible, EAL, and SEND, with this analysis being carried out in relation to two outcomes; conceptual vocabulary, as measured by the CELF Preschool-2 UK 'Basic Concepts' subtest (primary outcome); and early numeracy, as measured by the EY Toolbox ENA (secondary outcome).

Subgroup analysis 1: EYPP

The results for the EYPP subgroup analysis on CELF Preschool-2 UK 'Basic Concepts' scores are depicted in Table 30 and Figure 6 below. The results of the analysis on the EYPP-eligible subsample point towards a positive and statistically significant effect, with the effect size of 0.25 being indicative of an additional three months' progress in the primary outcome.

While the effect size observed in the EYPP-eligible subgroup (0.25) is larger than that in the primary analysis (0.18)—suggesting that the Concept Cat intervention may have a greater positive impact on EYPP-eligible children. The interaction model showed a similar overall effect size for EYPP-eligible children (0.23, presented in Appendix F). The raw regression output in Table F1 shows the estimated difference in outcome measures associated with being in the treatment group, receiving EYPP, and being both in the treatment group and receiving EYPP, from Equation 2. The produced interaction term coefficient is 0.16 ($p = 0.71$). As shown in Figure 6 below, the distribution of CELF Preschool-2 UK 'Basic Concepts' scores at endline among the EYPP-eligible subgroup, although still negatively skewed, does not exhibit as pronounced a ceiling effect. This greater variation in scores at the upper end of the scale within the EYPP subgroup may have led to the estimation of a larger effect size, as more significant improvements in conceptual vocabulary were less likely to be 'censored' compared to in the overall sample.

The results of the EYPP subgroup analysis on the EY Toolbox ENA are similarly presented in Table 30; these results show the intervention having no impact on early numeracy within the EYPP subgroup, as indicated by the effect size (-0.03) and wide associated CIs. The effect size estimated for the interaction model presented in Appendix F is similar. The interaction model results for the EYPP subgroup, presented in Table F2 produce an interaction term coefficient of -3.09 (p = 0.11). When examined in comparison to observed impacts among the full sample, this EYPP subgroup analysis suggests that while Concept Cat may help to improve early conceptual vocabulary among lower income children, the more latent beneficial effects on early numeracy are not experienced as intensely or immediately for this group.

Table 30: EYPP subgroup models

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	N (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	85 (17)	15.08 (14.52–15.64)	86 (14)	14.52 (13.95–15.1)	171 (85; 86)	0.25 (0.02–0.51)	0.03
EY Toolbox ENA	81 (21)	28.43 (26.02–30.84)	81 (19)	29.12 (26.88–31.35)	162 (81; 81)	-0.03 (-0.31–0.26)	0.84

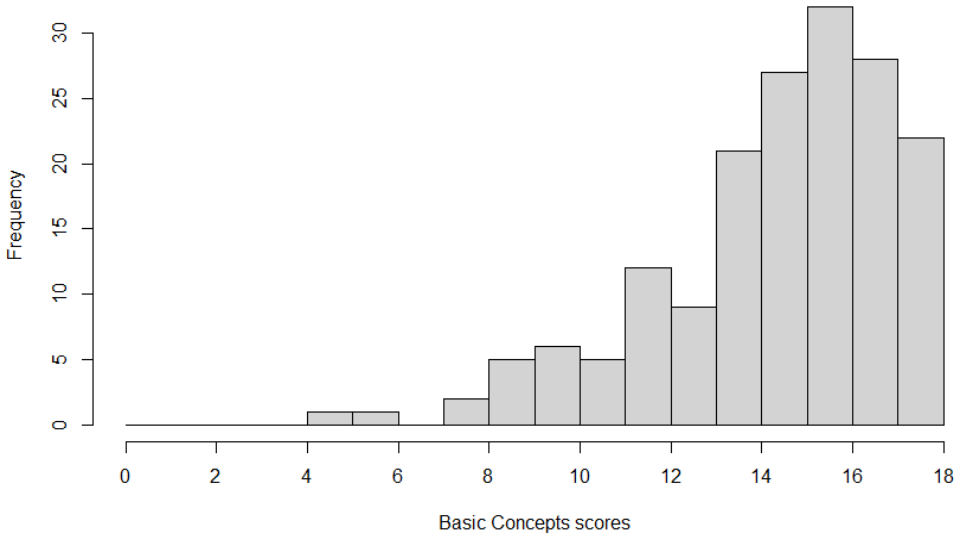


Figure 6: Frequency distribution of CELF Preschool-2 UK 'Basic Concepts' scores at endline (EYPP-eligible subsample)

Subgroup analysis 2: EAL

The results of the EAL subgroup analysis are displayed in Table 31 below. There is no evidence that the Concept Cat intervention has a differential impact on EAL students. The model run on the EAL subsample found no association between assignment to treatment and both CELF Preschool-2 UK 'Basic Concepts' scores and EY Toolbox ENA at endline, as indicated by the marginal effect sizes of 0.07 and 0.066, accompanied by wide CIs. In relation to the theory of change, therefore, this result potentially indicates that the proposed core mechanism of action for the intervention does not hold up for the EAL subgroup. As discussed in the 'IPE results' section below, the 'Focus children' aspect of delivery was frequently neglected by settings. This oversight likely had a negative impact on children with significantly lower levels of conceptual vocabulary development prior to the intervention, potentially affecting EAL children disproportionately. The overall effect sizes from the interaction models presented in Appendix F are similarly insignificant.

The interaction models observed no statistically significant differential impact of being EAL on the treatment effect. The raw regression outputs in Appendix F suggest that there is substantial uncertainty surrounding the coefficient on the

interaction term with wide bootstrapped CIs. It should be noted, however, that the detectability of an effect on the EAL subgroup is limited by the small sample sizes, on account of the small number of children classed as EAL in the trial.

Table 31: EAL subgroup analysis

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	109 (19)	14.27 (13.54–14.99)	103 (20)	13.88 (13.27–14.5)	212 (109; 103)	0.07 (-0.14–0.29)	0.52
EY Toolbox ENA	99 (29)	31.1 (28.29–33.92)	99 (24)	28.06 (25.41–30.71)	198 (99; 99)	0.066 (-0.16–0.30)	0.54

Subgroup analysis 3: SEND

A final subgroup analysis was conducted to explore Concept Cat's impact on children classed with SEND status. While this subgroup analysis found no evidence of SEND status having an impact on the intervention's efficacy in either the subsample model or the interaction model, the analysis was severely underpowered due to the low number of children classified as SEND in the trial (n=60). We therefore, advise caution in interpreting the results and have moved the analysis tables to Appendix C.

Additional analyses and robustness checks

Sensitivity analysis 1: Accounting for censoring in the primary outcome

As outlined in the 'Methods' section, we assessed the ceiling effect in the primary outcome using three different approaches. Figure 3 presents a histogram of the endline scores for the analytical sample, which immediately highlights a pronounced ceiling effect. This observation is further supported by the information presented in Table 32 below, where the strong negative Pearson's coefficient indicates significant skewness. Furthermore, at 58%, the proportion of children scoring within 1 SD dramatically exceeds the 25% threshold derived from Uttl (2005). As such, further sensitivity analysis to account for this ceiling effect was deemed necessary.

Table 32: Measures of ceiling effect in CELF Preschool-2 UK 'Basic Concepts' endline scores

Coefficient of skewness	Percentage of sample within 1 SD of the max.
-0.78	58%

The observed ceiling effect indicates that the CELF Preschool-2 UK may have been unfit for purpose, given how the measure's maximum score is unable to capture the true upper range of ability levels among children who completed the test. This makes it impossible to differentiate between children scoring at the upper end of the scale, rendering the accurate estimation of an effect size problematic. Therefore, as discussed in the 'Methods' section, the impact of the Concept Cat intervention on the primary outcome was re-estimated using a Tobit regression model as an additional sensitivity analysis, accounting for the censoring of the 'true' value of the latent construct of early conceptual vocabulary. The covariates and multilevel structure of this model were otherwise identical to that of the primary analysis model.

The results of this Tobit model are displayed in Table 33 below, with information used to calculate effect size presented in Table 34. The effect size observed from this model (0.23) is larger than that of the main primary analysis model (0.18) and demonstrates a similar level of statistical certainty ($p = 0.01$). This analysis in the presence of a ceiling effect is indicative of three month's additional progress in the primary outcome, supporting the theory the initial primary analysis may have underestimated the intervention's impact. It should be noted, however, that these results should be interpreted with caution, given the magnitude of the observed ceiling effect, which the Tobit model can only partially mitigate against.

Table 33: Primary analysis Tobit model (accounting for censoring)

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	460 (67)	15.57 (15.32–15.82)	442 (71)	14.94 (14.67–15.21)	902 (460; 442)	0.23 (0.15–0.31)	0.01

Table 34: Primary analysis Tobit model effect size estimation

Outcome	Unadjusted differences in means	Adjusted differences in means	Intervention group		Control group		Pooled variance
			n (missing)	Variance of outcome	n (missing)	Variance of outcome	
CELF Preschool-2 UK 'Basic Concepts' subtest	0.63	0.65	460 (67)	7.7	442 (71)	8.4	8.07

Sensitivity analysis 2: Controlling for age in the primary analysis model

As discussed in the 'Methods' section, further sensitivity analysis was conducted, where child age was introduced to the primary analysis model as a control variable. This allowed for the effect size to account for potentially heterogenous effects of Concept Cat across different age groups. The results of this analysis are presented in Table 35 below. The effect size and associated bootstrapped CIs demonstrate marginal change to those observed in the main primary analysis, suggesting that the primary analysis was not biased by age.

Table 35: Primary analysis with control for age

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	460 (67)	15.57 (15.32–15.82)	442 (71)	14.94 (14.67–15.21)	902 (460; 442)	0.17 (0.04–0.30)	0.01

Implementation and Process Evaluation

Compliance

Summary of findings

- Pupil attendance was collected for 72% of children (382 out of 527 children) in the intervention group of which all children from PVI settings met the compliance criteria (attended 15+ hours per week) and 98% of children (275 out of 282 of children) from Maintained settings met compliance criteria.
- Compliance was considered high for attendance at training across intervention sessions and moderate in terms of the number of words taught across the duration of the programme (30+ words).
- There were found to be low levels of compliance for attendance at group supervision sessions with 62% of settings (28 out of 45 settings) rated as low compliance.

In the impact evaluation, compliance was defined as the number of sessions attended by each pupil over the 30-week delivery period, along with the number of words taught in the setting over the same period. In addition, the IPE included two additional measures of compliance: i) attendance at training; and ii) attendance at group supervision meetings. These measures of compliance are discussed below.

Pupil attendance

Pupil attendance, for those who attended the setting for 15+ hours per week, was collected, at the end of the programme, for 72% of participating children (382 out of 527 in total; 282 out of 403 children from 32 Maintained settings [70%] and 100 out of 124 children from 12 PVI settings [80%]) within the intervention group (see Table 36). Of those for whom data was collected, eight children (out of 282 children) from four Maintained settings did not meet compliance criteria (i.e. did not attend for 15+ hours per week). Thus, where data was collected, 98% of children from Maintained settings (275 out of 282 children) met the compliance criteria and 100% of children from PVI settings (75 out of 75 children) met the compliance criteria.

While this indicates high compliance, caution should be taken when interpreting this data due to the low numbers of participating children for which attendance data was collected.

Table 36: Pupil attendance in the intervention group across setting type

Pupil attendance	Maintained n (%)	PVI n (%)
15+ hours attendance	275 (98)	75 (100)
<15 hours attendance	8 (2)	0 (0)

Words taught

As reported in the impact evaluation, word compliance data showed that 31 out of the 44 intervention settings for whom the evaluation team had complete data[13] (i.e. taught 30 words before endline testing was complete). The data shows that, across the intervention period, there were similar levels of compliance across setting type. Although there were slightly higher levels of recorded compliance among PVI settings (nine out of 12 settings) compared to Maintained settings (22 out of 32 settings). Given the small number of PVI settings in the sample, the evaluation team could not draw any firm conclusions from this finding (see Table 37).

[13] One setting had missing data.

Table 37: Number of words taught across setting types[a]

Number of words taught	Maintained n (%)[b]	PVI n (%)[b]
30 words taught	22 (69)	9 (75)
<30 words taught	10 (31)	3 (25)

[a] *N=44 (32 Maintained, 12 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*

Attendance at training

This section addresses: IPERQ4. 'Have practitioners attended mandatory training?'.

All EYPs in intervention settings were expected to attend an online training session. Lead practitioners were expected to attend a three-hour session while other EYPs were expected to attend a condensed, one-hour session.

Attendance at training was provided by the delivery team who collected this data on a session-by-session basis. Settings where the majority of EYPs attended a training session (i.e. 80% to 100% attendance) were rated as having a high level of compliance with regard to training. As agreed with the delivery team, settings with less than 80% attendance above 70% of EYPs attended training were classed as medium compliance, and those with below 70% staff attendance at training were classed as low compliance.

Table 38: Attendance at training across setting types[a]

Attendance at training	Maintained n (%)[b]	PVI n (%)[b]
80 to 100%	29 (91)	11 (84)
70 to 80%	1 (3)	0 (0)
Below 70%	2 (6)	2 (15)

[a] *N=45 (32 Maintained, 13 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*

As can be seen in Table 38, the majority of settings facilitated all lead practitioners and most other EYPs to attend training (40 out of 45 intervention settings) meaning that settings were highly compliant with regard to training attendance. Over the training period, the developers ran training sessions with an additional 162 practitioners. The additional training sessions were run for additional EYPs who wanted to receive the training and for staff who were new to the setting.

The endline survey asked setting managers if they found it difficult to release staff for training. Most setting managers (17 out of 21) stated that they did not find it difficult to release staff for training. Setting managers reported on the ease of releasing staff for training because of the short length of the training and the variation of the training times on offer:

> *Absolutely fine, there was a different schedule for timings of training so we just made sure that staff members who maybe are part-time just accessed this training that they could do in their work time. We got cover with that, between us all we covered it and made sure that was accessible really. It was very manageable actually to be honest, it worked really well.* (EYP interview, I003, Maintained)

In the interviews, all interviewees indicated that attendance at training was done within work time and was organised based on staff-child ratios at specific times of the day and/or already available cover, with the exception of one EYP who talked about their Maintained setting scheduling their training as a group at the end of the school day. Only one EYP mentioned, in their interview, a staff member's difficulty using technology as a barrier to attendance.

Attendance at group supervision

In addition to the mandatory training, EYPs were asked to attend five group supervision sessions over the course of the programme. Attendance at group supervision (Table 39) was collected by the delivery team as part of their monitoring data and shared with the evaluation team. Settings between four and five sessions were attended by at least one member of staff (i.e. 80 to 100% attendance) were rated as having a high level of compliance with regard to supervision. Settings with less than 80% attendance but more than 70% (three out of five) sessions were attended by at least one

EYP were classed as medium compliance and those with below 70% attendance at group supervision (less than three sessions) were classed as low compliance.[14] The majority of settings (17 out of 32 Maintained and 11 out of 13 PVI settings) were classed as having low compliance with regard to group supervision.

Table 39: Attendance at group supervision by setting type[a]

Attendance at group supervision	Maintained n (%)[b]	PVI n (%)[b]
80 to 100%	12 (38)	2 (15)
70 % 80%	3 (9)	0 (0)
Below 70%	17 (53)	11 (85)

[a] N=45 (32 Maintained, 13 PVI).
[b] Percentages may not sum to 100 due to rounding.

Training and support

Summary of findings

- Training was well received and EYP's felt the training prepared them well for implementation although some interviewees felt that all EYPs would benefit from the three-hour training.
- Coaches were highly valued and considered a key part of implementation although some settings sometimes struggled to accommodate the Concept Cat coach visits.
- While attendance at the group sessions was low, primarily due to staffing difficulties, EYPs felt they were a valuable opportunity to share ideas relating to implementation of Concept Cat.
- Settings found the resources provided were very helpful although sometimes there were difficulties in finding resources for certain concepts.

In addition to attendance at training, the IPE was designed to explore:

> IPERQ4. 'To what extent have training and resources supported practitioners' ability to effectively teach Concept Cat?'

This section reports findings from the endline EYP survey and interview data relating to practitioners' experiences and perceptions of training and support, including programme resources, Concept Cat coaches, and group support sessions.

Training

Observation data revealed that the training covered the key STAR elements of the Concept Cat programme (Teach, Activate, and Review) and emphasised the need for all parts of the programme to be delivered. Lead Concept Cat practitioners attended the three-hour training, and most other EYPs attended the one-hour training. In the interviews, some Lead Concept Cat practitioners reported attending both training sessions (the one-hour and the three-hour training).

In the endline EYP survey, respondents were asked a series of five-point Likert-scale questions relating to the Concept Cat training.

As can be seen in Table 40 below:

- three-quarters of EYPs (55 out of 73 respondents) strongly agreed that the training they received gave them enough information to be able to implement Concept Cat within their setting;

[14] This definition of low, medium, and high compliance was agreed with the delivery team.

- over three-quarters of EYPs (57 out of 73 respondents) strongly agreed that the Concept Cat training highlighted the key concepts of the STAR approach 'Teach, Activate, and Review';
- EYPs were particularly positive about the Concept Cat trainers with 60 out of 73 respondents strongly agreeing that they answered questions in a clear and helpful manner.

Table 40: EYPs views about the Concept Cat training (n=73)

	Strongly agree n (%)[a]	Agree n (%)[a]	Neither agree nor disagree n (%)[a]	Disagree n (%)[a]	Strongly disagree n (%)[a]
The training I received gave me enough key information to be able to implement Concept Cat within my setting	55 (75)	17 (23)	1 (1)	0 (0)	0 (0)
The training highlighted the key concepts of 'Teach, Activate, and Review'	57 (78)	14 (19)	2 (3)	0 (0)	0 (0)
The trainers answered questions in a clear and helpful manner	60 (82)	12 (16)	1 (1)	0 (0)	0 (0)

[a] *Percentages may not sum to 100 due to rounding.*

The EYPs, in their interviews, were also very positive about the training, stating that it explained the programme well, and they felt ready to deliver Concept Cat within their setting:

> *Everything was explained at the beginning and explained well.* (EYP interview, I004, PVI)

> *It actually left me feeling, like, raring to go.* (EYP interview, I006, PVI)

While the training was generally considered sufficient, the one-hour training was considered to be quite short, and some interviewees expressed the view that, for their qualified members of staff, the three-hour training would have also been beneficial. The longer (three-hour) training was seen as important since it equipped EYPs to understand more about the 'why' of the programme, as well as support other staff (who had taken the one-hour training) in their implementation:

> *After that if they had any questions they were able to ask me, because I'd had that extra training.* (EYP interview, I003, Maintained)

Two interviewees indicated that they had two members of staff take the three-hour training to provide them with more information about implementation of the programme and to provide additional implementation support within their setting:

> *I did the hour one first because [practitioner name] was going to take the lead because I'd got other things going on. Then we decided it's better if two of us did it so we both did it together.* (EYP interview, I005, PVI)

The EYP endline survey results also suggested that a third of practitioners felt that they would not have been able to implement the programme without the training (25 out of 73 respondents).[15] One interviewee highlighted how valuable the training was compared to having the Word Aware 2 book (Parsons and Branagan, 2016) alone (which their setting had previously acquired):

> *Our previous head bought it [the Word Aware 2 book] and said, 'This will be really good', and we all looked at it and was like, 'Oh yeah, right, okay', and then it got put in a cupboard. And it was like, 'Yeah it probably is a really good thing, but what do I do with it? Where am I coming from? Have I got time for it?'…And it is because from the training we know why, we know it's easy, we've got the resources.* (EYP interview, I002, Maintained)

[15] As this was a reversed question (i.e. it prompted respondents to answer using the scale in the opposite way), further analysis was carried **out** to identify whether respondents had selected the same answer for all the statements. Half (14 respondents) selected this response for all the statements. It is therefore, possible that the responses recorded for this question are not a true reflection of the respondents' thoughts.

Concept Cat coaches

Concept Cat coaches were embedded in the implementation support provided as part of the Concept Cat programme. Settings were expected to receive six Concept Cat coach visits during the implementation year (i.e. approximately one visit each half-term). As seen in Table 41, most settings received all six Concept Cat coach visits (41 out of 45 settings).

Table 41: Number of Concept Cat coach visits

Visits	Maintained n (%)[b]	PVI n (%)[b]
Six sessions	29 (90)	12 (92)
Five sessions	3 (10)	0 (0)
Two sessions	0 (0)	1 (7)

[a] *N=45 (32 Maintained, 13 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*

In the endline survey, EYPs were asked 'How helpful did you find the Concept Cat coaches in supporting you to implement what you had learnt during the training?' The majority (59 out of 72 respondents; 82%) indicated that they found the Concept Cat coaches 'very helpful', 12 respondents indicated that they found them 'helpful', and one respondent stated that they found them 'neither helpful nor unhelpful'.

In the interviews, EYPs talked about valuing the support provided by the Concept Cat coaches. For instance, EYPs discussed their usefulness as a sounding board, or a valuable source of advice. This was particularly true at the start of implementation, although in the later stages, some EYPs stated that Concept Cat coach visits helped them to keep on track with implementation:

> *Definitely, at the beginning it was nice to see the Concept Cat coach showing us, and validating that we're doing it right, we're doing what we need to be doing.* (EYP interview, I007, PVI)

> *I think without that, where there's been wobbly weeks we probably would have said, 'Leave it'. But knowing that she's coming in and she's going to check up…It's been good for us to keep to it and for me to say, 'No, we have to do it'. (*EYP interview, I005, PVI)

Concept Cat coaches were also seen as accessible in terms of contact and valuable in understanding setting needs:

> *It's nice to talk through with her, but she's always there for us to email as well.* (EYP interview, I007, PVI)

> *She really is quite flexible and very understanding of, that the nature of nurseries does change all the time, so she's really good.* (EYP interview, I006, PVI)

However, two EYPs mentioned the difficulties of finding time for Concept Cat coach visits or of finding time to talk during visits due to the busy demands of working in an EY setting:

> *Just sometimes it's like she's wanting to ask questions and, 'We'll be with you in a minute', and we felt like, 'God I feel awful', brushing her off, which we weren't but it is hard when you've got a lot of children.* (EYP interview, I005, PVI)

Group support sessions

Another element of support offered by the programme was group support sessions. In total, five sessions were held throughout the programme and were intended for settings to share ideas around implementation of the programme. At least one EYP per setting was expected to attend. Eight out of 45 settings (six Maintained, two PVI) attended all five group support sessions, five Maintained settings attended four sessions, three Maintained settings attended three sessions, seven settings (six Maintained and one PVI) attended two sessions, 11 settings (eight Maintained, three PVI) attended one session and seven settings (one Maintained and six PVI) did not attend any group support sessions.

Most of the EYPs in their interviews (five out of seven EYPs) discussed attending at least one of these meetings. For a small number of settings, finding the time to do so was an issue and two EYPs reported that no-one in their setting had attended any of the meetings, primarily due to staffing difficulties:

> *There's been a couple of them where we've thought, 'We'll be alright, one of us can do it', and it's just impossible.* (EYP interview, I005, PVI)

Sometimes, when EYPs did attend, attendance overall was low, which was also discouraging:

> *One we signed into nobody else was there!* (EYP interview, I004, PVI)

Generally, however, when settings did manage to attend, these meetings were found to be helpful resources for sharing good practice and discussing ideas:

> *It's always good to hear how it's working in other settings and what was working well for somebody, and also to have an opportunity if you've got something that you're finding a bit more tricky to fit in. And nice to be able to offer help and support maybe to someone who's having a difficulty with something that you've either encountered and found a solution for or have just got some suggestions.* (EYP interview, I001, Maintained)

> *I feel like that's an extra training in itself without even being trained in a sense.* (EYP interview, I003, Maintained)

Resources

Settings were pleased with the resources provided by the programme:

> *The Picky Puppet is great, the bag, it was really lovely to open and they're all so organised and it was really nice to have that.* (EYP interview, I006, PVI)

> *It's massive, it's amazing and I've possibly used not even a slither of it because our children, they were the words that our children have identified but we've got an amazing set for the future.* (EYP interview, I004, PVI)

However, in the interviews, some EYPs, talked about how, as their setting progressed through the programme to Level 3 words, they increasingly found themselves having to be more resourceful in organising resources for programme delivery. This included printing off additional sets of words and finding resources for the word of the week. EYPs also discussed purchasing additional Concept Cats. It should be noted, however, that overall EYPs did not report finding this an unreasonable requirement:

> *It's a little bit different now obviously because we've moved on to the [L]evel 3 words, so I just need to make sure that we've got a full folder on our online system of every single activity to send home everything, so it's all there, it's just a case of printing if off, laminating it if we need to, cutting it out…Very much still easy really just to get, just so accessible definitely.* (EYP interview, I003, Maintained)

Fidelity

Summary of findings

- The use of a three- and five-day implementation schedule along with varied starting points for the level of words taught supported the needs of the children within the setting and supported implementation.
- The key elements of Concept Cat were implemented similarly across setting types.
- The Concept Cat programme was mostly implemented with fidelity and quality across setting types although the 'Activate' element of the STAR approach was implemented with medium fidelity.
- Implementation fidelity with focus children was low across setting types.
- Children and practitioners alike clearly enjoyed the Concept Cat programme, which was facilitated by the ease of implementation for most of the elements of the programme.
- The concept words sometimes proved to be a barrier for implementation, particularly the Level 3 words, which were harder to incorporate into practice.

This section addresses the following research questions:

> IPERQ1. How closely does the Concept Cat programme, as implemented in settings, follow the intended model (implementation fidelity), as outlined in the TIDieR framework including extended implementation for focus children? What are the barriers and facilitators to implementation and how do these differ, if at all, between setting type (PVI/Maintained)?
>
> **IPERQ4.** What is the quality of delivery (i.e. how well are different components of the intervention delivered; Education Endowment Foundation, 2024., p. 6)?

It does so by presenting findings from the endline EYP survey, monitoring data, and interview data. This is supplemented by data collected through the embedded observations, which sought to explore implementation of the key elements of the programme.

Patterns of implementation

As part of the programme, settings could choose whether to implement Concept Cat over a three-day week, a five-day week, or both, depending on setting need. Data from the endline EYP survey revealed that, across setting types, both implementation schedules were used (Table 42) suggesting that the options to implement over a three-day week and a five-day week was needed to support implementation of the programme. The five-day week schedule was most commonly used across both setting types.

Table 42: Days of implementation across setting type[a]

Days	Maintained n (%)[b]	PVI n (%)[b]	Total n (%)[b]
Three-day week	15 (24)	9 (14)	24 (36)
Five-day week	29 (43)	12 (18)	41 (62)
Both	1 (2)	0 (0)	1 (2)

[a] *N=66 (45 Maintained 21 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*

Levels of words taught

At the start of the programme, Concept Cat coaches carried out assessments of the children to understand, which level of words was a suitable starting point for the setting. This was the 'Select' element of the STAR approach. Settings could then move through the word levels as needed.

The endline EYP survey investigated, which levels of words were implemented across settings (Table 43). The data shows that, in most settings (37 out of 59), all three levels of words were implemented. The disparity of word across setting types demonstrates the variability of the programme to meet the needs of the children within the setting.

Table 43: Levels of words implemented according to setting type[a]

Levels of words	Maintained n (%)[b]	PVI n (%)[b]	Total N (%)[b]
Level 1	1 (3)	2 (10)	3 (5)
Level 1 and Level 2	9 (23)	2 (10)	11 (20)
Level 2	2 (5)	0 (0)	2 (4)
Level 2 and Level 3	2 (5)	1 (5)	3 (5)
Level 3	2 (5)	1 (5)	3 (5)
Level 1, Level 2, and Level 3	23 (59)	14 (70)	37 (63)

[a] *N=59 (39 Maintained, 20 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*

Implementation of key programme elements

Following the embedded observations in four settings (two Maintained, two PVI), a count analysis was performed on the observation data, collected over the three days, to look at implementation fidelity across all the main elements of the programme, as a whole and across setting types (Table 44). For each element, researchers marked whether they observed each area of implementation; the researchers marked the element as observed only if it was observed consistently through the whole of the activity taking place. All four settings were observed during the same stage of delivery in the Summer Term.

Table 44: Elements of Concept Cat observed during the three-day embedded setting visits[a]

Inputs	Implementation	Consistency measure[b]	Observed	
			PVI No. of settings	Maintained No. of settings
Teach 1	A sign/gesture is used to portray the word	Yes	2	2
	The symbol on the card is shown	No	2	2
	Children are encouraged to say the word	Yes	2	2
	The word song is used either sung by teacher or the video is used	No	2	2
Teach 2	The story is told using Concept Cat toy either by teacher or video (Concept Cat is a clever, clever cat…)	No	2	2
Teach 3	Activity is done with all children to demonstrate the word	No	2	2
Activate	The word used consistently during everyday activities	Yes	1	2
Quality	Practitioners use comments rather than questions	Yes	0	2
	The word used in sentences?	Yes	2	2
	Practitioners use word/non-word?	Yes	1	2
Review	Objects are put into a bag to demonstrate the word	No	0	2
	Picky Puppet is used to repeat words using objects	No	0	2
	Review activities are used to support word/example correspondence	No	1	2
Focus children	The story repeated?	Yes	0	1
	Practitioners ensure they are hearing the word during play	Yes	1	1
Environment[c]	The word is displayed in a prominent place?		2	2
	There are activities available where the word can be used		2	2
	There is a Concept Cat house with objects and pictures so children can retell the story		1	2
	A display poster with words, for families is in a prominent place		2	2

[a] *N=4 (2 Maintained, 2 PVI).*
[b] *The consistency measure refers to whether or not the programme element observed was required to be implemented once within the observed time period (denoted as 'No') or several times (denoted as 'Yes').*

[c] *Grey area refers to not a measure of consistency*

Teach

For the 'Teach' element of the STAR approach, all key strategies were observed across setting types. Researchers observed that practitioners and children were familiar with the Concept Cat stories and songs demonstrating that implementation had been taking place consistently:

> *Children are clearly excited about everything Concept Cat. (*Researcher notes, Embedded observation I002, Maintained)

> *Children [are] clearly familiar with the song and story about Concept Cat.* (Researcher notes, Embedded observation I004, PVI)

Activate

To establish the implementation of the 'Activate' element of the STAR approach, the observation looked for consistency of the word of the week being used during everyday activities. This was observed in both Maintained settings. In contrast, this was observed in only one out of the two PVI settings. The researcher commented in the observation notes that children were selected for the Concept Cat *'Teach section and then the items that were being used to express the meaning of the concept were put away'* (Embedded observation, I006, PVI).

Review

For the 'Review' element of the STAR approach, all strategies were used in Maintained settings. In the PVI settings, no 'Review' activities were observed, with the exception of 'Review' activities to support word/example correspondence in one of the two settings. Where this did take place, the researcher reported that activity was implemented for ten minutes, after, which some children lost interest. However, the monitoring data (reported below) suggests that fidelity of the 'Review' element was only moderate across both setting types.

Focus children

Observations of additional activities with focus children were limited across both setting types. In one Maintained setting, the researcher noted that they had been advised that the provision for the focus children took place in an afternoon (the observation took place in the morning). In one PVI setting, the researcher noted that:

> *All children are included equally [it is] not clear who the focus children are.* (Researcher notes, Embedded observation I004, PVI)

The data suggests that, across setting types, the additional activities for focus children may not be taking place. This aligns with the monitoring data (reported below), which suggests low implementation fidelity for focus children in that focus children did not receive the additional input that was intended. However, it should be noted that the additional activities carried out with focus children are not central to implementation of the programme hence, why they are not included separately to other children in the logic model. In both PVI settings, the researcher reported that SEND children were excluded from Concept Cat, although it is unclear if these children were also children selected to be focus children.

In their interviews, EYPs discussed difficulties in delivering Concept Cat as a small group intervention for focus children:

> *We might not have done that because with the focus children that we've got, if they were already in a play situation elsewhere in the room and we were able to incorporate that word into what they were already doing, then we would much rather do that than remove them from something and take them away from that.* (EYP interview, I001, Maintained)

> *I used to take out the group of focused children and that always seemed to be a bit tricky when we got a bit busier and things like that…[So], we said we'd make that we're all aware of who those focused children are, and in provision we can implement it and just do that little bit extra with them in provision. And they're getting the extra groups as well like at group time and before dinner. Still*

getting lots of [A]ctivate and [T]each throughout, but a bit extra within provision as well. (EYP interview, I003, Maintained)

EYPs also talked about choosing their focus children. Such choices were influenced not only by language needs but also pragmatic considerations, including attendance patterns and/or a nominated staff member for Concept Cat delivery:

So we looked and we had an idea who we thought we might choose, there's like a criteria chart that you follow. But then we kind of decided that it would be easier if they were in one staff's key group, apart from one little girl, the global delay little girl who we thought had to be in it from the start. Just because then if it's one person's ownership to do it, it's just easier. (EYP interview, I007, PVI)

Environment

Across all setting types there was evidence that key elements of the programme had been implemented within the setting environment during the embedded observation. For example:

The Concept Cat house is a large cat house/scratching post where all objects and Concept Cat is displayed, and children can interact with the objects. (Researcher notes, Embedded observation I002, Maintained)

Children were doing forest school and selecting sticks to do 'long/not long', the activity had been set up to support using the word. (Researcher notes, Embedded observation I003, Maintained)

Implementation quality and fidelity

Table 45: Quality and fidelity of implementation across setting types[a]

Key elements	Categories	Score		Overall Rating
		Maintained n (%)[b]	PVI n (%)[b]	
Quality:				
Teach	Delivery	186 (98)	67 (91)	high
Activate	Delivery	168 (89)	68 (91)	high
Review	Delivery	184 (97)	69 (93)	high
Families	Family sessions	109 (57)	24 (32)	low
Focus children	Selection of children	188 (99)	71 (95)	high
	Contact with families	113 (60)	51 (69)	low
	Delivery[c]	159 (84)	60 (81)	medium
Part-time children	Delivery	183 (97)	70 (94)	high
Fidelity:				
Teach	All part delivered	184 (97)	67 (90)	high
	All staff involved	170 (90)	65 (88)	high
Activate	Word wall and Concept Cat house	153 (81)	61 (82)	medium
	Comments used rather than questions[c]	167 (88)	65 (88)	high
Review	Delivered three times per week	154 (81)	60 (80)	medium
Families	Family poster	172 (90)	71 (95)	high
	Family symbol	176 (92)	71 (96)	high
Focus children	Story repeated	130 (68)	55 (74)	low
	Regular guidance	122 (64)	45 (60)	low
	Additional activities	137 (72)	57 (77)	medium
	Additional Picky Puppet sessions	126 (67)	53 (71)	low

[a] *N=45 (32 Maintained, 13 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*
[c] *Many of the logs were blank (i.e. not observed) for these categories so we are not confident that these scores are accurate and reflective of the practice taking place.*

The monitoring data[16] provided further detail about levels of implementation of Concept Cat across settings and setting type (Table 45). Monitoring data was collected by the delivery team over the implementation period (via the Concept Cat coach logs) and provided to the evaluation team at the end of the programme. The logs were completed by Concept

[16] NB: Monitoring data was collected by logs completed by the Concept Cat coaches and as such there is a potential risk of bias.

Cat coaches at each visit (six visits were intended). Therefore, fidelity and the quality of delivery could differ over time. Settings were regarded as delivering Concept Cat with high implementation quality/fidelity if settings achieved a score of between 4.5 and 6 for a category in the Concept Cat coach logs (total range between 85% and 100%), medium quality/fidelity if they scored between 3 and 4 (total range between 75% and 84%) and low quality/fidelity if settings scored below 3 (total <74%). Each category was summed and divided by the total number of sessions that were observed by Concept Cat coaches across settings (maximum 189 Maintained, maximum 74 PVI)[17] to give a quality/fidelity score.

In line with the observation data, the 'Teach', 'Activate', and 'Review' elements of the STAR approach in the programme were shown to be implemented with quality in both Maintained and PVI settings. Although delivery was of high quality, it was not always delivered consistently across the programme. As such, while 'Teach' was delivered with high fidelity, 'Activate' and 'Review' were only implemented with medium fidelity across setting types.

As indicated by the observation data, implementation with the focus children was of medium or low quality and fidelity across setting types. While focus children were still receiving the intervention, practitioners found it more difficult to include extra time for these children. The family element in terms of arranging sessions to explain the Concept Cat programme with families was also low although settings did engage parents in ways that best suited the parents from their settings. Given parental engagement was high and additional activities for focus group children are not part of the logic model this is not deemed to be a threat to implementation fidelity rather, it is an area of consideration for the delivery team.

In the endline survey, EYPs were asked specifically about the ease of implementation of specific elements of the Concept Cat programme using Likert-scale questions.

Table 46: Ease of implementation of Concept Cat in intervention settings (N=69)

Implementation	Rating[a]				
	Very easy n (%)	Quite easy n (%)	Sometimes easy / sometimes difficult n (%)	Quite difficult n (%)	Very difficult n (%)
Using a sign/gesture to portray the word	41 (59)	16 (23)	12 (17)	0 (0)	0 (0)
Telling the Concept Cat story	47 (68)	11 (16)	9 (13)	2 (3)	0 (0)
Providing an activity to demonstrate the word	36 (52)	22 (31)	10 (14)	0 (0)	1 (1)
Using the word consistently during play activities	41 (59)	21 (30)	6 (9)	1 (1)	0 (0)
Ensuring all practitioners were consistently using the word	29 (42)	33 (47)	6 (9)	1 (1)	0 (0)
Finding objects to put in the word bag	27 (39)	25 (36)	16 (23)	0 (0)	1 (1)
Using Picky Puppet to repeat the word using the object	40 (57)	19 (27)	8 (11)	2 (3)	0 (0)
Engaging children with Concept Cat activities	50 (72)	13 (18)	5 (7)	1 (1)	0 (0)
Engaging focus children with Concept Cat activities	36 (52)	17 (24)	14 (20)	1 (1)	1 (1)
Engaging families in using the word at home	10 (14)	21 (30)	21 (30)	14 (20)	3 (4)

[a] *Percentages may not sum to 100 due to rounding.*

As can be seen in Table 46, for most settings, each part of implementation was deemed to be relatively easy. For example:

- Over two-thirds of EYPs (47 out of 69 respondents) reported finding 'Telling the Concept Cat story' 'very easy', 11respondents stated it was 'quite easy', and 9respondents stated it was 'sometimes easy/sometimes difficult'.
- @Over half of EYPs (36 out of 69 respondents) reported finding 'Providing an activity to demonstrate the word' 'very easy', 22 respondents stated it was 'quite easy', and ten respondents stated it was 'sometimes easy/sometimes difficult'.

[17] NB: Numbers are not equal to six times (32 Maintained, 13 PVI) as not all settings had six sessions observed.

- Two-thirds of EYPs (41 out of 69 respondents) reported finding 'Using the word consistently during play activities' 'very easy', 21 respondents stated it was 'quite easy', and six respondents stated it was 'sometimes easy/sometimes difficult.'

- Over one-third of EYPs (29 out of 69 respondents) reported finding 'Ensuring all practitioners were consistently using the word' 'very easy', over a third (33 out of 69 respondents) stated it was 'quite easy', and six respondents stated it was 'sometimes easy/sometimes difficult'.

- Over half of EYPs (40 out of 69 respondents) reported finding 'Using Picky Puppet to repeat the word using the object' 'very easy', 19 stated it was 'quite easy', and eight respondents stated it was 'sometimes easy/sometimes difficult'.

- 'Engaging families in using the word at home' was an element that EYPs (14 and 3 out of 69 respondents, respectively) reported as being 'quite difficult' or 'very difficult' to implement and was the item reported by the highest number of EYPs as being 'sometimes easy/sometimes difficult' (21 out of 69 respondents).

More information on engaging parents can be found in the section on parental engagement. Respondents also reported that 'Engaging focus children with Concept Cat activities' (16 out of 69 respondents) and 'Finding objects to put in the word bag' (17 out of 69 respondents) was 'sometimes easy/sometimes difficult', 'quite difficult', or 'very difficult'. Further analysis was performed to look at these elements across setting types (Table 47).

Table 47: Difficulties in engaging families, engaging focus children, and findings objects to put in the word bag, across setting type[a]

	Maintained N (%)[b]	PVI N (%)[b]	Total N (%)[b]
Engaging focus children with Concept Cat activities	12 (26)	4 (17)	16 (23)
Engaging families in using the word at home	20 (43)	18 (78)	38 (55)
Finding objects to put in the word bag	12 (26)	5 (22)	17 (25)

[a] *N=69 (46 Maintained, 23 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*

The data in Table 47 suggests that EYP across both setting types reported difficulties with these three elements of the programme, with EYPs from Maintained settings (26%, 12 out of 46 respondents) reporting that they had difficulties in engaging focus children at a higher rate than EYPs from PVI settings (17%, 4 out of 23 respondents). In addition, EYPs from PVI settings (78%, 18 out of 23 respondents) more frequently reported having difficulties in engaging families to use the word at home, compared to EYPs from Maintained settings (43%, 20 out of 46).

In the endline survey, EYPs were also asked: 'Did you give additional support to your focus children?' A total of 67 out of 69 respondents replied 'yes' to this question. Two respondents replied 'no' to this question (both in Maintained settings). One respondent gave a reason for this and stated:

> *The timetable is packed. I took over half way through the year on a maternity leave and focus children had not really been set up. One went for an extended holiday to Pakistan. The focus children chosen were pitched too low; selective mutism, etc.* (Practitioner endline survey, Maintained)

These results are surprising, given the monitoring data presented. It is possible that while settings did feel that they gave additional support to focus children, this was not done as consistently as the programme intended. The interview data suggests that this may have been the case, since a small number of EYPs discussed not always delivering to the focus children as intended:

> *We possibly have tweaked the programme a little bit, because for the focus children it talked about taking them to the house and getting them to retell the story there maybe. We might not have done that because with the focus children that we've got, if they were already in a play situation elsewhere in the room and we were able to incorporate that word into what they were already doing, then we would much rather do that than remove them from something and take them away from that.* (EYP interview, I001, Maintained)

Barriers and facilitators of implementation

Training, resources, and coaches

In their interviews, EYPs discussed various factors, which facilitated implementation. As discussed above, the training was valued by EYPs, as were the Concept Cat coaches and the resources provided.

> *From our point-of-view as practitioners it's simple to understand and to implement.* (EYP interview, I001, Maintained)

> *With all the resources being met we could just run with it. It was actually quite powerful having the resources made, you didn't have to get up and, 'I've got to have them done and everything'. So the way that the training was, the timing and the resources, yeah absolutely ideal.* (EYP interview, I002, Maintained)

Enjoyment of the programme

The children's overall enjoyment of the different programme elements was also cited as a helpful factor in implementation:

> *We love Picky Puppet, they love him, they ask for him to come out because we make it silly and a game, but he pinches things and they laugh at him.* (EYP interview, I001, Maintained)

Children's enjoyment of the programme also supported parental engagement with the programme (see 'Parental engagement' section below).

Setting routines

A small number of EYPs indicated that the programme fitted well with their setting's established routine. The routine and repetition of the programme were both seen as beneficial for the children's engagement:

> *It's fitted in really well with a day nursery's timetable and ethos I think.* (EYP interview, I004, PVI)

> *The children then get to know that routine, they like the routine, they understand it…just having that consistent approach.* (EYP interview, I002, Maintained)

This also fits with the previous findings that showed that the alternative implementation schedules (five-day week and three-day week) along with the levels of words implemented to meet the children's needs, supported implementation. Consequently, the data suggests that the implementation schedules aligned well with the routines in both Maintained and PVI settings.

Child attendance and staff turnover

In contrast, child attendance patterns were seen as a barrier to implementation:

> *Like I say, even our full timers are three days or four days at the most, there's no one who does five. So, we can't deliver it like a school, we can't say, 'Every afternoon we will deliver the first part, the next part, the next part', because that wouldn't happen.* (EYP interview, I004, PVI)

> *There's two-and-a-half day patterns at the beginning of the week, two-and-a-half day patterns at the end of the week, AM and PM patterns. Everybody attends at different times, so obviously making sure everybody gets that teaching and learning as well is quite tricky to manage sometimes.* (EYP interview, I003, Maintained)

Staff turnover could also be a barrier to implementation, particularly during the 'Activate' phase of the STAR approach in the programme:

> *I might have had a few weeks where it wasn't quite as—the activation part of it wasn't quite as strong. Because you do obviously the [T]each element and the [R]eview element, but that's through the day when we had the new member of staff that didn't quite understand, hadn't had the training, weren't quite onboard with it the same…That was a bit of a challenge barrier to doing Concept Cat and it's not unknown for us to have a high turnover of staff.* (EYP interview, I007, PVI)

Words

One key element of Concept Cat, which proved problematic for some settings was related to the Concept Cat words. This was in terms of both resources—which was particularly apparent when settings progressed to Level 3 words where fewer resources were provided—and in terms of the actual words themselves (words mentioned in both these contexts included 'dark', 'narrow' and 'next'):

> *The hardest part was probably putting together the Word Bag, trying to think of something that was, and I still struggle to remember because I've shoved all this stuff in the bag and, 'My God, what is that concept? I can't quite remember', that's really tricky trying to find the resources and the time to think, 'Well that's a good one for that particular word'.* (EYP interview, I006, PVI)

> *I can't think of any off the top of my head, but depending on what the word was some of them were a bit more tricky to just have that constant conversation around, they cropped up less often in general practice so we use them less often.* (EYP interview, I001, Maintained)

Two EYPs discussed that, in the videos, the words were too fast for the children to focus on, which meant that they had to repeat them to the children or they chose not to use this element. The videos are not, however, a key aspect of the programme.

Adaptations

Summary of findings

- Settings made some acceptable adaptations to the programme including adapting the suggested activities in line with the resources the settings had as well as implementing the programme to meet the needs of the children.
- Some unacceptable adaptations were captured in relation to not implementing key aspects of the programme such as additional delivery to focus children, which aligns with the monitoring data.
- Some unacceptable adaptations were made in relation to changing key elements of the programme such as the implementation of non-Concept Cat words.
- For all adaptations that were made, Concept Cat coaches played a key role in ensuring the settings were aware of the correct implementation of the programme elements.

This section identifies acceptable and unacceptable adaptations to the programme to answer:

> IPERQ2: What, if any, adaptations have been made to the programme during implementation? Why were they made? What do they look like?

It does so by presenting findings from the observation data, which sought to uncover if any adaptations had been made to the programme during implementation, the EYP endline surveys and interviews, which also asked about any adaptations settings may have made in their delivery of the programme.

Acceptable adaptations

The observation data showed that the Concept Cat programme was adapted according to setting needs. The only adaptations observed were within the 'acceptable adaptation' guidelines as discussed below.

Whole class versus small group delivery

Researcher notes in the observation data suggests that Concept Cat activities were carried out in both whole class and small group delivery. In one PVI setting, the researcher also reported that:

> *[The] practitioner commented on the struggle to implement with a wide range of age groups, SEND and different children for [each] session. Strategies included doing the [T]each activities every day, separating children*. (Researcher notes, I004, Maintained)

In Maintained settings, the Concept Cat activities were more generally observed as a whole-class activity with small group repetition for part-time children to cover varying attendance patterns.

Suggested activities

Across setting types, there was evidence of the activities suggested by the Word Aware 2 book (Parsons and Branagan, 2016) being changed to meet the needs of the settings. For example, in one setting the researcher reported that:

> *[The] practitioner reports that the activity for 'enough' is to draw a sheep and give them 'not enough' to fill it in. They do not have the supplies for this so are improvising an activity. Children [are] given envelopes with little bits of paper to put in the envelopes. Children [are] asked 'do you have enough?'…They are given more until they have enough.* (Researcher notes, I005, PVI)

In another setting, the researcher reported:

> *The pictures for the Concept Cat book are not taken from home [and instead taken at school] as [EYP reports] this [sending pictures into the setting] is not done in this community.* (Researcher notes, I002, Maintained)

To more fully understand any adaptations settings had made to programme implementation, EYP respondents to the endline survey were asked a number of Likert-scale questions about changes to the programme, which were designed to understand if settings had made either acceptable or unacceptable adaptations to the programme during delivery. Table 48 shows the reported programme adaptations. Unacceptable adaptations were defined as those adaptations, which meant that key elements of the programme were not implemented (highlighted in orange) or those adaptations resulted in changes to key elements of the programme, which were essential for implementation fidelity (highlighted in green). Rows not highlighted in the table are adaptations deemed acceptable after discussion with the delivery team.

Table 48: Reported adaptations to Concept Cat delivery (N=69)

Adaptations	Reported frequency				
	Never n (%)[a]	A few times n (%)[a]	Some of the time n (%)[a]	Most of the time n (%)[a]	All of the time n (%)[a]
Changed the order in which you taught the words each week	25 (36)	23 (33)	14 (20)	1 (1)	6 (9)
Used additional words in replacement of Concept Cat words	54 (78)	9 (13)	2 (3)	1 (1)	3 (4)
Used the 'Teach' element more than once a week	0	11 (16)	16 (23)	25 (35)	17 (25)
Used a different teach activity to that identified in the book	17 (25)	16 (23)	14 (20)	6 (9)	9 (13)
Used gestures during everyday activities to portray the word	1 (1)	3 (4)	10 (14)	25 (35)	30 (43)
Used a Concept Cat house wall display	2 (3)	3 (4)	6 (6)	11 (16)	49 (71)
Took pictures of children doing the activities for a Concept Cat book	1 (1)	3 (4)	8 (12)	16 (23)	41 (59)
Used additional Concept Cat wall displays	9 (13)	9 (13)	16 (23)	8 (12)	27 (39)
Spoke with the families of focus children more often than non-focus children's families	4 (6)	10 (14)	22 (32)	19 (28)	14 (20)
Repeated the Concept Cat story with focus children	1 (1)	4 (6)	13 (19)	21 (30)	30 (43)
Used the Picky Puppet and word bag element with focus children twice a week	2 (3)	6 (9)	10 (14)	16 (23)	28 (41)
Had a plan for part-time children to access all elements of Concept Cat each week	1 (1)	2 (3)	2 (3)	21 (30)	30 (43)

[a] *Percentages may not sum to 100 due to rounding.*

Most settings indicated they only used acceptable adaptations, which was in line with the observation data. However, a small number of respondents indicated that they had implemented some unacceptable adaptations within their setting, which meant that they were not consistently implementing key elements of the programme (Table 48):

- Repeating the Concept Cat story to focus children was reported by 18 out of 69 respondents as 'never' being implemented (one respondent), only being implemented 'a few times' (four respondents), or only being implemented 'some of the time' (13 respondents).

- Speaking more to focus children families than non-focus group families was reported by 36 out of 69 respondents as 'never' being implemented (four respondents), only being implemented 'a few times' (ten respondents), or only being implemented 'some of the time' (22 respondents).

- The use of gestures being used in everyday activities was reported by 14 out of 69 respondents as 'never' being implemented (one respondent), only being implemented 'a few times' (three respondents), or only being implemented 'some of the time' (ten respondents).

- Having a plan for part-time children was reported by five out of 69 respondents as 'never' being implemented (one respondent), only being implemented 'a few times' (two respondents), or only being implemented 'some of the time' (two respondents).

In contrast, the monitoring data, shows low quality of implementation in terms of delivery to focus group children, a high quality of delivery to part-time children, and low fidelity with regard to repeating the story with focus children. This discrepancy between observation, monitoring, and endline EYP survey data potentially indicates that adaptations were more common at the start of the programme, which were then corrected by the Concept Cat coaches during their visits to the setting. However, it should also be noted that the survey findings are based on self-reported data.

In terms of unacceptable adaptations, in which key elements of the programme were changed, respondents reported (Table 48) the following instances:

- Using additional words in place of Concept Cat words was reported by some EYPs (15 out of 69 respondents). Nine respondents stated this happened 'a few times', two respondents reported this happened 'some of the time', one respondent reported it having happened 'most of the time', and three respondents reported it happening 'all of the time'.

- The use of the Concept Cat house as a wall display, rather than a physical display, was reported by EYPs (69 out of 71 respondents) as having happened 'a few times' (three respondents), happened 'some of the time' (six respondents), having happened 'most of the time' (11 respondents), or 'all of the time' (49 respondents).

The monitoring data showed that fidelity of the Concept Cat house was delivered with medium fidelity (in 12 out of 45 settings). The monitoring data also captured when settings were implementing words other than the Concept Cat words (in three out of 45 settings). In both instances where key elements of the programme were being adapted, the Concept Cat coaches commented that EYPs were advised of the correct implementation, and this was acted upon. Thus, the Concept Cat coaches played a key role in ensuring the programme was implemented as intended particularly at the start of the programme.

Where respondents indicated that they did make adaptations, responses were aggregated and analysed at the setting level (Table 49). The data shows similarities between Maintained and PVI settings in the adaptations made.

A small number of Maintained settings reported that they less regularly had a plan for part-time children to access all elements of the programme. A possible explanation for this could be due to the nature of the setting provision (e.g. longer hours) in PVI settings, which made implementation for part-time children easier in PVI settings compared to Maintained settings. However, monitoring data (see Table 44Table 44:) showed that the quality of delivery for part-time children was high in both Maintained and PVI settings suggesting part-time children received the programme as intended.

Table 49: Unacceptable adaptations by setting type[a]

Unacceptable adaptations	Maintained n (%)[b]	PVI n (%)[b]
Used gestures during everyday activities to portray the word	10 (22)	4 (17)
Spoke with the families of focus children more often than non-focus children's families	26 (56)	10 (44)
Repeated the Concept Cat story with focus children	13 (28)	25 (22)
Used the Picky Puppet and word bag element with focus children twice a week	9 (20)	4 (17)
Had a plan for part-time children to access all elements of Concept Cat each week	5 (11)	0 (0)

[a] *N=69 (46 Maintained, 23 PVI).*
[b] *Percentages may not sum to 100 due to rounding.*

The interview data supported the survey and observation data; most adaptations mentioned by EYPs were in line with those allowed by the programme. This included moving the order the words were taught in to accommodate other planned delivery, or changing from the five-day schedule to the three-day schedule due to time limitations:

> *If we're learning something in maths say for instance, like if we're learning the word 'more' and then the Concept Cat word is 'less' for that week, just making sure that we swap the words around on the list just so that it doesn't confuse things too much.* (EYP interview, I003, Maintained)

> *The only changes that I ever make is, as I said this week I'm using the three-day programme rather than the five to fit in with the fact that we're teaching Holy Week and the fact that afternoon nursery aren't here on Friday afternoons so they would have missed out on one bit. So I've done that a couple of times.* (EYP interview, I001, Maintained)

In relation to the words provided, one setting did discuss deciding to stop delivering one word altogether:

> *They only one word that we just struggled with and we stopped was 'fat'. We gave it a go, some settings were willing to give it a go from the Zoom things we'd have, we gave it a go but from I think it was Tuesday one of the children had called her mummy fat, so we stopped. Luckily the mum took it in jest, and I said we weren't linking it to people, we were linking it to animals, but…we thought, 'No', so we stopped. (*EYP interview, 007, PVI)

Other adaptations were minor, such as changing the resources to suit the setting or to make the story 'more interesting' as discussed in the 'Implementation' section above:

> *Some of the stories are quite old fashioned or plain and boring! I have stuck to the narrative and then just been a little bit more creative*. (EYP Interview, I004, PVI)

> *Obviously changing things like the resources, if we haven't got that specific resource.* (EYP Interview, I004, PVI)

Parental engagement

Summary of the findings

- Maintained and PVI settings used a variety of ways to engage parents with the Concept Cat programme.
- Maintained and PVI settings sent home a variety of Concept Cat activities for parents to do with their children.
- There is some evidence that the Concept Cat programme may have facilitated parental involvement with the setting.
- There is clear evidence that the Concept Cat programme enhanced parents understanding of their child's development and this was mediated by PVI settings.
- EYPs show recognition of the difficulties some families face when engaging with settings.
- Children's enjoyment of the Concept Cat programme facilitated parental engagement.

This section addresses parental engagement with the Concept Cat programme to answer:

> IPERQ6. To what extent have settings engaged families with the programme and in what ways? Are there differences between setting type (PVI/Maintained) in the ways settings have engaged with families? How is this linked, if at all, to child outcomes?

> IPERQ7. What are the barriers and facilitators for families in home implementation of the programme, particularly for focus children, disadvantaged children, and those who are EAL?

It does so by presenting findings from the EYP endline survey data and EYP interviews, which sought to uncover the ways settings had facilitated parental engagement, as well as identify barriers and facilitators to parental engagement. The baseline and endline parent/carer surveys, also sought to uncover any changes in parental involvement and engagement resulting from the Concept Cat programme.

Setting facilitation of parental engagement

The endline EYP survey asked respondents: 'What methods did you use to engage families in Concept Cat?'. All EYPs (64 out of 64 respondents) reported displaying the word of the week on parent notice boards and giving information on Concept Cat via newsletters or similar. Over half of respondents (35 out of 65) indicated that they conducted in-person workshops.

To investigate any difference/similarities in setting types, the data was analysed across type of setting (Table 50). Across setting types, there were few differences in the ways settings reported communicating with parents. Therefore, any differences in parental involvement or engagement cannot be attributed to the listed methods of communication.

Table 50: Methods of communication to parents by setting type[a]

Methods of communication to parents[b]	Maintained n (%)	PVI n (%)
Displaying the word of the week on the parent notice board	41 (100)	23 (100)
Giving information on Concept Cat through newsletters or similar	41 (100)	23 (100)
In-person sessions/workshops	24 (59)	11 (48)
Parent App e.g. Tapestry, class dojo, or social media	7 (17)	2 (9)
Other (please state)[c]	13 (31)	7 (30)

[a] N=64 (41 Maintained, 23 PVI).
[b] More than one response could be given.
[c] Other: 'Concept Cat home activity' = four (one Maintained, three PVIs); 'Stay and play sessions' = two (two PVIs); 'Evidence me (observation and assessment app)' = one (one Maintained); 'Parent meetings' = one (one Maintained).

Data from the endline parent survey also showed settings had communicated the Concept Cat activities in a variety of ways (see Table 51).

Table 51: Methods of communication reported by parents across setting types (N=64)

Communication of Concept Cat activities[a]	Maintained n (%)	PVI n (%)
Online platform/App	47 (54)	8 (35)
Events held in the setting	14 (16)	4 (17)
Newsletters	16 (18)	7 (30)
Other (please state)[b]	10 (11)	4 (17)

[a] More than one response could be given.
[b] Other: 'Home learning' = five (three Maintained settings, two PVIs); 'Word of the week poster' = five (three Maintained settings, two PVIs); 'Face to face conversations' = four (four Maintained settings).

In the endline parent/carer survey, intervention respondents were asked if they were aware their setting had been taking part in the Concept Cat programme. Nearly all respondents (98%, 103 out of 105 respondents) indicated 'yes' (80 out of 82 respondents from Maintained settings and 23 out of 23 respondents from PVI settings). Two respondents from Maintained settings answered 'no' to this question. This suggests that settings had successfully made parents aware of the Concept Cat programme.

The endline parent survey also asked parents/carers if the setting their child attended sent Concept Cat activities to complete at home. Eight respondents out of 74 from Maintained settings, and three respondents out of 21 from PVI settings, replied 'no' to this question. From the 84 respondents out of 95 who replied 'yes' to this question (66 out of 74 respondents from Maintained settings and 18 out of 21 respondents from PVI settings), 68 respondents gave an example of the types of activities. These included flashcards or worksheets, challenges, and games, finding objects to match the word of the week, wristbands, signs/symbols, a home learning book sent home with Concept Cat, and taking photos of children doing the activities. In the EYP interviews, one practitioner commented:

> *I think about two or three weeks into it, we did the word 'long'. Well, they were bringing tape measures from home, they were bringing in sticks they'd found on the street and I've never seen them so getting into the swing of what we're doing, and I would say that's carried on all through the year.* (EYP interview, I005, PVI)

While the monitoring data (see Table 44) shows that some settings struggled to engage more with the families of focus children and that many settings did not hold a family session to explain the purpose of the programme, it is clear from the survey data that settings were trying to engage families.

Data from the interviews also show that settings engaged parents in several other ways. In their interviews, EYPs talked about a variety of methods used to engage parents with Concept Cat. These included: i) showing parents the Concept Cat video or PowerPoint presentation provided by the development team; ii) sending information or resources home including Concept Cat and wristbands and stickers with the word of the week on; or iii) displaying information clearly on the door or noticeboard. EYPs discussed having informal conversations with parents at the start or the end of the day or more organised opportunities, such as stay and play sessions, parents' evenings, and parent workshops. Four EYPs discussed using a parent communication app to convey information. Settings also encouraged families to send in pictures with Concept Cat, and/or of items representing the concept and for children to bring in examples from home.

Parental involvement and engagement

To understand whether the Concept Cat programme had any influence on parents' 'involvement' and 'engagement'[18] with settings, all parents (from both the intervention and control groups) were asked, in both the baseline and endline surveys how involved they were with their child's setting (see Table 52).

Table 52: Reported involvement of parents with settings across control and intervention groups

Involvement	Baseline N=215		Endline N=190	
	Intervention n (%)[a]	Control n (%)[a]	Intervention n (%)[a]	Control n (%)[a]
Very involved	87 (61)	36 (49)	54 (52)	43 (50)
Involved	48 (34)	31(42)	40 (38)	30 (35)
Somewhat involved	5 (4)	6 (8)	9 (8)	11 (13)
Somewhat not involved	2 (1)	0 (0)	1 (2)	2 (2)
Total	142	73	104	86

[a] *Percentages may not sum to 100 due to rounding.*

- At baseline, the majority of parents across intervention and control settings rated themselves as 'very involved' or 'involved' with their child's setting (95%, or 135 out of 142 respondents and 91%, or 67 out of 73 respondents, respectively).

- At endline, there was a drop in parents' ratings of their involvement with their child's setting in both groups although this was slightly less pronounced for parents from intervention settings (90%, or 94 out of 104 respondents) when compared to parents from control settings (85%, or 73 out of 86 respondents).

To investigate this further, analysis was undertaken on the responses from parents who completed the survey at both baseline and at endline (Table 53).

Table 53: Reported involvement of parents with settings across control and intervention groups (parents who completed surveys at baseline and endline only)

Involvement	Baseline N=29		Endline N=29	
	Intervention n (%)[a]	Control n (%)[a]	Intervention n (%)[a]	Control n (%)[a]
Very involved	10 (43)	2 (33)	12 (52)	2 (33)
Involved	10 (43)	3 (50)	9 (39)	3 (50)
Somewhat involved	3 (13)	1 (16)	2 (9)	1 (16)
Somewhat not involved	0 (0)	0 (0)	0 (0)	0 (0)
Total	23	6	23	6

[a] *Percentages may not sum to 100 due to rounding.*

[18] Parental involvement is defined as taking part in setting activities while parental involvement is about understanding children's development so they can be engaged with setting activities.

- At baseline the majority of parents across intervention and control settings rated themselves as 'very involved' or 'involved' with their child's setting (86%, or 20 out of 23 respondents and 83%, or five out of six respondents, respectively) (Table 53).
- At endline, there was a rise in parental ratings of involvement with intervention settings (91%, or 21 out of 23 respondents) but parental involvement ratings in control settings (83%, or five out of six respondents) remained the same (Table 53).

Consequently, given this data is from the same participants, it can be suggested that the Concept Cat programme may have influenced parental involvement with settings. However, given the small number of respondents completing at both baseline and endline, particularly in the control group, this interpretation should be treated with caution.

Further analysis was conducted to look at any differences in overall responses at baseline and endline between setting types (Table 54):

- At baseline, parents with children attending Maintained settings in the intervention group (99%, or 84 out of 85 respondents) rated themselves as slightly more involved compared to parents with children attending PVI settings in the intervention group (90%, or 51 out of 57 respondents). This was similar to ratings from parents with children from Maintained settings in the control group who rated themselves as more involved (96% or 41 out of 43 respondents) compared to parents with children from PVI settings in the control group (87%, or 26 out of 30 respondents).

- At endline, parents with children attending Maintained settings from the intervention group (93%, or 76 out of 82 respondents) rated themselves as more involved compared to PVI settings from the intervention group (81%, or 18 out of 22 respondents). However, parents with children attending Maintained settings in the control group (85%, or 53 out of 62 respondents) rated themselves similarly to parents with children attending PVI settings in the control group (84%, or 20 out of 24 respondents).

Table 54: Involvement of parents with settings across control and intervention groups and setting types[a]

Involvement	Baseline N=215				Endline N=190			
	Intervention		Control		Intervention		Control	
	Maintained n (%)[b]	PVI n (%)[b]	Maintained n (%)[b]	PVI n (%)[b]	Maintained n (%)[b]	PVI n (%)[b]	Maintained n (%)[b]	PVI n (%)[b]
Very involved	54 (64)	33 (58)	27 (63)	9 (30)	44 (54)	10 (45)	28 (45)	15 (63)
involved	30 (35)	18 (32)	14 (33)	17 (57)	32 (39)	8 (36)	25 (40)	5 (21)
Somewhat involved	0 (0)	5 (9)	2 (5)	4 (13)	5 (6)	4 (18)	8 (13)	3 (13)
Somewhat not involved	1 (1)	1 (2)	0 (0)	0 (0)	1 (1)	0 (0)	1 (2)	1 (4)
Total	85	57	43	30	82	22	62	24

[a] Baseline = 142 intervention (85 Maintained, 57 PVI), 73 control (43 Maintained, 30 PVI); Endline = 104 intervention (82 Maintained, 22 PVI), 86 control (62 Maintained, 24 PVI).
[b] Percentages may not sum to 100 due to rounding.

Overall, the data shows that in both Maintained and PVI settings, parents' rating of their involvement with their child's setting declined from baseline to endline. There were also slight differences between Maintained and PVI settings in parents' ratings of their involvement. In both the intervention group and the control group parents with children from Maintained settings were more likely to rate themselves as 'very involved' and 'involved' compared to PVI settings at baseline and at endline.

Table 55: Understanding of children's development (engagement) across control and intervention groups[a]

Understanding of development	Baseline		Endline	
	Intervention n (%)[b]	Control n (%)[b]	Intervention n (%)[b]	Control n (%)[b]
Very good	52 (57)	38 (47)	20 (32)	37 (46)
Good	28 (31)	33 (42)	38 (60)	33 (41)
Somewhat good	7 (7)	10(12)	5 (8)	9 (11)
A little / None	3 (3)	0 (0)	0 (0)	1 (1)
Total	90	81	63	80

[a] N=171 baseline (90 intervention, 81 control), N=143 endline (63 intervention, 80 control).
[b] Percentages may not sum to 100 due to rounding.

Respondents to the parent survey were also asked at both timepoints (baseline and endline) to rate their understanding of their child's development (Table 55):

- At baseline, parents whose children attended intervention settings (88%, or 80 out of 90 respondents) rated their understanding of their child's development as 'very good' or 'good'. Parental ratings of their understanding increased at endline (92% or 58 out of 63 respondents) rating their understanding as either 'good' or 'very good'.

- At baseline, parents whose children attended control settings (89%, or 71 out of 81 respondents) rated their understanding of their child's development as 'very good' or 'good', which was similar to their ratings at endline (87%, or 70 out of 80 respondents) rating their understanding as 'good' or 'very good'.

The data shows that, while parents' rating of their understanding of their child's development increased for intervention settings between baseline and endline, there was a slight decrease overall for control settings, which could be a result of the small number of parents from control settings completing at baseline and endline (n=6). This suggests that the Concept Cat programme may have raised parents' understanding of their child's development; however, these results are taken with caution given the relatively small number of parents from the intervention group also completing at baseline and endline (n=19).

To investigate this further, additional analysis was performed to understand if there were any differences between setting type (Table 56):

- At baseline, a higher proportion of parents with children from Maintained settings in the intervention group (95%, or 67 out of 70 respondents) rated their understanding of their child's development as 'good' or 'very good' compared to parents with children from PVI settings (65%, or 13 out of 20 respondents) in the intervention group.

- At endline, a higher proportion of parents with children from Maintained settings in the intervention group (95%, or 40 out of 42 respondents) rated their understanding of their child's development as 'good' or 'very good' compared to parents with children in PVI settings (72%, or 22 out of 28 respondents) in the intervention group.

- At baseline, a lower proportion of parents with children from Maintained settings in the control group (85%, or 51 out of 60 respondents) rated their understanding of their child's development as 'good' or 'very good' compared to parents with children from PVI settings (95%, or 20 out of 21 respondents) in the control group.

- At endline, a lower proportion of parents with children from Maintained settings in the control group (85%, or 51 out of 60 respondents) rated their understanding of their child's development as 'good' or 'very good' compared to parents with children from PVI settings (95%, or 19 out of 20 respondents) in the control group.

Table 56: Understanding of children's development (engagement) across control and intervention groups in Maintained and PVI settings[a]

	Baseline				Endline			
	Intervention		Control		Intervention		Control	
	Maintained n (%)[b]	PVI n (%)[b]	Maintained n (%)[b]	PVI n (%)[b]	Maintained n (%)[b]	PVI n (%)[b]	Maintained n (%)[b]	PVI n (%)[b]
Very good	44 (62)	8 (40)	26 (43)	12 (57)	14 (40)	3 (11)	26 (43)	11 (55)
Good	23 (33)	5 (25)	25 (42)	8 (38)	19 (45)	19 (61)	25 (42)	8 (40)
Somewhat good	1 (1)	6 (24)	9 (15)	1 (5)	2 (5)	3 (10)	8 (13)	1 (5)
A little / None	2 (3)	1 (5)	0 (0)	0 (0)	0 (0)	1 (3)	1 (2)	0 (0)
Total	70	20	60	21	35	28	60	20

[a] *N=171 baseline (90 intervention, 81 control), N=143 endline (63 intervention, 80 control).*
[b] *Percentages may not sum to 100 due to rounding.*

Overall, the data shows that in the intervention group, while parents rating of their understanding of their child's development remained the same at baseline and endline for Maintained settings, parents' self-ratings rose between endline and baseline in PVI settings. Thus, the increase in parents' self-rating of their understanding child's development

seen in Table 56 came from parents with children who attended PVI settings. Interestingly, while ratings remained the same across Maintained and PVI settings from baseline to endline in the control group, parents with children from control PVI settings rated their level of understanding of their child's development higher than Maintained settings in the control group and the same as Maintained settings in the intervention group.

Barriers and facilitators of parental engagement

To understand any challenges parents experienced completing the Concept Cat activities at home, the parent endline survey asked respondents to list any barriers they experienced in implementation at home. Only ten out of 64 parents indicated that there were barriers, suggesting that the remaining parents did not perceive any barriers to implementation at home. Of those who did indicate barriers, seven respondents stated that time was the biggest barrier, one respondent said they forgot, one respondent mentioned the attention span of the child, and one respondent said they did not always know what the concept was or how to teach it.

During the EYP interviews, there was also a recognition that, for many parents, engagement (with settings generally) was limited, primarily due to external factors:

> *You've probably heard this from many settings, getting parents on board is difficult and it's very difficult and they're all working or grandparents are bringing them but they've taken this on board.* (EYP interview, I005, PVI)

> *One of the focus children, his mum is a single parent so she's spinning all sorts of plates by herself.* (EYP interview, I006, PVI)

In addition, while some parents were seen to be very engaged, sending in photographs of Concept Cat at home or their children with objects identified with the word of the week, EYPs recognised that it was difficult to ascertain the extent of this engagement:

> *So even if they don't necessarily send us things in and send us the photos, when you talk to parents they are talking about the word at home, they're just not necessarily getting round to that sending something for us to see or for the child to share.* (EYP interview, I001, Maintained)

Two EYPs specifically discussed engaging parents of children with EAL:

> *We do have some EAL learners, so it was just making sure that it's emphasised that they obviously use it in their home language as well.* (EYP interviews, I003, Maintained)

> *We have those discussions with the parents, we speak to them in home language, and if they're doing it at home in home language, that understanding in home language and then bringing it into nursery and we can deal with the English.* (EYP interview, I002, Maintained)

Overall, EYPs discussed the difficulties of parental engagement, particularly over the longer term:

> *I think keeping the parents' sort of momentum with the at home activities has been a bit more tricky as we've got through the year, and engaging all of the parents in that.* (EYP interview, I001, Maintained)

However, interview data also suggests that children were seen as the main facilitator of parental engagement:

> *They know what the word is in the week, and then seeing their child understanding it, and using it at home, has really engaged them parents. (*EYP interview, I007, PVI)

Perceived outcomes

Summary of the findings

- EYPs reported that the Concept Cat programme was helpful in supporting practitioner skills to identify children with higher language needs and support children's conceptual language and communication.
- EYPs reported that the Concept Cat programme was helpful in improving practitioner knowledge about conceptual language and communication development.
- EYPs reported that the Concept Cat programme was helpful in improving children's conceptual and expressive vocabulary as well as improving their numerical development.
- The HLEI data showed that the Concept Cat programme had a positive impact on improving the HLE.
- The programme resulted in some positive unintended consequences such as supporting the vocabulary of EAL parents.

The IPE was designed to explore the perceived outcomes of the Concept Cat programme on both practitioners and children, as well as any unintended consequences. This section addresses the following research questions:

> IPERQ5a. To what extent have practitioners developed their knowledge about conceptual vocabulary and skills in identifying and supporting the conceptual vocabulary development of children with higher language needs (i.e. those identified as focus children)?
>
> IPERQ7. What, if any, are the wider impacts on the HLE?
>
> IPERQ8. To what extent does Concept Cat result in positive or negative unintended consequences for settings, practitioners, children, families, and the HLE?

It does so through data provided by the EYP endline survey and interviews, which sought to uncover any perceived impact on practitioners and children. This section also draws on data from the parent/carer baseline and endline surveys to understand impacts on the HLE.

Practitioner outcomes

The EYP endline survey asked respondents various questions related to any perceived changes the Concept Cat programme had on practitioners' skills and knowledge (Table 57).

Table 57: Helpfulness of Concept Cat in improving practitioner knowledge and skills (N=69)

Helpfulness of the Concept Cat programme in	Scale				
	Very helpful n (%)[a]	Helpful n (%)[a]	Neither helpful / nor unhelpful n (%)[a]	Not helpful n (%)[a]	Not helpful at all n (%)[a]
Improving practitioners' skills to identify children with higher language needs	26 (38)	33 (48)	9 (13)	1 (1)	0 (0)
Improving practitioners' skills to support children's conceptual language and communication	34 (49)	33 (48)	2 (3)	0 (0)	0 (0)
Improving practitioner knowledge about children's conceptual language and communication development	37 (54)	30 (43)	1 (1)	1 (1)	0 (0)

[a] *Percentages may not sum to 100 due to rounding.*

- Over three-quarters of EYPs (59 out of 69 respondents) thought the Concept Cat programme had been 'very helpful' (26 respondents) or 'helpful' (33 respondents) in improving their skills in identifying children with higher language needs.
- Almost all EYPs (67 out of 69 respondents) reported that they found the programme 'very helpful' (34 respondents) or 'helpful' (33 respondents) in improving their skills to support children's conceptual language and communication.

- Almost all EYPs (67 out of 69 respondents) reported that they found the programme 'very helpful' (37 respondents) or 'helpful' (30 respondents) in improving their knowledge about conceptual language and communication development.

Given the high number of respondents rating the programme as either 'very helpful' or 'helpful' in improving their knowledge and skills, further analysis across setting type was not deemed appropriate.

The main outcome of the training for practitioners mentioned in the EYP interviews was improved knowledge relating to teaching one concept at a time and not teaching opposites together:

> *It's just made us think completely differently about the way that you teach one word and not saying too many words, and using the right terminology with that word and the right concept for the word. And knowing that one word can have more concepts as well, and knowing how far do we go with this? Definitely, it's just made our knowledge completely different as well in a more focused way.* (EYP interview, I003, Maintained)

Some practitioners also discussed having an increased awareness of the needs of children with lower language skills and knowledge and improved understanding of the needs and experiences of EAL children in their setting:

> *I think we already knew they'd got higher language needs but as I say I do think it's certainly, certainly for me, highlighted that they don't understand what I think they understand...Even children who are really, really bright*. (EYP interview, I005, PVI)

> *That was something that was really interesting, because what I found was by speaking to those parents and understanding how they view it and see it, some of the words mean different things in their languages.* (EYP interview, I003, Maintained)

Interestingly, one EYP specifically linked their learning from Concept Cat with their approach to the children's maths learning:

> *I always find maths, at this age, really, really hard to teach and I've taken a step back personally and just do it in my day to day language. Even, quite often we'll go to the woods on a Monday…I've started just doing it in, 'Right, who's the furthest away in the line?' instead of trying to teach numbers and adding up and taking away. I used to be doing it at snack time, 'Right, I've taken those away so how many have you got left?' but it's boring for a 2 year old, isn't it? But it's not boring to see who's at the front or who's the highest if you've climbed up?* (EYP interview, I005, PVI)

Perceived child outcomes

To understand any perceived change the Concept Cat programme had on children's knowledge, the EYP endline survey asked respondents pertinent Likert-scale questions. The data shows that:

- Almost all EYPs (68 out of 69 respondents) thought the Concept Cat programme had helped to improve children's conceptual and expressive vocabulary.
- The majority of EYPs (86%, 59 out of 69 respondents) reported that it had helped to improve children's numerical development.

In their interviews, EYPs talked about a range of perceived child outcomes resulting from the programme. These were both immediate, in terms of vocabulary and language development and more distal, including school readiness, and socio-emotional development:

> *Sentence structure as well, there's lots of sentences are becoming more complex, yes, there has been.* (EYP interview, I004)

> *I think we've got a sense of real confidence in sending these children off to school that they're going to do better in things like that, so not just English and their speech and language but maths and...Spatial awareness. You know, the whole science…we will be sending these children to school with real foundations of concepts across the board what they need going forward...It boosts their PSED [Personal, Social and Emotional Development] as well.* (EYP interview, I007, PVI)

Two EYPs also linked Concept Cat to maths in term of child outcomes:

> *I do definitely think that, especially because some of the concept words do really link to mathematical language as well, which is really helpful. We really teach maths throughout, we have focused teaching of maths, and from what I've seen with my focused teaching of maths some of those word that we have used or learnt through Concept Cat and being taught to the children are coming out more and they're using that language more.* (EYP interview, I003, Maintained)

The endline EYP survey also asked respondents if they felt the programme had a particular impact on any specific group of children. Forty-eight out of 69 respondents identified group(s) of children they perceived the programme to have a particular impact on: 17 respondents indicated the programme had a distinct influence on 'children with higher language needs'; 22 respondents indicated 'EAL children'; and nine respondents indicated 'EYPP children'.

We further analysed the data to see if the influence on groups of children was perceived as the same across setting types:

- Higher proportions of EYPs in Maintained settings (13 out of 46 respondents) than in PVI settings (four out of 23 respondents) reported that they perceived the programme to be particularly beneficial for children with higher language needs.
- A higher proportion of EYPs from Maintained settings (20 out of 46 respondents) than in PVI settings (two out of 23) said that they perceived the programme to be particularly beneficial for EAL children.
- Eight out of 46 respondents in Maintained settings and one out of 23 respondents in PVI settings said that they perceived the programme to be particularly beneficial for EYPP children.

However, it should also be noted that Maintained settings also reported having higher proportions of EAL, EYPP, and children with higher language needs compared to PVI settings (see Table 9 in the 'Methods' section).

When EYPs were probed further in the interviews about whether or not Concept Cat had benefits for any particular group of children, the findings were more mixed. Furthermore, responses included not only children with higher language needs and those with EAL (which were asked about in the survey) but also older children and those with existing higher language abilities. For example:

> *I think it's had a bigger impact on the older children.* (EYP interview, I005, PVI)

> *For some children who maybe might be say for instance having speech and language therapy it's definitely something that I've noticed that those children say those words a lot more clearly.* (EYP interview, I003, Maintained)

> *I do think Concept Cat has helped them [EAL children] with the words, yes, and with understanding.* (EYP interview, I005, PVI)

Although only a few EYPs identified EYPP children as receiving particular benefits in the survey, focus children were seen to particularly benefit from the programme. One EYP, in their interview, explicitly conflated EYPP with focus children:

> *Sentence structure as well, there's lots of sentences are becoming more complex, yes, there has been. Especially with focus children, they've come on so much, they really, really have.* (EYP interview, I004, PVI)

> *We do have quite a high percentage of Early Years Pupil Premium children at nursery as well, so in terms of that it's really been picked out as well alongside those focused children with Concept Cat as well, which is something that showed a bit of a correlation in a way.* (EYP interview, I003, Maintained)

Two EYPs indicated that there may have been differential impacts based on gender:

> *Particular boys who sometimes their attitude's a bit wobbly to learning, but we've found those boys in particular have engaged. (*EYP interview, I002, Maintained)

Finally, the EYP endline survey asked practitioners: 'Are there any groups of children you feel the Concept Cat programme is not suited for?': 52 respondents replied 'no' to this question; 11 respondents (six from Maintained settings, five from PVI settings) said they did not think it was suitable for children with higher language needs and three respondents (all from Maintained settings) said it was not suitable for EAL children.

Only one EYP discussed SEND children particularly in their interview (other than those with speech, language, and communication needs discussed above), and they did not feel that these children benefited from the programme as intended:

> *When she was doing a SEND, we tried it. It didn't work…I suppose at least they're learning that sitting skill but I don't think they've, no, they've not had anything else.* (EYP interview, I005, PVI)

This was reinforced by a further EYP who discussed the difficulties of implementing Concept Cat with SEND children:

> *Most of our SEND are high needs and haven't really been involved. They don't come to the carpet for an adult-led intervention anyway.* (EYP interview, I004, PVI)

HLE

The HLEI (Melhuish *et al.,* 2013) was embedded in the parent survey at baseline and endline to evaluate any changes within the home during the implementation period of the Concept Cat programme (Table 58).

Table 58: HLEI outcomes

Outcome	Intervention		Control		Total n	P-value
	n	Mean	n	Mean		
Baseline	143	28	74	27	217	.609
Endline	108	26	97	28	205	<.05
Baseline	19	36			19	.208
Endline	19	38			19	

[a] *As there were only six respondents in the control group at baseline and endline this data was not included.*

At baseline, there were no significant differences between control (mean=27, SD=10) and intervention groups (mean=28, SD=9), $t(127) = -0.513$, $p = 0.609$. At endline, there were significant differences between control (mean=26, SD=9) and intervention groups (mean=28, SD=10), $t(201) = -2.33$, $p<0.05$.

Since the overall data set included a number of different participants at baseline and endline, a separate analysis on the total scores was performed on responses given from parents who completed the HLEI at baseline (mean=36, SD=9) and endline (mean=38, SD=10) in the intervention group. A paired sample t-test showed no significant differences within the intervention group at baseline and endline, $t(17) = -1.3$, $p = 0.208$. As there were only six parents in the control group who completed the HLEI at baseline and endline, an analysis between the control and intervention group could not be performed.

Given that the mean score at baseline in the smaller sample (those completing both baseline and endline surveys) (mean=36) is higher than the mean score in the larger sample at baseline (all intervention respondents)(mean=28), it can be seen that respondents in the intervention group who completed the HLEI at both endline and baseline had much higher scores on the HLEI at baseline. As such, a significant difference between baseline and endline cannot be seen within this group (who completed surveys both at baseline and endline), the significant difference at endline between control and intervention groups (i.e. those who completed at least one survey; which was not apparent at baseline) suggests that the Concept Cat programme did have an influence on the HLE.

This data also supports the findings that settings were engaging parents with the Concept Cat programme, which was facilitated by child engagement in the programme.

Unintended consequences

To understand any unintended consequences of the Concept Cat programme on settings EYP, survey respondents were asked at endline: 'Have there been any changes in the setting during the evaluation period? (e.g. increase or decrease in staff turnover, improvement in parental engagement)'. Four out of 20 settings answered 'yes' to this question. One setting stated there had been a decrease in staff (i.e. a decrease in staff turnover), and one setting said there has been an increase in parental engagement; however, this is part of the theory of change and so is not unexpected. Sixteen out of 20 respondents selected 'no' to the question indicated above.

In analysing the interview data, two additional unexpected consequences emerged. In one setting, an EYP talked about the programme also assisting in EAL parents' vocabulary acquisition, particularly their concept learning. Another EYP discussed how useful the programme had been in assisting with a staff members degree programme (in EY education):

> *We have a lot of parents going for ESOL [English for Speakers of Other Languages] classes, so they're picking up the concept as well and they're picking up that language and understanding what it means…Our parents are learning alongside the children.* (EYP interview, I002, Maintained)

> *We did use a lot of the activities for her degree observations which was nice, because we had something structured and she knew exactly what she was planning around that.* (EYP interview, I005, PVI)

Usual practice

Summary of the findings

- There was a wide use of interventions being implemented across control and intervention settings and Maintained and PVI settings during the evaluation period demonstrating a pragmatic landscape over the evaluation period.
- A large majority of settings use the WellComm toolkit to assess children's language needs and as an intervention, this was particularly evident for PVI settings.
- Intervention settings reported some difficulties in presenting words as opposites given the large proportion of settings using the WellComm toolkit where concepts are taught using opposites and concept words are taught closer together.
- There is evidence of some crossover effects of the Concept Cat programme in a very small number of control settings.
- Control settings' practice was different from practice implemented as part of the Concept Cat programme.
- Where control settings used similar teaching and environmental strategies as those used in the Concept Cat programme, intervention settings used them more frequently.

The IPE sought to understand business as usual in EY settings to answer:

> IPERQ3. What is the nature of business as usual with regard to vocabulary instruction? How does this differ between control and intervention settings? What are the similarities/differences between setting type (PVI/Maintained)? How does programme delivery differ from business as usual?

The baseline and endline EYP surveys were designed to capture similarities/differences in the implementation of other programmes and interventions, including the identification of children with higher language needs. The setting observations, along with the endline EYP survey, sought to capture data on whether key strategies used in the Concept Cat programme were being used within control settings.

Implementation of programmes and interventions

As part of the EYP baseline and endline survey, respondents in both control and intervention settings were asked about the whole-class programmes they had been implementing in their setting prior to the evaluation (at baseline) and throughout the evaluation (at endline; see Table 59).

Table 59: Whole-class programmes across control and intervention groups at baseline and endline[a]

Whole-class programmes[b]	Baseline		Total	Endline		Total
	Intervention n (%)	Control n (%)		Intervention n (%)	Control n (%)	
WellComm – The Big Book of Ideas	30 (73)	30 (70)	60	40 (95)	31 (82)	71
Concept Cat	0 (0)	0 (0)	0	35 (83)	4 (11)	39
Communication Friendly Settings	3 (7)	4 (9)	7	4 (10)	14 (37)	18
Letters and Sounds	0 (0)	0	0	4 (10)	14 (37)	18
Makaton and British Sign Language[c]	N/A	N/A		5 (12)	11 (29)	16
Word Aware	0 (0)	0 (0)	0	9 (21)	0 (0)	9
Early Talk Boost	2 (5)	3 (7)	5	5 (12)	4 (11)	9
Communication Champions	0 (0)	0 (0)	0	1 (2)	5 (13)	6
I-Can Early Language Development	1 (2)	2 (5)	3	1 (2)	5 (13)	6
BLAST (Boosting Language Auditory Skills and Talking)	1 (2)	0 (0)	1	2 (5)	1 (3)	3
NELI	1 (2)	0 (0)	1	2 (5)	1 (3)	3
Talking Time©	0 (0)	0 (0)	0	1 (2)	0 (0)	1
Picture Exchange Communication System™ (PECS)	0 (0)	0 (0)	0	0 (0)	3 (8)	3
Learning Language and Loving It™ – The Hanen Program® (Hanen LLLI)	0 (0)	1 (2)	1	1 (2)	0 (0)	1
Talking Partners	0 (0)	0 (0)	0	0 (0)	2 (5)	2

[a] N= 86 baseline (41 intervention, 43 control), N= 80 endline, (42 intervention, 38 control).
[b] More than one response could be given.
[c] Question only asked at endline.

The data indicates that slightly higher numbers of EYPs reported that whole-class programmes were being implemented in their setting (in both control and intervention settings) by the end of the Concept Cat evaluation period than at the start. As seen in Table 59:

- Over the duration of the evaluation there was a rise in the reported use of WellComm The Big Book of Ideas in both the intervention group (73%, or 30 out of 41 respondents) at baseline, compared to 95%, or 40 out of 42 respondents at endline) and in the control group (70%, or 30 out of 43 respondents) reported using WellComm The Big Book of Ideas at baseline compared to (82%, or 31 out of 38 respondents) at endline.

- At baseline, no EYPs (out of 41 respondents in the intervention group and 43 respondents in the control group) reported using Concept Cat. At endline, 11% of EYPs in the control group (four out of 38 respondents from four different settings) reported using Concept Cat. This suggests some contamination of the intervention within the control group.

- 7% of EYPs (three out of 41 respondents) in intervention settings reported using Communication Friendly Settings at baseline, compared to 10% (four out of 42 respondents) at endline. In the control group, 9% of EYPs (four out of 43 respondents) reported using Communication Friendly Settings at baseline compared to 37% (14 out of 38 respondents) at endline.

- In both intervention and control groups, no EYPs (out of 41 respondents in the intervention and 43 respondents in the control) reported using Letters and Sounds at baseline, while at endline, 10% of

EYPs (four out of 42 respondents) in the intervention group reported using the programme and 37% of EYPs (14 out of 38 respondents) in the control group reported using the programme.

Overall, the data suggests that more control settings were using Communication Friendly Settings and Letters and Sounds throughout the evaluation period, when compared to intervention settings. However, it should be noted that these interventions focus on other areas of speech, language, and communication. The data also shows that there was a slightly larger increase in the number of intervention settings using WellComm The Big Book of Ideas during the evaluation period when compared to control settings. Although slightly more settings in the group were implementing WellComm The Big Book of Ideas compared to the control group, the use of concepts is only a small part of this intervention and it teaches concepts as opposites opposed to singular (which is a key component of Concept Cat), the researchers are confident that any additional impact on children seen in the evaluation were due to the implementation of the Concept Cat programme.

Contamination of the Concept Cat programme occurring in control settings was investigated by the evaluation team and the delivery team. It was found that one control setting was implementing the programme. The setting reported incidentally using the programme in nursery and reception years but only used the video or story once a week rather than daily. Due to this, the sensitivity analysis conducted as part of the impact evaluation took this into account (see Table 28).

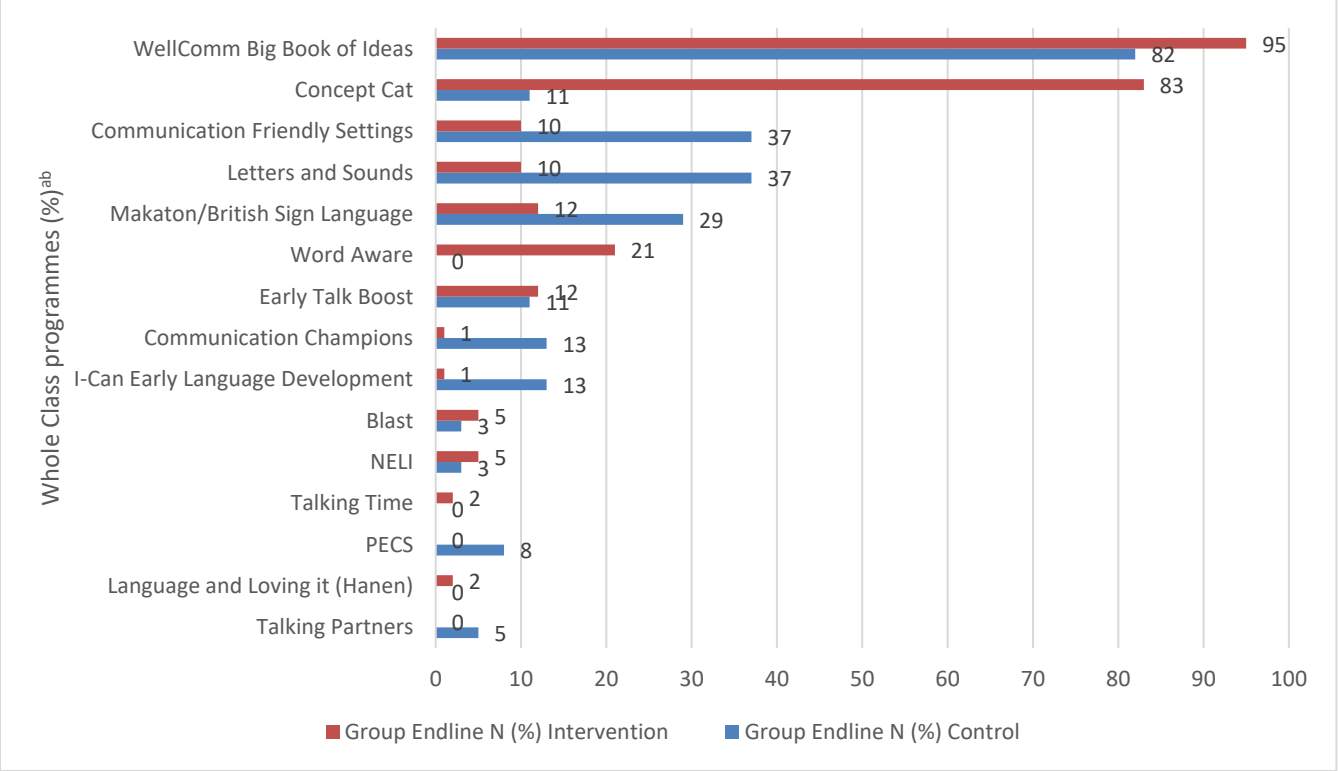


Figure 7: Whole-class programmes across control and intervention groups at endline.[a] [a] N=80 endline, (42 intervention, 38 control). [b] More than one response could be given

The interviews with EYPs in intervention settings highlighted conflicts between existing interventions and Concept Cat. In particular, the WellComm toolkit uses opposite concepts in its teaching and teaches concept words, which are closer together (e.g. long, longer, longest) whereas these concept types in Concept Cat would typically be taught separately (and taught in different levels). One EYP discussed a similar issue in the context of the Early Years Foundation Stage (EYFS):

> *We've got Wellcomm assessments next week and we know we've got long, longer and longest coming up, but because of Concept Cat it's long and not long…So we are doing a little bit of, 'It's long. Ooh, that's longer', so we're starting with the 'long'...There's just one person who does the Wellcomm assessment, so she's picked out, 'Can I just make everyone aware, long, longer, longest, tall, tallest, taller'.* (EYP interview, I002, Maintained)

> *The EYFS framework, because if you look at it and then you do it, heavy, light, long, short, you think, 'I've got to put these concepts together'.* (EYP interview, I002, Maintained)

One EYP interviewee from the control group also explicitly linked the use of WellComm in their setting to the teaching of opposites:

> *I think most of the concepts that are covered in the WellComm book, lots of them already, so...I mean our focus has been the last two or three weeks in WellComm there's been opposites.* (EYP interview, C001, Maintained)

In contrast, some EYPs in intervention settings described other existing practices and interventions as complementary to their implementation of Concept Cat:

> *We do Birth to 5 Matters and most of those concepts are within our maths or our understanding of the world anyway, so we would have been teaching them anyway.* (EYP interview, I004, PVI)

> *In terms of looking at vocabulary and things like we have our core story that we have, and we use the key vocabulary from the core story and use those words in provision because those words are planned for and they're meaningful words that we use from that core story. Other than those no, I've not done anything like this.* (EYP interview, I003, Maintained)

The baseline and endline data were also analysed across setting types (Table 60).

Table 60: Whole-class programmes baseline and endline across setting types[a]

Whole-class programmes[b]	Baseline		Total	Endline		Total
	Maintained n (%)	PVI n (%)		Maintained n (%)	PVI n (%)	
WellComm The Big Book of Ideas	40 (66)	20 (74)	60	50 (86)	21 (95)	71
Concept Cat	0 (0)	0 (0)	0	27 (47)	12 (54)	39
Communication Friendly Settings	4 (7)	3 (11)	7	15 (26)	3 (14)	18
Letters and Sounds	0 (0)	0	0	17 (29)	1 (5)	18
Makaton and British Sign Language[c]	N/A	N/A		15 (25)	1 (5)	16
Word Aware	0 (0)	0 (0)	0	5 (9)	4 (18)	9
Early Talk Boost	4 (7)	1 (4)	5	7 (12)	2 (9)	9
Communication Champions	0 (0)	0 (0)	0	1 (2)	5 (23)	6
I-Can Early Language Development	3 (5)	0 (0)	3	3 (5)	3 (14)	6
BLAST	0 (0)	1 (4)	1	0 (6)	3 (14)	3
NELI	1 (2)	0 (0)	1	2 (3)	1 (5)	3
Talking Time©	0 (0)	0 (0)	0	0 (0)	1 (5)	1
PECS	0 (0)	0 (0)	0	0 (0)	3 (14)	3
Hanen LLLI	1 (2)	0 (0)	1	1 (2)	0 (0)	1
Talking Partners	0 (0)	0 (0)	0	2 (3)	0 (0)	2

[a] N= 88 baseline (61 Maintained, 27 PVI); N= 80 endline, (58 Maintained, 22 PVI).
[b] More than one response could be given.

The data in Table 60 shows a mixed picture for Maintained and PVI settings at baseline and endline. Of particular interest:

- More EYPs from both Maintained and PVI settings reported using WellComm The Big Book of Ideas at endline than baseline. In Maintained settings, 66% of EYPs (40 out of 61 respondents) reported using WellComm The Big Book of Ideas at baseline compared to 86% (50 out of 58 respondents) at

endline. In PVI settings, 74% of EYPs (20 out of 27 respondents) reported using WellComm The Big Book of Ideas at baseline compared to 95% (21 out of 22 respondents) at endline.

- More EYPs from Maintained settings, reported using Communication Friendly Settings at endline, compared to baseline. In Maintained settings, 7% of EYPs reported using Communication Friendly Settings (four out of 61 respondents) at baseline compared to 26% (15 out of 58 respondents) at endline. In PVI settings, 11% of EYPs (three out of 27) reported using Communication Friendly Settings at endline compared to 14% (three out of 22 respondents) at endline.

- Only 5% of EYPs from PVI settings (one out of 22 respondents) said they used Makaton and British Sign Language while 25% of EYPs from Maintained settings (15 out of 58) said they used Makaton and British Sign Language at endline.

In their interviews, two EYPs from settings allocated to the control group (one PVI and one Maintained) discussed the Communication Friendly Settings training other staff members were receiving, although in both cases cascading was only beginning to occur:

> *She's done the Elklan, she's done that, so she's fed bits down to us during team meetings…she's fed a lot of that down just about giving children the time to think before, to answer, extending that language, adding one more word and giving them time to think. Not bombarding them with questions, things like that, that as a practitioner you do sometimes do.* (EYP interview, C002, PVI)

> *We can I think apply for a Community Friendly Setting status. So we've had two meetings with that, there's six meetings and it lasts a year, so she's training us. (*EYP interview, C004, Maintained)

In the EYP interviews, two Maintained intervention settings also spoke specifically about the use of Makaton and British Sign Language being used within the school as did two EYPs from control group Maintained and PVI settings:

> *We're Makaton. We want to become a communication for all school, so we do Makaton, we're all Makaton trained. (*EYP interview, I002, Maintained)

> *We've been doing Makaton for a long time with the children, we tend to use most of it in singing as a whole group, and then when we work with specific children or specific groups that we know need that language support we use Makaton signs with them as well, especially are children that have language delays and communication issues, they can be supported by Makaton.* (EYP interview, C001, PVI)

Table 61 shows how practitioners reported assessing children's language needs across setting types.

Table 61: How language needs are assessed by setting allocation and setting type[a]

Identification of language needs[b]	Baseline N=85				Endline N=81			
	Intervention		Control		Intervention		Control	
	Maintained n (%)	PVI n (%)	Maintained n (%)	PVI n (%)	Maintained n (%)	PVI n (%)	Maintained n (%)	PVI n (%)
Practitioner knowledge	15 (58)	7 (50)	15 (48)	3 (25)	24 (86)	10 (78)	29 (91)	6 (75)
Speech and language assessment	5 (18)	2 (14)	7 (23)	2 (17)	5 (18)	4 (31)	17 (53)	4 (50)
Screening tool	22 (52)	6 (43)	18 (58)	5 (42)	19 (68)	11 (85)	22 (69)	7 (88)
Other	6 (23)	4 (29)	5 (16)	3 (25)	1 (4)	2 (15)	4 (13)	2 (25)

[a] Baseline = 42 intervention (28 Maintained, 14 PVI), 43 control (31 maintained, 12 PVI); Endline, 41 intervention (28 Maintained, 13 PVI), 40 control (32 Maintained, 8 PVI).
[b] More than one response could be given.

- At baseline, EYPs reported using practitioner knowledge to assess children's language needs similarly in the intervention group across both Maintained (58%, or 15 out of 28 respondents) and

PVI settings (50%, or seven out of 14 respondents). Similar levels of using practitioner knowledge to assess children's language needs were also reported by EYPs in Maintained settings in the control group (48%, or 15 out of 31 respondents). However, only 25% of EYPs in PVI settings allocated to the control condition (three out of 12 respondents) reported using practitioner knowledge.

- Across setting types, all EYPs were more likely to report using practitioner knowledge at endline compared to baseline: More EYPs in PVI settings in the intervention group reported using practitioner knowledge at endline (78%, or ten out of 13 respondents), when compared to baseline (50% or seven out of 14 respondents). More EYPs in PVI settings in the control group reported using practitioner knowledge at endline (75%, or six out of eight respondents), when compared to baseline (25%, or three out of 12 respondents). More EYPs in Maintained settings in the intervention group (86%, or 24 out of 28 respondents) reported using practitioner knowledge at endline when compared to baseline (58%, or 15 out of 28 respondents), and more EYPs in Maintained settings in the control group reported using practitioner knowledge at endline (91%, or 29 out of 32 respondents) when compared to baseline (48%, or 15 out of 31 respondents).

- At baseline, the use of a speech and language assessments to assess children's language needs was reported by EYPs similarly across control and intervention groups in Maintained and PVI settings. In Maintained settings in the intervention group, EYPs reported use of speech and language assessments as being the same at baseline and endline (18%). In PVI intervention settings, more EYPs reported the use of speech and language assessments at endline (31%, or four out of 13 respondents) when compared to baseline (14%, or two out of 14 respondents). In control settings, more EYPs in PVI settings reported the use of speech and language assessments at endline (50%, or four out of eight respondents), compared to baseline (17%, or two out of 14 respondents), and more EYPs in Maintained settings in the control group reported using speech and language assessments at endline (53%, or 17 out of 32 respondents) compared to baseline (23%, or seven out of 28 respondents).

- At baseline, the use of a screening tool to assess children's language needs was reported by similar proportions of EYPs across control and intervention groups and setting types. At endline in Maintained settings, there was an increase in the intervention group from baseline (52%, or 22 out of 28 respondents) to endline (68%, or 19 out of 28 respondents), as well as in the control group from baseline (58%, or 18 out of 31 respondents) to endline (69%, or 22 out of 32 respondents). For PVI settings, this rise was much more prominent with PVI settings in the intervention group, increasing from baseline (43%, or six out of 14 respondents) to endline (85%, or 11 out of 13 respondents), and PVI settings in the control group raising 46% from baseline (42%, or five out of 12 respondents) to endline (88% or seven out of eight respondents).

To understand more about similarities/differences between control and intervention settings in the way they assess children's language needs we asked EYPs in the endline survey to specify what speech and language assessments were used to monitor speech and language difficulties in children. Across intervention and control groups settings were most likely to report using the WellComm (98%, or 45 out of 46 respondents and 83%, or 38 out of 46 respondents, respectively). EYPs were also asked to specify what screening tools were used to assess those who may have speech and language difficulties. The data shows again that WellComm was the most widely reported screening tool used across control (52%, or 16 out of 31 respondents) and intervention settings (58%, or 14 out of 24 respondents).

The data suggests no differences between Maintained and PVI settings in the way in which they assessed children's language needs. However, more EYPs in PVI settings reported the use of a screening tool when compared to baseline PVIs and PVIs in Maintained settings at endline.

Table 62 shows the higher language needs interventions being implemented across control and intervention groups.

As with whole-class programmes, there were more programmes reported by EYPs as being implemented for children with higher learning needs at endline compared to baseline in both control and intervention conditions:

- At baseline fewer EYPs in intervention settings reported the use of WellComm for children with higher language needs (68%, or 28 out of 41respondents) than at endline (88%, or 37 out of 42 respondents). Fewer EYPs in the control settings also reported the use of WellComm at baseline (49%, or 21 out of 43 respondents) compared to endline (74%, or 28 out of 38 respondents). Overall, EYP's in the intervention settings reported using WellComm more than EYPs in the control settings at baseline and endline.

- EYPs in intervention settings did not report using Communication Friendly Settings as an intervention for children with higher language needs at baseline while at endline, 17% (or seven out of 42 respondents) reported using Communication Friendly Settings. More EYPs in control settings reported using Communication Friendly Settings at endline (34%, or 13 out of 38 respondents) compared to baseline (5%, or two out of 41 respondents).

- No EYPs in either control or intervention settings (out of 84 respondents) reported using Concept Cat in the baseline survey. In the endline survey, 8% of EYPs in the control group (three out of 38 respondents from three settings) reported using Concept Cat. This suggests some contamination of the intervention within the control group.

Table 62: Interventions for higher learning needs across control and intervention groups[a]

Higher learning needs interventions[b]	Baseline		Total	Endline		Total
	Intervention n (%)	Control n (%)		Intervention n (%)	Control n (%)	
WellComm The Big Book of Ideas	28 (68)	21 (49)	39	37 (88)	28 (74)	55
Concept Cat	0 (0)	0 (0)	0	28 (67)	3 (8)	31
Communication Friendly Settings	0 (0)	2 (5)	2	7 (17)	13 (34)	20
Early Talk Boost	2 (5)	3 (7)	5	4 (10)	11 (29)	15
Early Talk 0 to 5	0 (0)	0 (0)	0	1 (2)	0 (0)	1
I-CAN Early Language Development	0 (0)	0 (0)	0	2 (5)	2 (5)	4
Hanen LLLI	0 (0)	1 (2)	1	2 (5)	0 (0)	2
Every Child a Talker (ECAT)	0 (0)	0 (0)	0	2 (5)	3 (8)	5

[a] N= 84 baseline (41 intervention, 43 control); N= 80 endline, (42 intervention, 38 control).
[b] More than one response could be given.

Figure 8 shows the interventions for children with higher language needs across control and intervention conditions at endline.

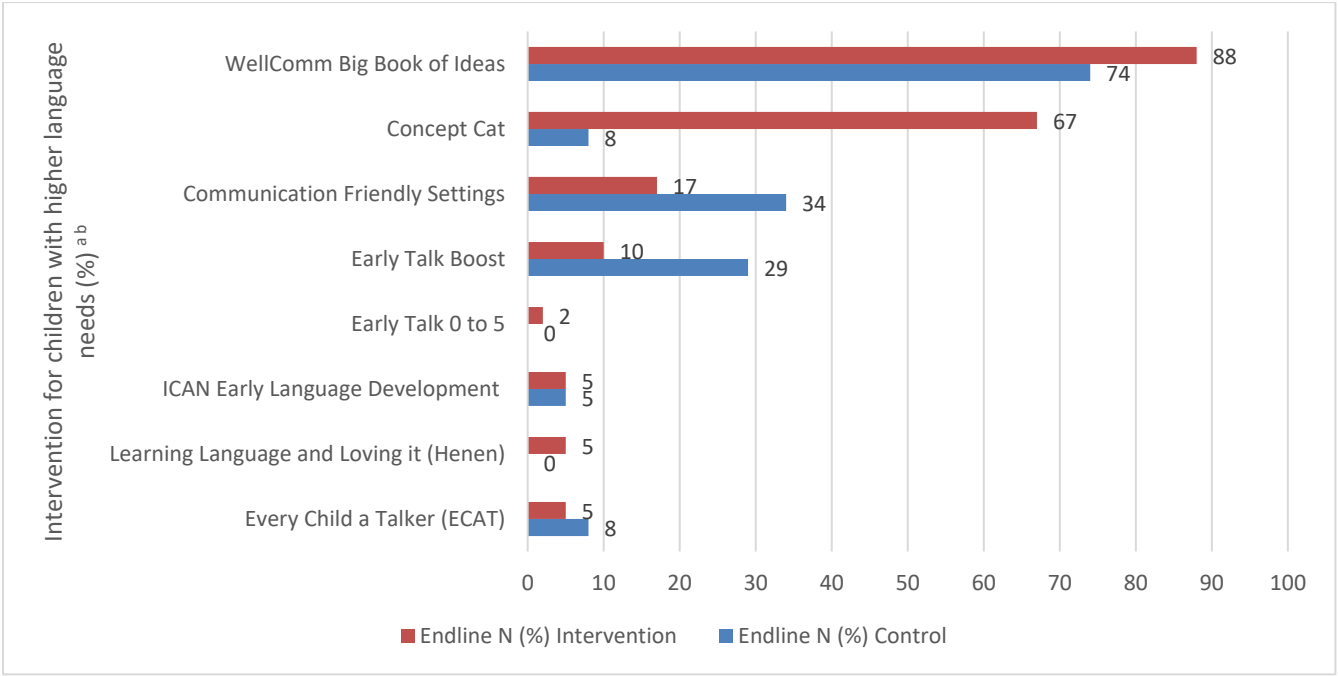


Figure 8: Programmes for children with higher language needs across control and intervention groups at endline. a N=157 endline, (79 intervention, 78 control). b More than one response could be given.

WellComm The Big Book of Ideas was mentioned in the interviews as being more frequently used to support children with speech and language delay/SEND than as a general intervention for all children, and one EYP explicitly linked this to EYPP children in their setting:

> *Generally, EYPP children are a little bit of language delay so that's noted anyway, we know that anyway so they're possibly the ones that we would be giving WellComm to and offering possibly speech and language intervention through their own health visitors is already in place anyway.* (EYP interview, I004, PVI)

An EYP in the control group also discussed specifically targeting EYPP children and EAL:

> *We're very aware of who our EYPP children are, so we know what we're doing with them, we're targeting them, we're trying to offer them as much language, etc. The EAL children as well, I think now probably are making far more progress than they ever have before, because it's such a structured approach actually.* (EYP interview, C001, PVI)

The baseline and endline data were also analysed across setting type (Table 63).

Table 63: Interventions for higher learning needs across setting types[a]

Higher learning needs interventions[b]	Baseline		Total	Endline		Total
	Maintained n (%)	PVI n (%)		Maintained n (%)	PVI N (%)	
WellComm The Big Book of Ideas	33 (56)	6 (23)	39	48 (83)	17 (77)	65
Concept Cat	0 (0)	0 (0)	0	20 (34)	11 (50)	31
Communication Friendly Settings	1 (2)	1 (4)	2	15 (26)	5 (23)	20
Early Talk Boost	5 (8)	0 (0)	5	12 (21)	3 (14)	15
Early Talk 0 to 5	0 (0)	0 (0)	0	1 (2)	0 (0)	1
I-CAN Early Language Development	0 (0)	0 (0)	0	2 (4)	2 (9)	4
Hanen LLLI	1 (2)	0 (0)	1	2 (4)	0	2
ECAT	0 (0)	0 (0)	0	2(4)	3 (14)	5

[a] N=85 baseline (59 Maintained, 26 PVI); N=80 endline, (58 Maintained, 22 PVI).
[b] More than one response could be given.

- WellComm The Big Book of Ideas was reported as being used by more EYPs in Maintained settings at baseline (56%, or 33 out of 59 respondents) compared to PVI settings at baseline (23%, or six out of 26 respondents). At endline, EYPs in Maintained settings reported using WellComm The Big Book of Ideas (83%, or 48 out of 58 respondents) in similar proportions to EYPs in PVI settings (77%, or 17 out of 22 respondents). Thus, the rise seen in the use of WellComm The Big Book of Ideas for children with higher language needs was largely from PVI settings but also from Maintained settings.

- Similar proportions of EYPs from Maintained and PVI settings reported using Communication Friendly Settings at baseline (2%, or one out of 59 respondents and 4%, or one out of 26 respondents, respectively). At endline more EYPs in Maintained settings (26%, or 15 out of 58 respondents) reported using Communication Friendly Settings than in PVI settings (23%, or five out of 22 respondents). Thus, the rise in Communication Friendly Settings being implemented across intervention and control settings came mainly from Maintained settings but also from PVI settings.

Overall, the data from the EYP baseline and endline surveys appears to suggest that more whole class and individualised interventions were being implemented in control and intervention settings and across Maintained and PVI settings at endline compared to baseline.

Key strategies used in settings

One of the elements of Concept Cat is sharing best practice through good practice networks. However, as seen in Table 64, good practice networks were not as well-attended. The endline EYP survey explored whether the sharing of best practice was similar across control and intervention settings and across setting types (Table 64).

Table 64: Sharing of best practice across control and intervention settings and setting types[a]

Sharing of best practice	Intervention		Control	
	Maintained n (%)[b]	PVI n (%)[b]	Maintained n (%)[b]	PVI n (%)[b]
Do not share best practice	16 (28)	5 (23)	8 (13)	6 (38)
Share with local setting	22 (39)	11 (50)	30 (49)	5 (31)
Share with sister settings	7 (12)	3 (14)	7 (11)	5 (31)
Other[c]	12 (21)	3 (14)	16 (26)	0 (0)

[a] N=79 intervention (57 Maintained, 22 PVI); N=77 control (61 Maintained, 16 PVI).
[b] Percentages may not equal 100 due to rounding.
[c] Other: Multi Academy Trusts = seven (four Maintained, three PVIs); Across settings =nine (five Maintained, four PVIs); Local clusters = five (four Maintained, one PVI); other Maintained settings (ten).

- Majority of EYPs reported that their settings shared best practice with other local settings. Lower proportions of EYPs in Maintained settings in the intervention group (39%, or 22 out of 57 respondents) reported sharing best practice with local settings than in Maintained settings in the control group (49%, or 30 out of 61 respondents) and then EYPs in PVI settings in the intervention group (50%, or 11 out of 22 respondents).

- A similar proportion of EYPs from Maintained (28%, or 16 out of 57 respondents) and PVI settings (23%, or five out of 22 respondents) in the intervention group reported not sharing best practice with other settings. More EYPs from PVI settings (38%, or six out of 16 respondents) in the control group when compared to the intervention group (13%, or eight out of 61 respondents) and compared to Maintained settings in the control group (13%, or eight out of 61 respondents) reported that they did not share best practice with other settings.

- More EYPs from PVI settings in the control group reported sharing best practice with sister settings (31%, or five out of 16 respondents) compared to EYPs from PVI settings in the intervention group (14%, or three out of 22 respondents).

Control practice compared to intervention practice

EYPs in the control group were asked, as part of the endline survey, questions linked to the key elements of the Concept Cat programme.

EYPs in control group settings were asked if they specifically taught children concepts in the setting. Forty-four percent (34 out of 77 respondents) selected 'no' to this question, 53% (41 out of 77 respondents) selected 'yes, all children', and 4% (three out of 77 respondents) selected 'yes, key children'. For those who specifically taught concepts to 'all children' 83% (or 34 out of 44 respondents) taught concepts as opposites (e.g. long/short), while 16% (or seven out of 44 respondents) said they taught concepts separately (e.g. long/not long). Of the three EYPs who reported specifically teaching concepts to 'key children', all three (out of three respondents) said they taught concepts as opposites (e.g. long/short). This demonstrates that usual practice is sufficiently different in the way children are taught concepts across control settings.

EYPs in the control settings were also asked if they used the STAR approach in their settings; 92% (or 70 out of 76 respondents) said 'no' to this question. Three percent (or two out of 76 respondents) answered 'yes, I have used the STAR approach in a different setting'. Five percent (of four out of 76 respondents) from two settings said they are currently using the STAR approach, and 1% (or one out of 76 respondents) said that they have previously used the STAR approach in the setting.

Where respondents said they were currently using the STAR approach, they were asked how they used the approach within the setting. One of the two EYPs replied, stating: '*We identify key vocabulary that all staff use with the children and these words are shared with parents'* (Open response, EYP endline survey). This data is similar to the baseline survey data where 93% (or 41 out of 44 respondents) stated they did not use the STAR approach, 2% (or one out of 44 respondents) said they previously used the STAR approach and 5% (or two out of 44 respondents) said they currently used the STAR approach. This suggests that the STAR approach is not widely used in settings allocated to the control condition.

In their interviews, two EYPs in intervention settings mentioned using the STAR approach prior to Concept Cat, although they both indicated that this usage was limited, and a further EYP indicated that they had, to some extent, not been teaching opposites together:

> *I'd worked with somebody a few years ago on supporting children with EAL and so I'd already been introduced to the idea of not using opposites by her because of the confusion. So it was something that I was already doing and we were already doing, staff members who'd been around since then were already doing. But it was new to some people. And we've definitely done it more since implementing Concept Cat than we would have done before.* (EYP interview, I001, Maintained)

> *Occasionally I've done it [the STAR approach] with my SEND children, occasionally but it's just different. You go on different courses and they use different teaching methods don't they, so yes, I probably looked at it but I wouldn't say we've used it generally with the children, no.* (EYP interview, I005, PVI)

In addition, one EYP from a setting allocated to the control condition discussed using the STAR approach in their interview:

> *The vocabulary, the concepts that we want to teach, that is put onto the planning and then within the planning then, I basically script out how we're going to use it within our direct teaching time. So the staff will then during that direct time and group time, they will teach those words and they use repetition to get the children to all say it. We say it numerous times and we'll use objects to bring it to life…so we'll use concrete objects or work go out in the garden to look at what we're talking about, if it's something like we're doing planting or whatever, to activate that word…we're covering the same topic repeatedly for a week it could be or two weeks, so they're practicing using that vocabulary.* (EYP interview, C003, Maintained)

EYPs from the control group were asked in the endline survey if they used the Word Aware approach in their settings. One EYP (out of 78 respondents) stated that they currently used the Word Aware approach and further stated that they use it '*to introduce vocabulary tiered words during topic time'* (Open response, EYP endline survey). Five percent of EYPs (or four out of 78 respondents) said they had previously used the STAR approach, and 95% (or 73 out of 78 respondents) replied 'no' to this question. This data is similar to the baseline survey data where 7% of EYPs (three out of 44 respondents) reported using the Word Aware approach and 93% (or 41 out of 44 respondents) said they did not use the STAR approach. The control group EYP who used the STAR approach also reported having the Word Aware 2 book (Parsons and Branagan, 2016), with both the STAR approach and the book being recommended through Talk Nursery training. However, despite their planning containing elements of Concept Cat they did not report using the book extensively '*we've kind of dibbed in and out*' (EYP interview, C003, Maintained) and this did not influence the teaching of opposites within their setting:

> *So I'm aware that we shouldn't be teaching the opposites, but they just seem to naturally go together sometimes within story.* (EYP interview, C003, Maintained)

The midpoint setting observations with control and intervention settings specifically looked for similarities in practice with Concept Cat (Table 65). When compared to the observations in intervention settings it can be seen that:

- Intervention settings were using all teaching strategies (as listed in Table 65) almost all of the time (38 out of a possible 40) whereas control settings were using the strategies less than half of the time (15 out of a possible 40).
- Intervention settings were using environmental strategies (as listed in Table 65) more than control settings (ten out of a possible 12 compared to seven out of a possible 12, respectively).
- Overall, intervention settings were using Concept Cat strategies over twice as much as control settings (48 out of a possible 52 compared to 22 out of a possible 52) meaning that while all settings tended to use the strategies in normal teaching practice, intervention settings were using them much more frequently.

Table 65: Strategies used in intervention and control settings[a]

Strategies	Control			Intervention		
	Yes	No	Not observed	Yes	No	Not observed[b]
Teaching:						
Emphasis on teaching a word[c]	1	3		4	0	
Sign/gesture used to portray the word	4	0		4	0	
Children are encouraged to use specific words[c]	1	3		4	0	
A story or song is used to emphasise words[c]	0	4		3	0	1
There is an emphasis on teaching concept words specifically[c]	2	2		4	0	
If concepts are taught, they do not say the opposite e.g. long/short[c]	0	4		4	0	
All practitioners emphasise the word or concept[c]	2	2		4	0	
The word/concept used during everyday activities[c]	2	2		4	0	
Practitioners use comments rather than questions	1	3		3	1	
Key concepts/words are used in sentences[c]	2	2		4		
Total	15	25		38	1	1
Environment:						
There are activities available where the words can be used	3	1		4		
Key words on display in the setting	3	1		4		
There is a poster/notice board which is used to communicate words being taught for parents	1	3		2	2	
Total	7	5		10	2	0
Overall total	22	30		48	3	1

[a] *N= 8 (four control, four intervention).*
[b] *NB: Observed indicator was used as the researchers asked the practitioners whether elements not observed were normally used in practice.*
[c] *It is important to note that this does not indicate poor practice in control settings, only that settings were not implementing key strategies used in the Concept Cat programme.*

In teaching practice in intervention settings, there was much more emphasis on the strategies related to Concept Cat: teaching a specific word; encouraging children to use a specific word; using songs or stories to emphasise a word; and not using opposite concepts, compared to control settings. Both control and intervention settings used some kind of gesture or word to portray words. Further analysis showed that, in two control and intervention settings, Makaton and British Sign Language was used during carpet time and in the other two control and intervention settings, it was used by practitioners during free play. Thus, signs were used similarly across control and intervention settings.

In the environment, intervention settings were more likely to use strategies related to Concept Cat: displaying words being used in poster format (or similar) for parents to see and use at home. Three of the four control settings also provided activities to encourage the use of words during free play and had key words on display although the extent to which they reported using these in the EYP interviews varied:

> *We use the boards in the areas as well, in the concept provision areas we'll have the little blackboards just to prompt the practitioners of the words as well that we want to teach or a concept that we want to teach, so we'll have those out as well.* (EYP interview, C003, Maintained)

> *We did have some little picture cards actually, I've still got them somewhere, we need to get them up.* (EYP interview, C002, PVI)

Further analysis showed that in control settings, while play areas could encourage the use of words, the practitioners did not always use specific words with children during these activities. In intervention settings, it seemed more likely that the activities had been set-up with the objective of practitioners and children both using the word. For example, in one

setting, a circular course had been set-up in the outside area to encourage children and practitioners to use the word 'around', while the children used the scooters to go around the course.

Continued implementation of Concept Cat

Summary of the findings

- Practitioners in both Maintained and PVI settings spoke about continued implementation of the Concept Cat programme in their settings.
- Practitioners from both Maintained and PVI settings were motivated to continue implementing Concept Cat and spoke about the positive impact of the programme.

The following section addresses the following research question:

> IPERQ5b. To what extent are practitioners motivated to implement, and continue to implement, Concept Cat? Is this motivation different across setting type (PVI/Maintained) and if so, why?

The practitioner interviews and endline EYP survey were designed to capture data on practitioners and children's enjoyment of the programme and whether settings intended to continue implementing Concept Cat.

Continued implementation of the Concept Cat programme

In their interviews, all EYPs in the intervention group indicated that they would be continuing to implement Concept Cat in the next academic year. Similarly, EYPs surveyed at endline also commented on their plans to continue implementing the programme:

> *A fantastic programme that we have really enjoyed taking part in and we will continue to do next year. The children love Concept Cat and have engaged with it really well and used the words well through play across the setting. A big thank you to [Coach] who has been really supportive throughout the journey.* (EYP survey, Maintained)

In addition, all EYPs interviewed stated that they would recommend Concept Cat to other EY settings, and some discussed already having done so:

> *I think I would, because the children have been engaged, so they've been interested. It's simple and doesn't require a lot of resources. From our point-of-view as practitioners it's simple to understand and to implement.* (EYP interview, I001, Maintained)

> *I'm doing this Early Years Development Programme now and I've brought it up a lot on my online, and I know there's people from all over the country have looked it up because of what I've said, they've said, 'Oh, we'll look that up, it sounds really good'.* (EYP interview, I005, PVI)

Many of these interviewees (four out of seven EYPs interviewed in intervention settings) also discussed extending their use of Concept Cat in the following academic year, particularly to younger age groups, although one EYP also indicated that they anticipated continuity when the children entered reception next academic year (having already shared Concept Cat with reception staff):

> *I think as well it's something that we'd like to hopefully develop with the 2-year-old[s].* (EYP interview, I003, Maintained)

> *Obviously when this cohort of children go through into Reception, they will have had a year of Concept Cat, so it will be even more meaningful for them if some of those tier three words, for example, crop up, that the same approach could be used.* (EYP interview, I001, Maintained)

Motivation to implement, and continue to implement Concept Cat

In the EYP endline survey and the interviews, practitioners gave a number of reasons relating to their motivation to implement Concept Cat, and to continue to implement the programme beyond the intervention year. As discussed in

the 'perceived outcomes' section above, children's enjoyment of the programme and the ease of implementation of Concept Cat were both cited as important factors:

> *The programme is amazing! The children have loved the interactive stories with Concept Cat and the staff have found the toolkit really easy to use. We have loved it.* (EYP interview, I004, PVI)

EYPs also mentioned the positive impact the programme was perceived to have had on the children in their setting within this context:

> *We have found that the project has benefited all the children and have responded really positively to it.* (EYP survey, Maintained)

Overall, the data suggests that practitioners across setting types were motivated to implement Concept Cat and there is evidence to suggest that settings will continue to implement the programme.

Cost

The cost estimates presented in Table 66 and Table 67 are derived from data collected by the delivery team regarding expenses incurred by both themselves and the settings involved in this trial.

Table 66: Cost of delivering Concept Cat in this trial

Item	Type of cost	Cost (average per setting in this trial)	Total expected cost over three years	Total expected cost per child per year over three years[a]
Three-hour training session for lead practitioner	Start-up	£60	£60	
One-hour training session for other staff	Start-up	£47.04	£47.04	
Cost of covering staff attending training – 50% costs covered by the EEF trial	Start-up	£220[b]	£220[c]	
Resources costs (books, puppets, bags, cats, word bags, assembly, and postage)	Start-up	£363.60	£363.60	
Concept Cat coach visits (seven visits)	Start-up	£1,372.80	£1,372.80	
Concept Cat coach costs (training, travel, supervision, administration)	Start-up	£426	£426	
Delivery team administrative costs (project monitoring and delivery coordination)	Start-up	£330.96	£330.96 (in the first year)	
Delivery team administrative costs (project monitoring and delivery coordination)	Recurring		£240 (£120 each subsequent year)	
Top-up training if change in new staff[d]	Recurring		£120 overall (£60 per training per year)	
Cost of covering staff attending top-up training	Recurring		£440 overall (£220 per year)	
Total		£2,820.40	£3,620.40	£30.94

[a] *Based on total numbers of children aged three to four in treatment schools, as reported by settings to Concept Cat.*
[b] *Actual costs to settings in trial with 50% cover by the EEF would be (on average) £110.*
[c] *Actual costs to settings in trial with 50% cover by the EEF would be (on average) £110.*
[d] *Assume new staff trained once a year in year two and year three.*

Table 67: Cumulative costs of Concept Cat

Programme	Year one	Year two	Year three
Concept Cat	£2,820.40	£3,220.40	£3,620.40

Thus, the average cost of delivering Concept Cat per setting over three years would be £1,206.80.

In calculating the costs for the delivery of Concept Cat we have made the following assumptions:

- For the trial, the EEF paid 50% of staff cover costs so that practitioners could attend training. On average, settings spent £220 on training, with £110 being subsidised by the EEF. For the purposes of this cost calculation £220 was used.
- Resources to deliver the programme with a cost of £363.60 per setting.
- Seven visits for Concept Cat coaches to visit each setting in the first year, averaging 28.6 hours overall, with a cost per setting of £1,372.80.
- Administration costs averaging £330.96 per setting.
- Supervision of coaches by Concept Cat team, comprising five one-on-one sessions, and six group sessions, plus travel, £426 per setting.
- We have assumed settings will require top-up training once a year (£220 per year) due to the relatively high turnover found in EY settings. The cost covers training and costs for covering staff to attend training.
- We assume that the wear and tear on physical resources are minimal with no requirement to repurchase items over the three-year period.

Owing to the costs of administering testing, only 11.7 children were included per setting at randomisation. However, given that Concept Cat is a whole-class intervention, the total cost per setting is based on the number of children per year expected to benefit from Concept Cat. To calculate the child per year estimate of cost, Concept Cat provided the evaluation team with the number of children aged three to four in each setting, the average of this was 38.5 children per setting, rounded up to 39 for the purposes of the cost calculations, with a range of 9 to 105 children per setting, and a mode of 32 children per setting. We assumed that each year Concept Cat would reach 39 children, for a total reach of 117 children over the course of three years.

While survey data on additional expenses borne by settings during the evaluation were collected as part of the IPE, the low response rate led to the decision to exclude these figures from the cost evaluation.

While not typically considered a ‘direct’ cost for settings, the time practitioners are expected to spend preparing and delivering Concept Cat sessions is an important factor when evaluating broader resources needed for the intervention. This information is detailed in Table 68 below. During the initial six weeks, it is anticipated that practitioners will spend approximately 30 minutes per week on planning. However, as practitioners become more familiar with the session structure, this planning time is expected to decrease to 15 minutes per week. Additionally, the delivery of the sessions themselves is expected to require about one hour per week, although this may be less for smaller class sizes. As such, the total estimated time required by practitioners in delivering Concept Cat to children is 1.25 hours per week.

Table 68: Staff time required in preparing and delivering Concept Cat

Item	Type of cost	Hours spent per week (average per setting)
Time spent preparing for sessions	Recurring cost	0.25
Time spent delivering sessions	Recurring cost	1.0
	Total:	1.25

Conclusion

Table 69: Key conclusions

Key conclusions	
1.	Children in Concept Cat settings made, on average, two months' additional progress in understanding conceptual vocabulary, compared to children in other settings. This result has a moderate to high security rating.
2.	Among children with Early Years Pupil Premium (EYPP), those in Concept Cat settings made three months' additional progress in conceptual vocabulary compared to those in other settings. These results may have lower security than the overall findings because of the smaller number of EYPP children.
3.	Children in Concept Cat settings demonstrated, on average, two months' additional progress in their early numeracy development, compared to children in other settings.
4.	Compliance and fidelity to the programme design were moderate to high across settings. However, it was found that settings were not delivering additional activities for focus children, as set out in the design, indicating low fidelity in this regard.
5.	Concept Cat appears to have facilitated parents' involvement in the setting and understanding of their children's learning and development.

Impact evaluation and IPE integration

The results of this evaluation suggest that participating in Concept Cat is associated with two months' worth of progress in children's understanding of conceptual vocabulary and early numeracy skills. The impact appears to be accentuated for settings that implement the intervention as intended. However, these results need to be taken with some caution. Results of the conceptual vocabulary test suggest the test was too easy (i.e. notable ceiling effects). As such, the test may not be capturing the full upper range of scores—that is, scores representing higher ability levels. The evaluation provides evidence of a positive impact on children's conceptual vocabulary attainment, however, due to ceiling effects with the primary outcome, this may be an underestimation, and the true impact of the programme could be greater.

However, given that Concept Cat is a language development intervention, it is hard to conceive that the programme could improve children's numeracy independently of their conceptual vocabulary development. This mediating effect of conceptual vocabulary on numeracy is substantiated by previously discussed meta-analytical literature (see Lin *et al.,* 2021). As such, the impact of Concept Cat on numeracy being larger than its impact on conceptual vocabulary—despite the latter being a mediator for the former—supports the notion of ceiling effects obscuring the true impact of Concept Cat on conceptual vocabulary outcomes.

The evaluation also found an effect size for the EYPP subgroup in terms of impact on conceptual vocabulary development, but further analysis found no indication that being EYPP-eligible had a mediating effect on this impact. The effect size observed in the EYPP subgroup may also be partly influenced by the fact that less ceiling effects were observed for this subsample; more significant improvements in conceptual vocabulary were less likely to be 'lost' off the top, compared to in the overall sample. These findings suggest that the test instrument used to measure the primary outcome was unfit for purpose, potentially underestimating the impact of Concept Cat.

Positive appraisals of the programme by surveyed EYPs in treatment settings—all but one reporting that Concept Cat helped improve children's conceptual vocabulary—also potentially suggest that Concept Cat is indeed more effective than what the primary outcome measure has demonstrated. Additionally, the findings that both control and intervention settings were using other interventions, which target other aspects of speech, language, and communication demonstrate the ease of implementation of the Concept Cat programme alongside other programmes.

It is worth noting, however, that the WellComm toolkit, which teaches opposite concepts opposed to singular concepts, was concurrently implemented in a slightly larger number of intervention settings when compared to control settings. Nevertheless, while the Wellcom toolkit may influence practitioners' competences, there is little to suggest any impact on child-level outcomes (Dysart and Code, 2024). As such, given the difference between control and intervention settings was minimal, it is unlikely that the WellComm toolkit has contributed to the improved outcomes observed in children who received Concept Cat. In sum, the overall effect of Concept Cat is potentially larger than what was found in this trial, though it is impossible to say this with certainty.

Compliance with the required training was high, with participating EYPs expressing that the training adequately prepared them for implementation. In addition, Concept Cat coaches were described by EYPs as integral to the support provided during programme implementation, despite a small number of settings struggling to accommodate coaching visits due to scheduling conflicts. Group supervision sessions were also characterised by EYPs as being helpful for exchanging knowledge and good practices, despite low compliance with this support modality, owing mainly to staffing issues.

The training and support provided to EYPs are likely to have translated into the generally high degree of fidelity and quality observed in the implementation of Concept Cat. The findings show that while, at the beginning of programme implementation, the 'Activate' and 'Review' elements of the STAR approach in the programme were implemented with some degree of inconsistency—that is, with only medium fidelity, Concept Cat coaches played a key role to help establish higher implementation fidelity. Once the programme was successfully embedded the need for the Concept Cat coaches was minimal, which supports the sustainability of the programme.

EYPs reported that while focus children did receive the Concept Cat programme, most settings did not provide additional support such as repeating the Concept Cat story and speaking more to families of these children. Nonetheless, despite low fidelity in terms of providing additional support for focus children, overall compliance in terms of the required week-on-week child attendance (i.e. at least 15 hours per week) was high. However, it is important to note that attendance data was available for only a small number of participating children, and the findings regarding compliance should be understood in this context. Broadly speaking, however, the high degree of fidelity and quality in implementing Concept Cat may be attributed to the programme's ease of implementation, as reported by surveyed EYPs.

The parent/carer surveys also indicate that the outcomes of Concept Cat may also extend to children's families. In line with the hypotheses laid out in the theory of change, Concept Cat was found to have increased parents' engagement in their children's learning, as well as parents' involvement with settings, between baseline and endline. Settings used a variety of modalities in their efforts to engage with parents, but the modality itself mattered less than the extent to which it suited children's families' context and needs. Additionally, engagement with the Concept Cat programme at home was also driven by children's enthusiasm of the programme.

In sum, the impact evaluation has shown that Concept Cat has a positive impact on the intended outcomes of conceptual vocabulary and numeracy. However, the trial may have underestimated this impact, owing to measurement failure. The IPE further supports the notion of programme effectiveness, with respect to both children and their families. Moreover, the fact that all survey EYPs said they were going to continue implementing the programme after the trial, and that children clearly enjoy Concept Cat and this facilitated home implementation of the programme, shows the desirability of the programme overall.

Recommendations

Given the findings outlined above, the evaluation team make the following recommendations for the programme moving forward and have amended the theory of change based on these recommendations (see Appendix Q):

- consider three-hour training for all EYPs;
- amend the Concept Cat coach logs to reflect the differing ways in which settings provide families with introductory information and how settings regularly engage families in activities;
- reflect on how best to engage EYPs in good practice networks, which takes in to account the impact of staffing issues impacting on engagement;
- give guidance to EYPs on how to manage the teaching of opposite concepts outlined in the WellComm toolkit;
- revisit the appropriateness of additional activities currently required for focus children and/or how training and coach support may facilitate this;
- reflect on how training may incorporate support for SEND children; and

- consider how long Concept Cat coaches are needed to support settings and whether there should be variation in visits at the start of the programme compared to the end of the programme.

Limitations and lessons learned

Measurement failure remains the biggest limitation of this evaluation. As discussed earlier, the instrument used to measure conceptual vocabulary development at baseline and endline displayed ceiling effects. Since the instrument could not capture the full range of higher scores, the impact of Concept Cat on the primary outcome is likely to have been underestimated. While this instrument was reported to have been psychometrically tested with similar populations, the evaluation could have benefited from a pilot administration to ensure this. However, this would not have been possible, given the evaluation's budget and resourcing constraints.

Therefore, in future evaluations of Concept Cat, alternative testing instruments to measure conceptual vocabulary development should be considered.

This qualitative data collected from EYPs has also indicated a mixed picture of Concept Cat's effectiveness on EAL and SEND children. The impact evaluation is unable to clarify this mixed picture, given that the trial has not been powered to detect an impact on the EAL and SEND subgroups. However, given the mixed picture offered by the IPE, subsequent evaluations with sufficient sample sizes for detecting an effect in the EAL and SEND subgroups would also be warranted.

The lack of actual child-level attendance data is also an important limitation of this evaluation. While the evaluation team initially sought to use actual child attendance records collated by settings as the pupil-level compliance metric, it was found that not all settings collected attendance data to this level of granularity. As such, to ensure consistency in measuring compliance, the evaluation team opted instead to use attendance patterns reported by settings—that is, a binary variable indicating whether a child was scheduled to attend the setting at least 15 hours each week. Even in collecting this proxy variable, the evaluation still encountered a moderate degree of missingness in the child-level compliance data (i.e. 27%).

To complement data on attendance patterns and make the compliance analysis more robust, the evaluation team combined data on attendance patterns with data on the number of words taught by settings over the delivery period (with 30 words being the compliance threshold). These two data points were taken together to create a binary compliance variable, where compliance meant that a child attended the setting for at least 15 hours a week and that their setting taught at least 30 words throughout the duration of Concept Cat. In the absence of actual child attendance data from EY settings, this was a reasonable proxy that could be used in subsequent evaluations of Concept Cat and other EY trials.

Future research and publications

Existing meta-analytical evidence has shown an association between conceptual vocabulary and numeracy development (see Lin *et al.,* 2021), and the results of this evaluation appear to corroborate this evidence, noting how Concept Cat has produced improvement in the conceptual vocabulary and numeracy of participating children. Qualitative data collected from EYPs has also indicated a mixed picture of Concept Cat's effectiveness on EAL and SEND children. The impact evaluation is unable to clarify this mixed picture, given that the trial has not been powered to detect an impact on the EAL and SEND subgroups. However, given the mixed picture offered by the IPE, subsequent evaluations with sufficient sample sizes for detecting an effect in the EAL and SEND subgroups would also be warranted.

In addition, future studies could examine whether gains in language and numeracy development deriving from children's participation in Concept Cat are sustained in later years. The cumulative nature of learning advantage (or disadvantage) beginning in earlier life is well-documented in the existing literature. As such, a longitudinal analysis of relevant outcomes for children who have participated in Concept Cat can inform decisions around future design and implementation of the programme, as well as other interventions with similar target outcomes.

References

Blachowicz, C. and Fisher, P. (2015) '*Teaching Vocabulary in All Classrooms*, *5th edition'.* New York: Pearson.

Clarke, D., Romano, J.P. and Wolf, M. (2019) *'The Romano-Wolf Multiple Hypothesis: Correction in Stata'.* IZA Institute of Labor Economics, Article 12845. Bonn: IZA Institute of Labor Economics. Available at: https://docs.iza.org/dp12845.pdf (accessed 09/07/2025).

Clements, D.H., Sarama, J., Spitler, M.E., Lange, A.A. and Wolfe, C.B. (2011) *'Mathematics Learned by Young Children in an Intervention Based on Learning Trajectories: A Large-Scale Cluster Randomized Trial'.* Journal for Research in Mathematics Education, 42: 2, 127–166.

Data Protection Act 2018. Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents (accessed 09/07/2025).

Dawson, A., Stokes, L., Huxley, C., Runge, J., Takala, H., Manzoni, C., Hudson-Sharp, C., and Williams, C. (2020) '*Early Years Toolbox: Pilot Report*'. London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Early_Years_Toolbox_Report_(final).pdf (accessed 09/07/2025).

Dimova, S., Illie, S., Rosa Brown, E., Broeks, M., Culora, A., and Sutherland, A. (2020) '*The Nuffield Early Language Intervention. Evaluation Report*'. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Nuffield_Early_Language_Intervention.pdf (accessed 09/07/2025).

Dong, N. and Maynard, R. (2013) 'PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies'. *Journal of Research on Educational Effectiveness*, 6: 1, 24–67. https://doi.org/10.1080/19345747.2012.673143

Dysart, E. and Code, A. (2024) 'The WellComm Toolkit: Impact on Practitioner Skills and Knowledge and Implications for Evaluation Research'. *Education Sciences*, 14: 3, 263. https://doi.org/10.3390/educsci14030263

Eadie, P., Cattram, N., Carlin, J., Bavin, E., Bretherton, L. and Reilly, S. (2014) 'Stability of Language Performance at 4 and 5 Years: Measurement and Participant Variability'. *International Journal of Language & Communication Disorders*, 49: 2, 215–17. https://doi.org/10.1111/1460-6984.12065

Education Act 1996, s.10. Available at: https://www.legislation.gov.uk/ukpga/1996/56/contents (accessed 09/07/2025).

Education Endowment Foundation (EEF). (2022) *'Statistical Analysis Guidance for EEF Evaluations'.* London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1751533489 (accessed 09/07/2025.

Education Endowment Foundation (EEF). (2023a) '*Early Years Measures Database'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-outcome-measures-and-databases/early-years-measures-database-2 (accessed 09/07/2025).

Education Endowment Foundation (EEF). (2023b) *'Cost Evaluation Guidance for EEF Evaluations'.* London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/Cost-Evaluation-Guidance-Feb_2023.pdf?v=1743094546 (accessed 09/07/2025).

Education Endowment Foundation (EEF). (2024). *Putting Evidence to Work - A School's Guide to Implementation*. EEF. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/implementation (accessed 09/07/2025)

Fernald, A., Marchman, V.A. and Weisleder, A. (2013) 'SES Differences in Language Processing Skill and Vocabulary are Evident at 18 Months'. *Developmental Science*, 16: 2, 234–48. https://doi.org/10.1111/desc.12019

Fricke, S., Bowyer-Crane, C., Haley, A., Hulme, C. and Snowling, M. (2012) 'Efficacy of Language Intervention in the Early Years'. *Journal of Child Psychology and Psychiatry*, 54: 3, 280–90. https://doi.org/10.1111/jcpp.12010

GDPR. (2016a). Regulation (EU) 2016/679 of the European Parliament and of the Council. '*Article 6: Lawfulness of Processing. 1, (f)'.* legislation.gov.uk. Available at: https://www.legislation.gov.uk/eur/2016/679/article/6# (accessed 09/07/2025).

GDPR. (2016b). Regulation (EU) 2016/679 of the European Parliament and of the Council. '*Article 9: Processing of Special Categories of Personal Data 2, (g), (j)'.* legislation.gov.uk. Available at: https://www.legislation.gov.uk/eur/2016/679/article/9 (accessed 09/07/2025).

Golinkoff, R.M., Hoff, E., Rowe, M.L., Tamis-LeMonda, C.S., and Hirsh-Pasek, K. (2019) 'Language matters: Denying the existence of the 30-million-word gap has serious consequences'. *Child Development*, 90: 3, 985-992. https://doi.org/10.1111/cdev.13128

Hart, B. and Risley, T.R. (1995) '*Meaningful Differences in the Everyday Experience of Young American Children'.* New York, NY: Paul H Brookes Publishing.

Hedges, L.V. (2007) 'Effect Sizes in Cluster-Randomized Designs'. *Journal of Educational and Behavioral Statistics*, 32: 4, 341–70. https://doi.org/10.3102/1076998606298043

Hoff, E. (2013) *Interpreting the Early Language Trajectories of Children From Low-SES and Language Minority Homes: Implications for Closing Achievement Gaps.* Developmental Psychology, 49: 1, 4–14. https://doi.org/10.1037/a0027238

Hopkins, T., Harrison, E., Coyne-Umfreville, E. and Packer, M. (2022) 'A Pilot Study Exploring the Effectiveness of a Whole-School Intervention Targeting Receptive Vocabulary in the Early Years: Findings From a Mixed Method Study Involving Students as Part of a Practice-Based Research Placement'. *Child Language and Teaching Therapy*, 38: 2, 212–29. https://doi.org/10.1177/02656590221088210

Howard, S.J., Neilsen-Hewett, C., de Rosnay, M., Melhuish, E.C. and Buckley-Walker, K. (2022) 'Validity, Reliability and Viability of Pre-School Educators' Use of Early Years Toolbox Early Numeracy'. *Australasian Journal of Early Childhood*, 47: 2, 92–106. https://doi.org/10.1177/18369391211061188Hutchison, D. and Styles, B. (2010). *A Guide to Running Randomised Controlled Trials for Educational Researchers.* Slough: NFER.

Law, J., Charlton, J., Dockrell, J., Gascoigne, M., McKean, C. and Theakston, A. (2017) *'Early Language Development: Needs, Provision, and Intervention for Preschool Children From Socio-economically Disadvantage Backgrounds'.* London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/guidance/Law_et_al_Early_Language_Development_final.pdf?v=1745504439 (accessed 09/07/2025).

Lin, X., Peng, P. and Zeng, J. (2021) 'Understanding the Relation Between Mathematics Vocabulary and Mathematics Performance: A Meta-Analysis'. *The Elementary School Journal*, 121: 3, 504–40. https://doi.org/10.1086/712504

LeFevre, J.A., Skwarchuk, S.L., Smith-Chant, B.L., Fast, L., Kamawar, D. and Bisanz, J. (2009) 'Home Numeracy Experiences and Children's Math Performance in the Early School Years'. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 41, 2, 55–66. https://doi.org/10.1037/a0014532

McBee, M. (2010) 'Modeling Outcomes With Floor or Ceiling Effects: An Introduction to the Tobit Model'. *Gifted Child Quarterly*, 54: 4, 314–20. https://doi.org/10.1177/0016986210379095

Melhuish, E., Quinn, L., Sylva, K., Sammons, P., Siraj-Blatchford, I. and Taggart, B. (2013) 'Preschool Affects Longer Term Literacy and Numeracy: Results From a General Population Longitudinal Study in Northern Ireland'. *School Effectiveness and School Improvement*, 24: 2, 234–50. https://doi.org/10.1080/09243453.2012.749796

Parsons, S. and Branagan, A. (2016) '*Word Aware 2: Teaching Vocabulary in the Early Years'*. Abingdon, Oxon: Routledge.

Purpura, D. J., Schmitt, S. A., and Ganley, C. M. (2017). Foundations of mathematics and literacy: The role of executive functioning components**.** *Journal of Experimental Child Psychology*, *153,* 15-34.

Rowe, M.L. and Zuckerman, B. (2016) 'Promoting Language Development: Timing and Technique'. *Pediatrics,* 138: 4, e20161332.

Scarborough, H.S. (2009) 'Connecting Early Language and Literacy to Later Reading (Dis)abilities: Evidence, Theory, and Practice'. In F. Fletcher-Campbell, J. Soler, and G. Reid (eds.) *Approaching Difficulties in Literacy Development: Assessment, Pedagogy and Programmes, First Edition*. *Part One: Theoretical Understandings: Implications for Practice.* London: SAGE Publications Ltd.

Schafer JL. (1999) Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, 8: 1, 3–15. https://doi.org/10.1177/096228029900800102

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Maturi, T., Higgins, S., Einbeck, J., Culliney, M. and Demack. S. (2023) 'Improving Power Calculations in Educational Trials'. London: Education Endowment Foundation. Available at: Work_Package_2023-WP6_18_09_2023_FINAL.pdf (accessed 09/07/2025).

Torgerson, D., & Torgerson, C. (2008). Designing and running randomised trials in health, education and the social sciences. United Kingdom: Palgrave Macmillan.

Uttl, B. (2005) 'Measurement of Individual Differences: Lessons From Memory Assessment in Research and Clinical Practice'. *Psychological Science*, 16: 6, 460–67. https://doi.org/10.1111/j.0956-7976.2005.01557.x

Verdine, B.N., Golinkoff, R.M., Hirsh-Pasek, K. and Newcombe, N.S. (2014) 'Finding the Missing Piece: Blocks, Puzzles, and Shapes Fuel School Readiness'. *Trends in Neuroscience and Education*, 3: 1, 7–13. https://doi.org/10.1016/j.tine.2014.02.005

Wiig, E.H., Semel, E. and Secord W.A. (2017). 'Clinical Evaluation of Language Fundamentals – Fifth Edition: CELF-5 UK'. Pearson Clinical Assessment UK. Available at: www.pearsonclinical.co.uk/en-gb/Store/Professional-Assessments/Speech-%26-Language/Clinical-Evaluation-of-Language-Fundamentals---Fifth-Edition/p/P100009245?msclkid=2103f9a960b91f23b65f91c2ebbf9723&utm_source=bing&utm_medium=cpc&utm_campaign=SLT_products&utm_term=celf5&utm_content=CELF-5 (accessed 09/07/2025).

Appendix A: EEF cost rating

Table A1: Cost rating

Cost rating	Description
£ £ £ £ £	*Very low:* less than £80 per child per year.
£ £ £ £ £	*Low:* up to about £200 per child per year.
£ £ £ £ £	*Moderate:* up to about £700 per child per year.
£ £ £ £ £	*High:* up to £1,200 per child per year.
£ £ £ £ £	*Very high:* over £1,200 per child per year.

Appendix B: Security classification of trial findings

OUTCOME: Early conceptual vocabulary (CELF Preschool-2 UK 'Basic Concepts' subtest)

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%	**4**				
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%			Adjustment for threats to internal validity **[-1]**		**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%					
0	No comparator	>=0.6	>50%					

Threats to validity	Risk rating	Comments
Threat 1: Confounding	**Moderate**	There was an imbalance of .11 SD on the pre-test measure, but relative balance is achieved on other school and pupil characteristics between intervention and control groups.
Threat 2: Concurrent Interventions	**Low**	There was some evidence of concurrent interventions, but this in line with normal activity in EY settings.
Threat 3: Experimental effects	**Low**	No evidence of experimental effects.
Threat 4: Implementation fidelity	**Low**	Intervention is reasonably well implemented although some practitioners struggled to deliver additional optional activities for focus children.
Threat 5: Missing Data	**Low**	67 and 72 children lost to follow-up, but no loss of settings, no further missing data to assess. Average of 13.3% attrition across the groups.
Threat 6: Measurement of Outcomes	**Moderate**	Ceiling effects observed for the primary outcome and accounted for analytically. The observed positive impact may underestimate the true effect of the programme.
Threat 7: Selective reporting	**Low**	There appears to be no issues related to selective reporting.

- **Initial padlock score: 4 padlocks –** two-armed, waitlisted cluster randomised trial; MDES of .253 at randomisation stage and 13.3% pupil attrition led to a loss of 1 padlock.
- **Reason for adjustment for threats to validity**: -1 padlock. There is a moderate risk associated with confounding given a baseline imbalance in the pre-test measure, but which is accompanied by high statistical uncertainty. A second moderate risk is associated with the

measurement of the primary outcome as highlighted by the evaluator. There is a similar direction of bias associated with these two moderate risks which on balance is sufficient to justify the reduction of 1 padlock.

- FINAL PADLOCK SCORE: 3

Appendix C: Effect size estimation

Table C1: Primary analysis effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.63	0.5	460 (67)	7.7	442 (71)	8.4	8.07

Table C2: Secondary analysis effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Concepts and Following Directions' subtest	0.45	0.23	460 (67)	20.31	442 (71)	19.16	19.71
EY Toolbox ENA	2.88	1.68	444 (83)	171.05	416 (97)	168.46	160.78

Table C3: EYPP subgroup effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.56	0.67	85 (17)	6.7	86 (14)	7.22	6.97

Table C4: EYPP subgroup effect size estimation (ENA)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
EY Toolbox ENA	-0.68	-0.35	81 (21)	118.9	81 (19)	97.92	108.37

Table C5: EAL subgroup effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.38	0.24	109 (19)	14.62	103 (20)	9.83	12.32

Table C6: EAL subgroup effect size estimation (ENA)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
EY Toolbox ENA	3.04	0.91	99 (29)	199.19	99 (24)	178.04	188.51

Table C7: SEND subgroup effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	1.38	-0.05	28 (1)	18.1	32 (7)	17.86	17.98

Table C8: SEND subgroup effect size estimation (ENA)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
EY Toolbox ENA	6.38	0.13	23 (6)	208.51	31 (8)	147.37	173.19

Table C9: EYPP interaction model effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.65	0.50	460 (67)	7.7	431 (82)	8.42	8.07

Table C100: EYPP interaction effect size estimation (ENA)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
EY Toolbox ENA	2.93	-0.35	444 (83)	171.05	405 (108)	170.01	170.56

Table C11: EAL interaction model effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Prewschool-2 UK 'Basic Concepts' subtest	0.63	0.57	460 (67)	7.7	442 (71)	8.4	8.07

Table C12: EAL interaction effect size estimation (ENA)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
EY Toolbox ENA	2.88	0.91	444 (83)	171.05	416 (97)	168.46	169.78

Table C13: SEND interaction model effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.66	0.47	447 (80)	7.8	442 (71)	8.4	8.12

Table C14: SEND interaction effect size estimation (ENA)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
EY Toolbox ENA	3.06	0.13	431 (96)	169.71	416 (97)	168.46	169

Table C15: Primary analysis with control for age effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.63	0.50	460 (67)	7.7	442 (71)	8.4	8.07

Table C16: Primary analysis Tobit model effect size estimation (accounting for censoring)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.63	0.65	460 (67)	7.7	442 (71)	8.4	8.07

Table C17: CACE analysis effect size estimation

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	N (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.86	0.6	338 (189)	6.58	442 (71)	8.4	7.6

Table C18: Primary analysis effect size estimation (without contamination)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	N (missing)	Variance of outcome	Pooled variance
CELF Preschool-2 UK 'Basic Concepts' subtest	0.61	0.5	460 (67)	7.7	432 (81)	8.34	8.01

Table C19: SEND subgroup analysis (relegated from main report)

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	28 (1)	13.79 (12.14–15.44)	32 (7)	12.41 (10.88–13.93)	60 (28; 32)	-0.01 (-0.51–0.50)	0.99
EY Toolbox ENA	23 (6)	28.35 (22.1–34.59)	31 (8)	21.97 (17.51–26.42)	54 (23; 31)	0.01 (-0.48–0.51)	0.95

Table C20: SEND interaction models raw regression outputs

Outcome	Variable	Raw coefficient	Standard error	95% CI	P-value
CELF Preschool-2 UK 'Basic Concepts' subtest	Treatment	0.47	0.19	0.08 – 0.85	0.018
	SEND status	-1.26	0.41	-2.08 – -0.47	0.002
	Treatment SEND interaction	0.24	0.60	-0.88 – 1.36	0.674
EY Toolbox ENA	Treatment	1.67	0.82	0.10 – 3.31	0.038
	SEND status	-3.60	1.97	-7.45 – 0.43	0.081
	Treatment SEND interaction	-0.32	2.96	-6.24 – 5.53	0.906

Table C21: SEND interaction model (relegated from main report)

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	447 (80)	15.6 (15.34–15.86)	442 (71)	14.94 (14.67–15.21)	889 (447; 442)	0.25 (-0.287–0.77)	0.67
EY Toolbox ENA	431 (96)	33.98 (32.75–35.21)	416 (97)	30.92 (29.67–32.17)	847 (431; 416)	0.10 (-0.47–0.64)	0.91

Appendix D: Residual plots from analysis models

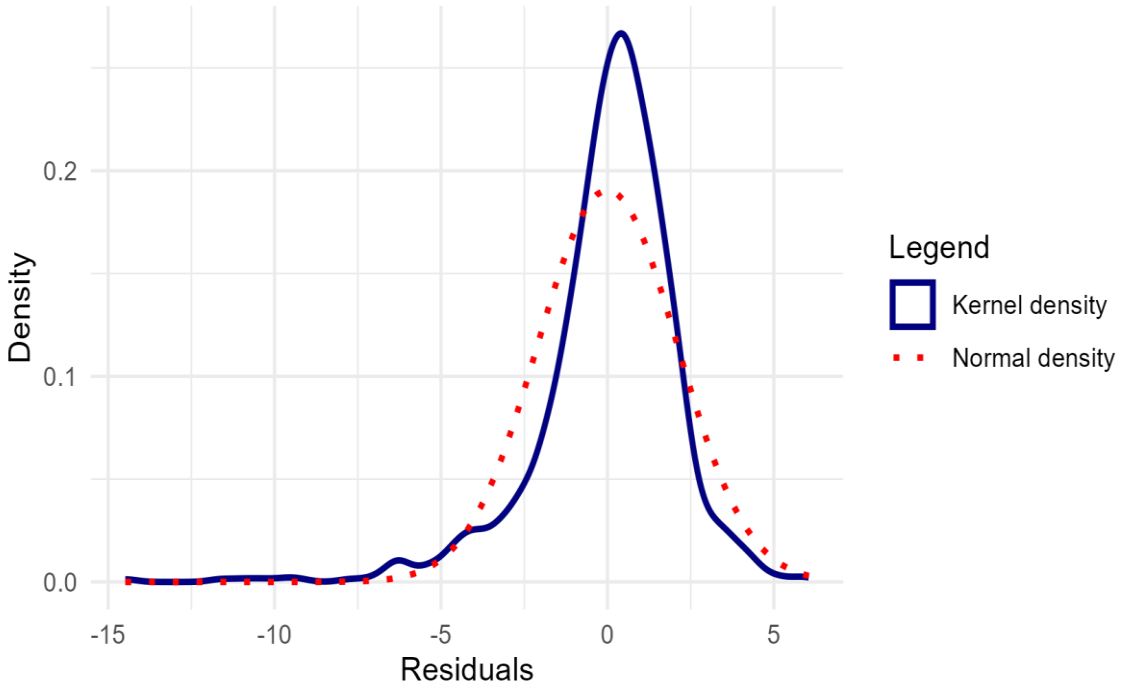


Figure D1: Primary analysis residual density plot

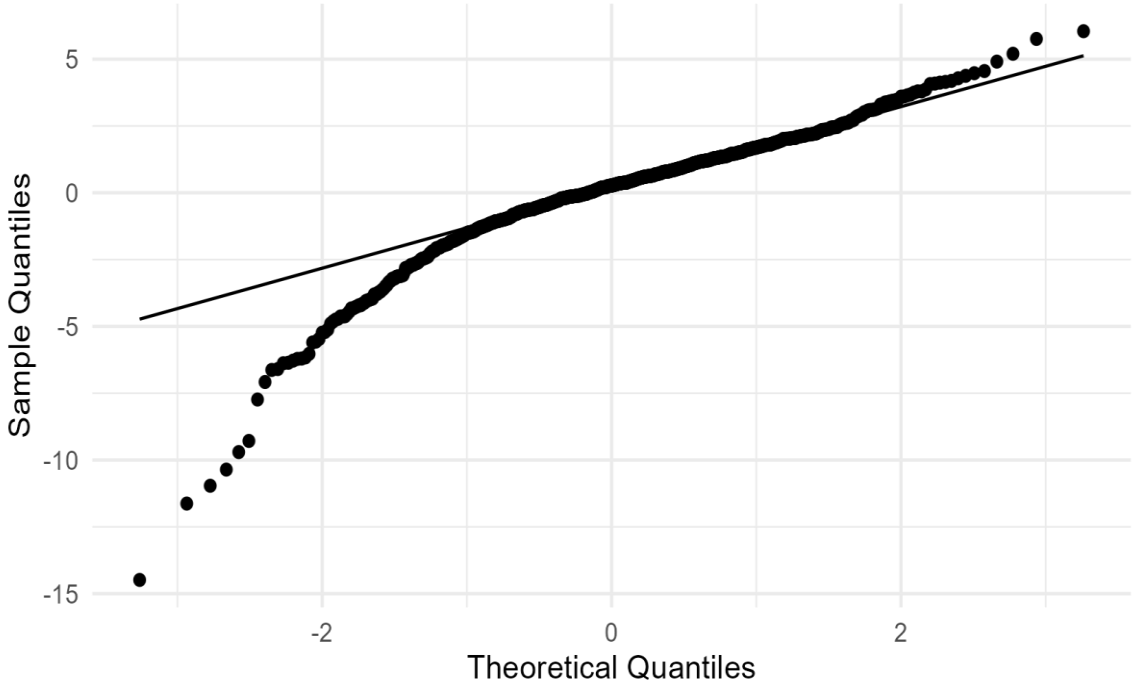


Figure D2: Primary analysis Q-Q (Quantile–Quantile) residual plot

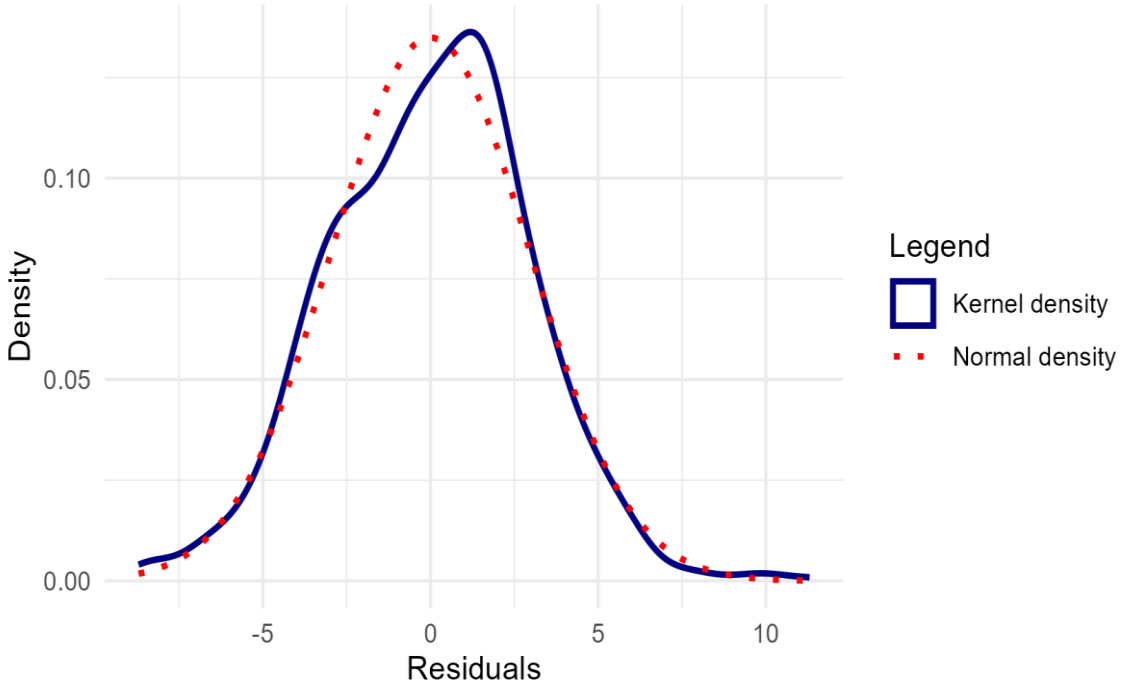


Figure D3: Secondary analysis residual density plot (CELF Preschool-2 UK 'Concepts and Following Directions' subtest)

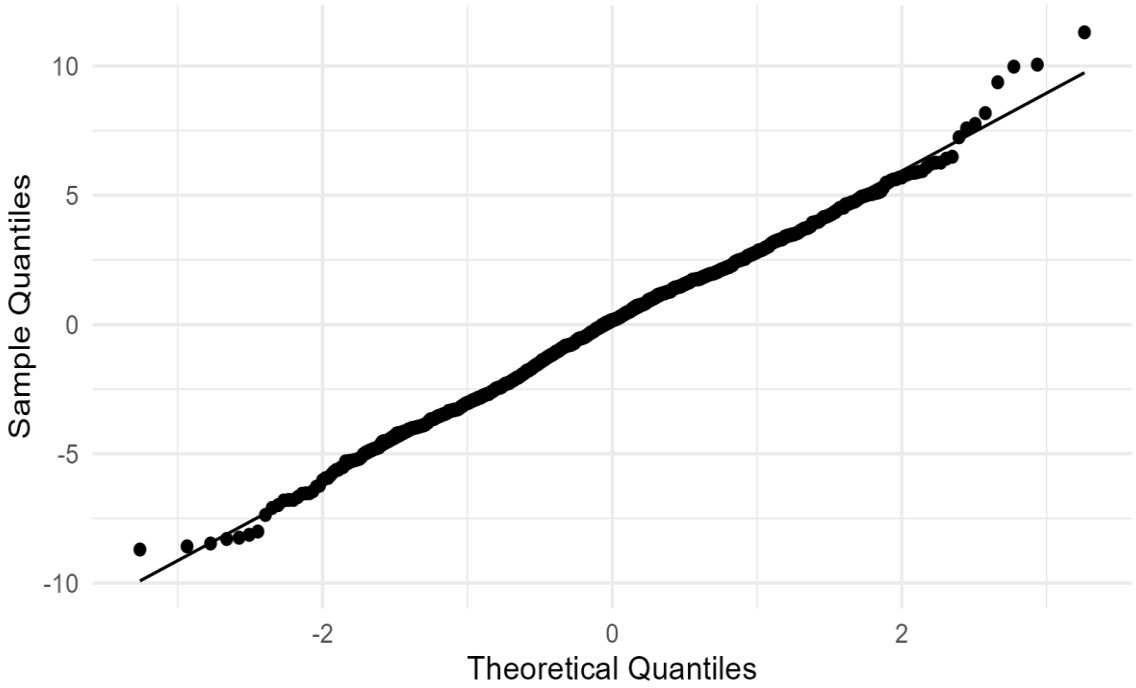


Figure D4: Secondary analysis Q-Q residual plot (CELF Preschool-2 UK 'Concepts and Following Directions' subtest)

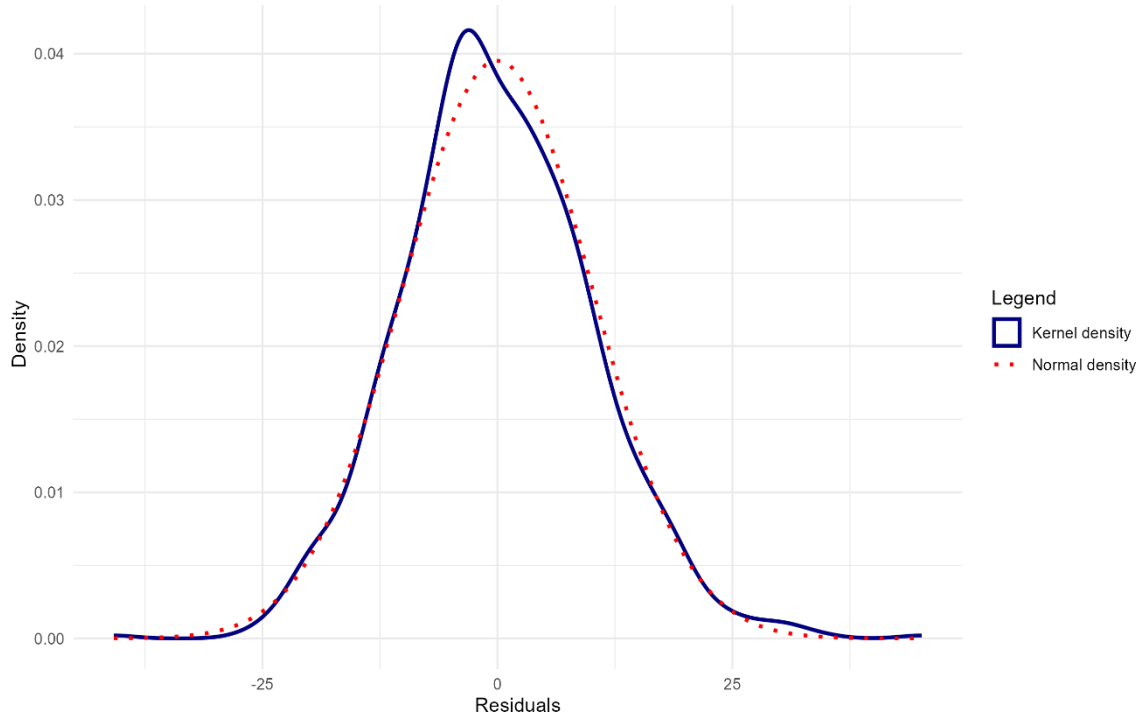


Figure D5: Secondary analysis residual density plot (EY Toolbox ENA)

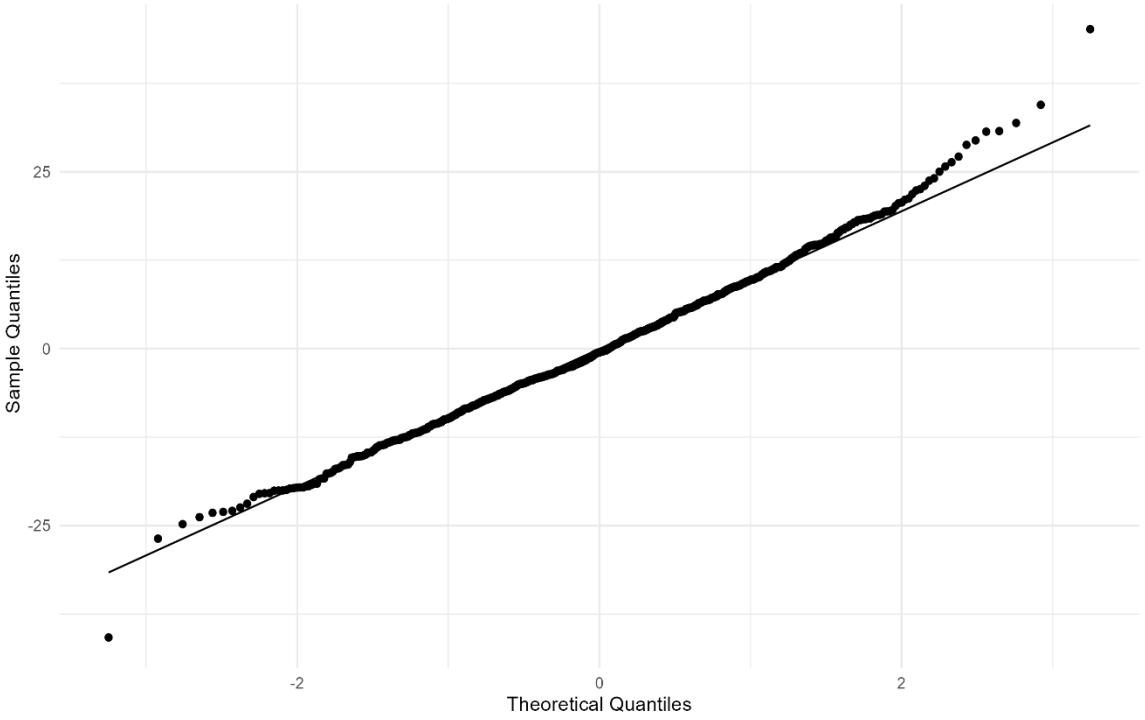


Figure D6: Secondary analysis Q-Q residual plot (EY Toolbox ENA)

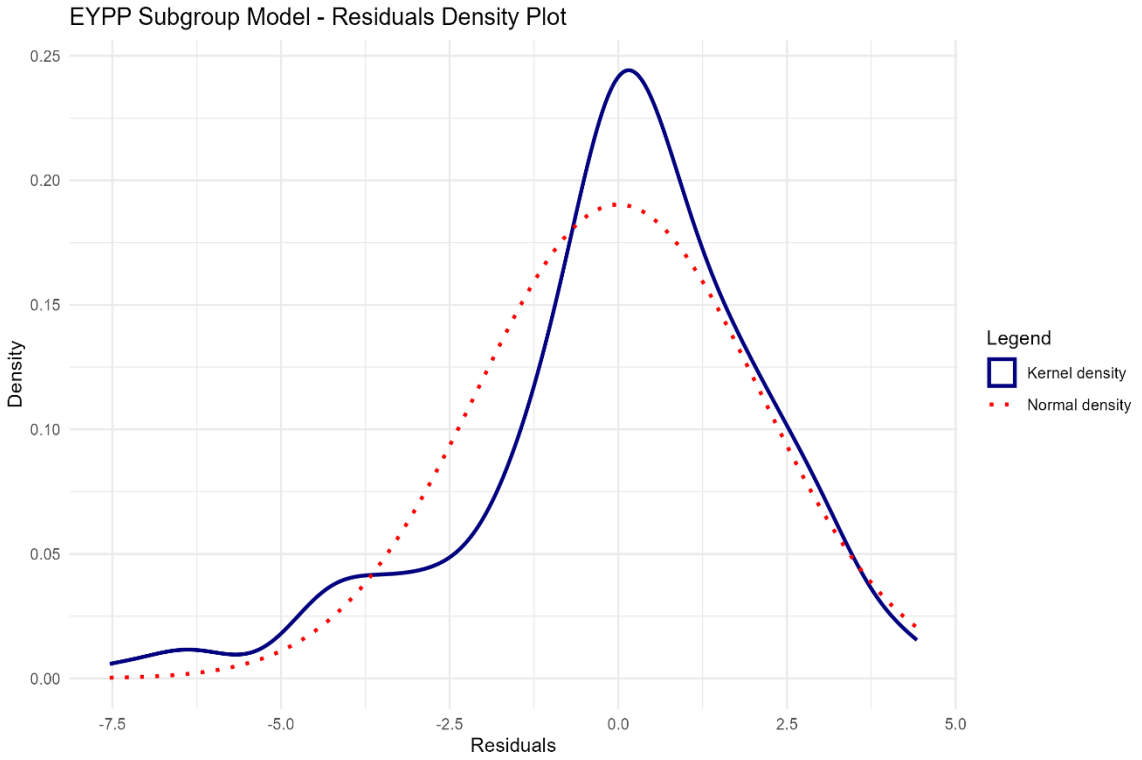


Figure D7: Residual kernel density plot (EYPP subgroup)

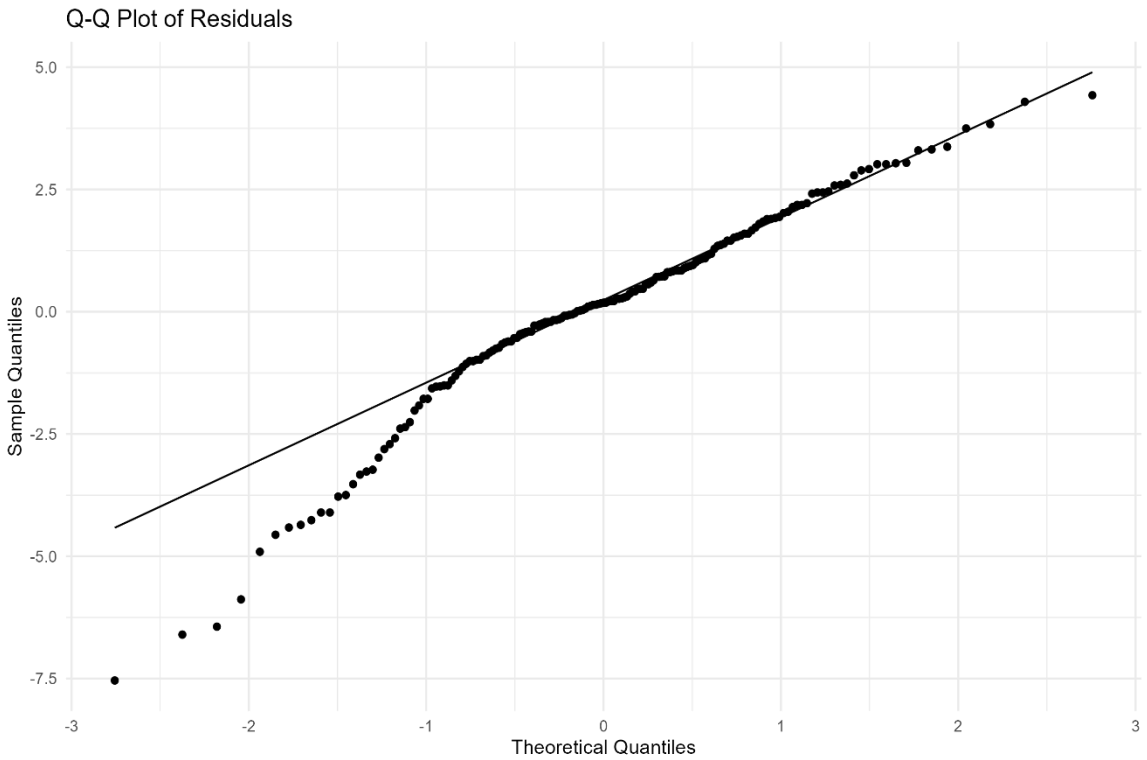


Figure D8: Residual Q-Q plot (EYPP subgroup)

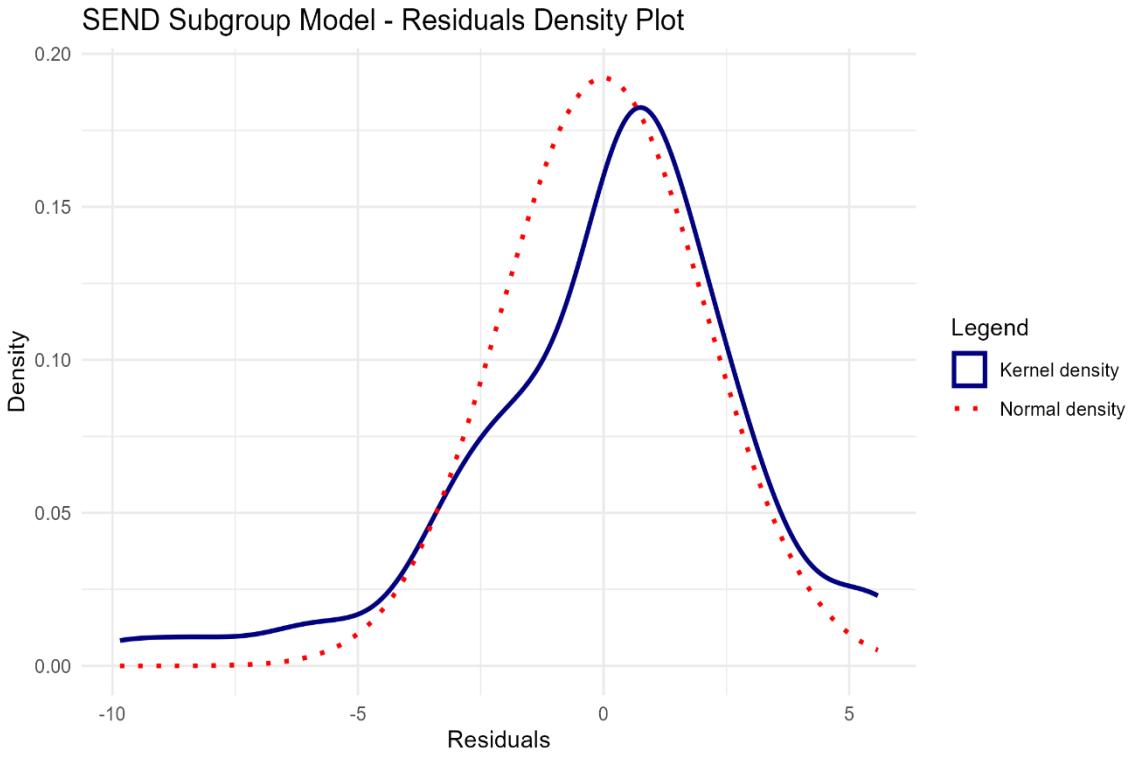


Figure D9: Residual kernel density plot (EAL subgroup)

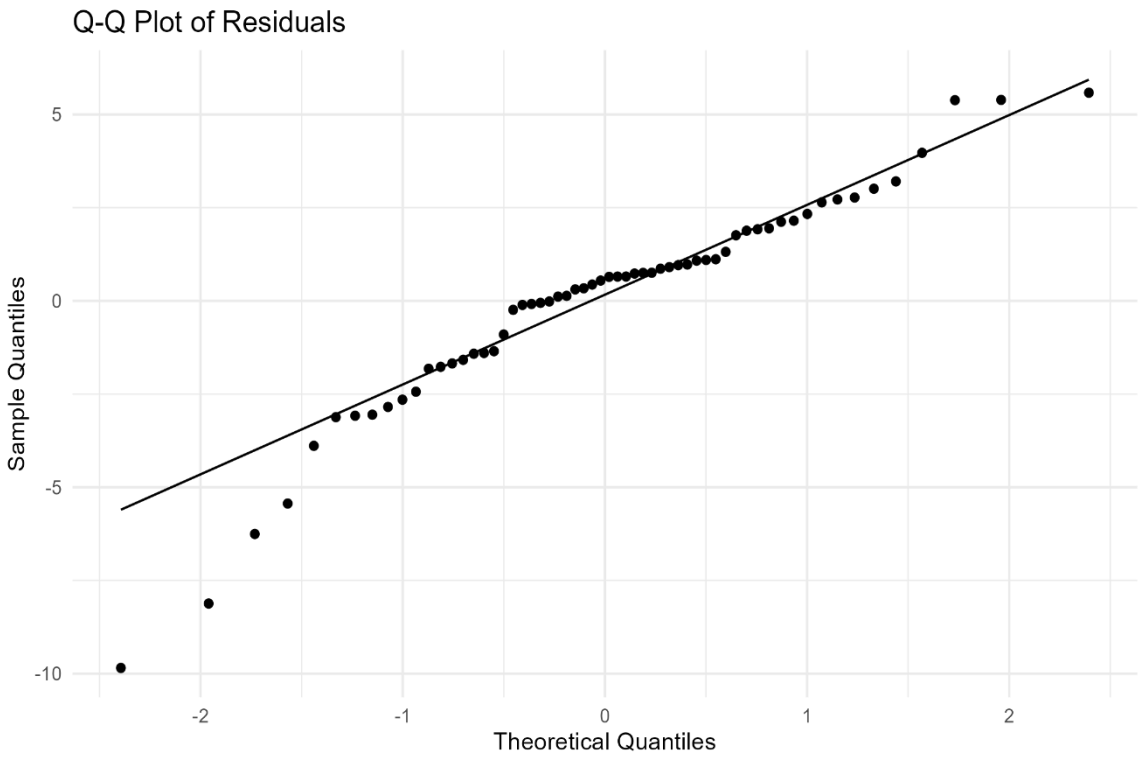


Figure D10: Residual Q-Q plot (EAL subgroup)

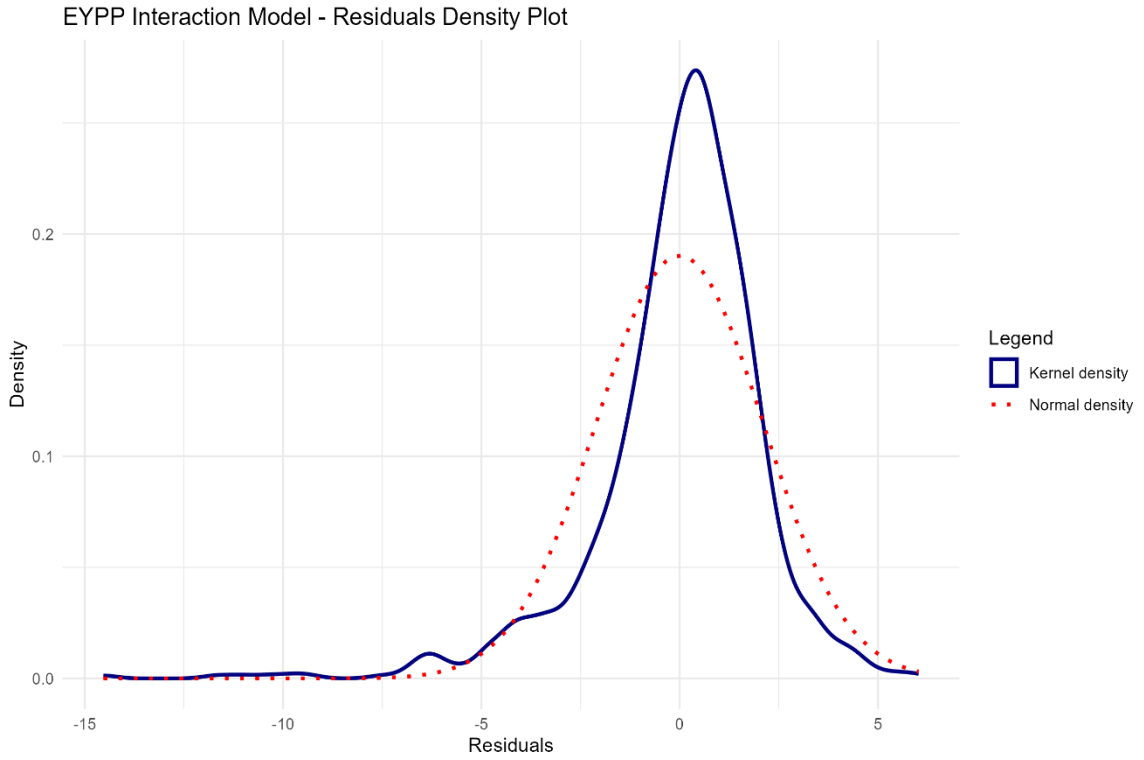


Figure D11: Residual kernel density plot (EYPP interaction model)

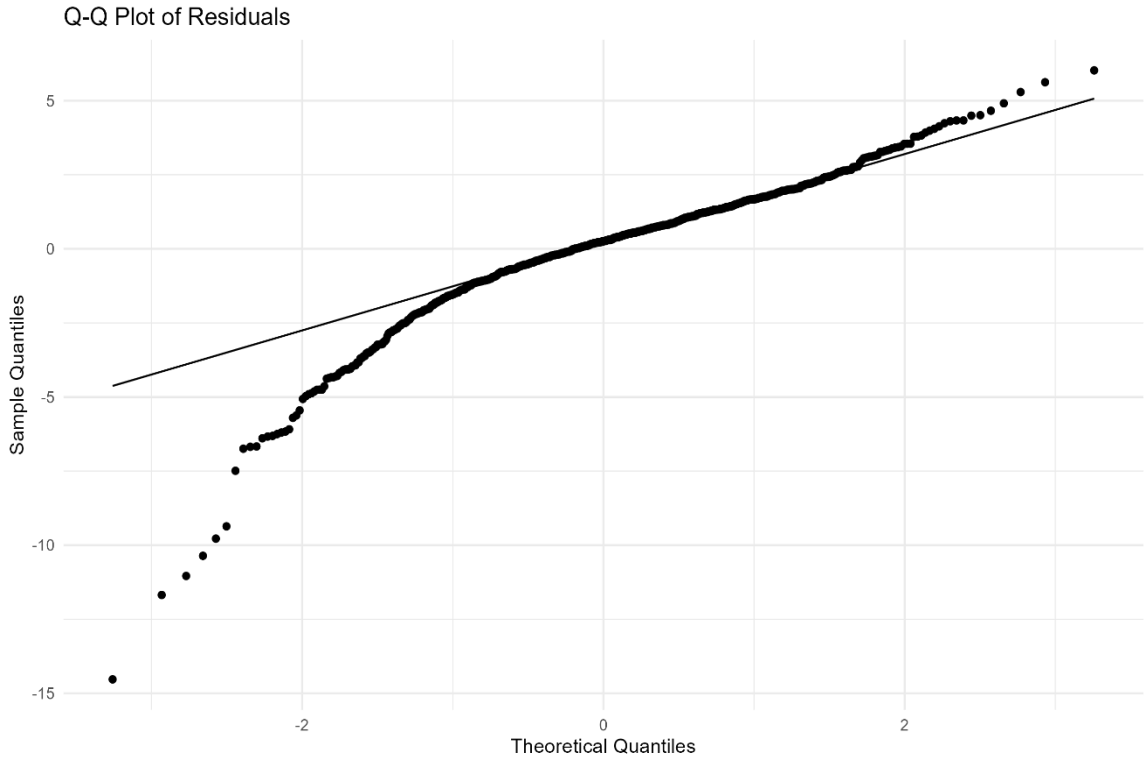


Figure D12: Residual Q-Q plot (EYPP interaction model)

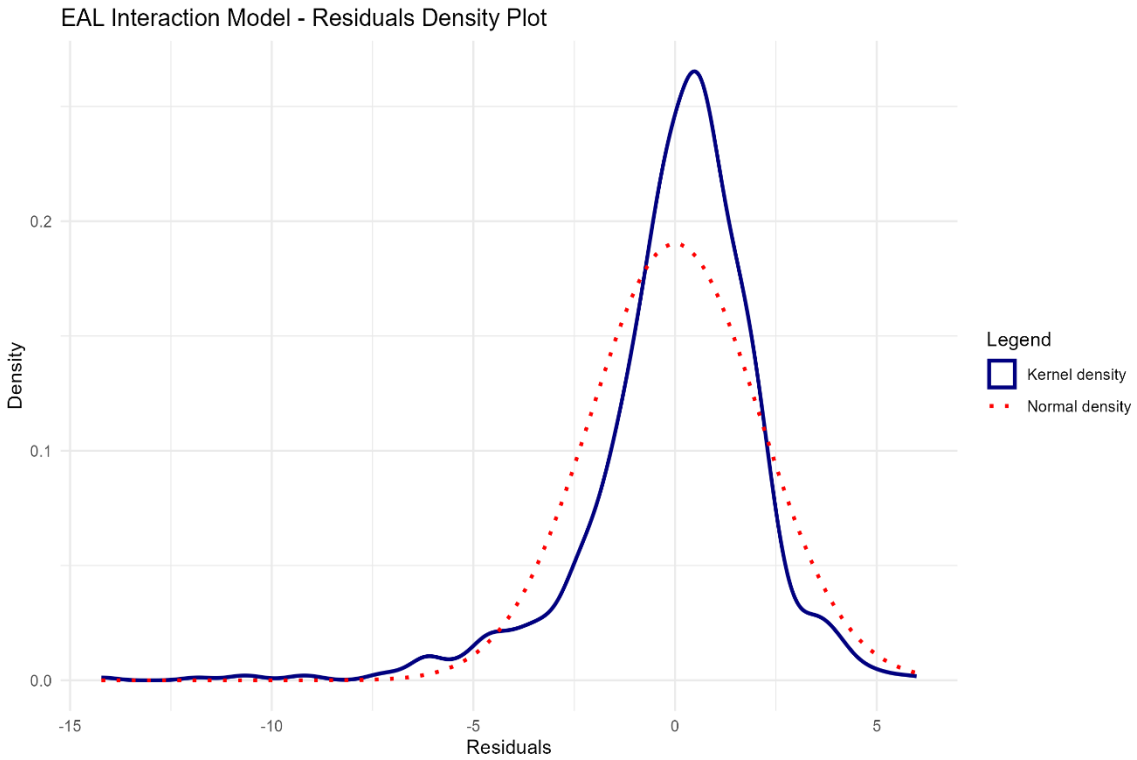


Figure D13: Residual kernel density plot (EAL interaction model)

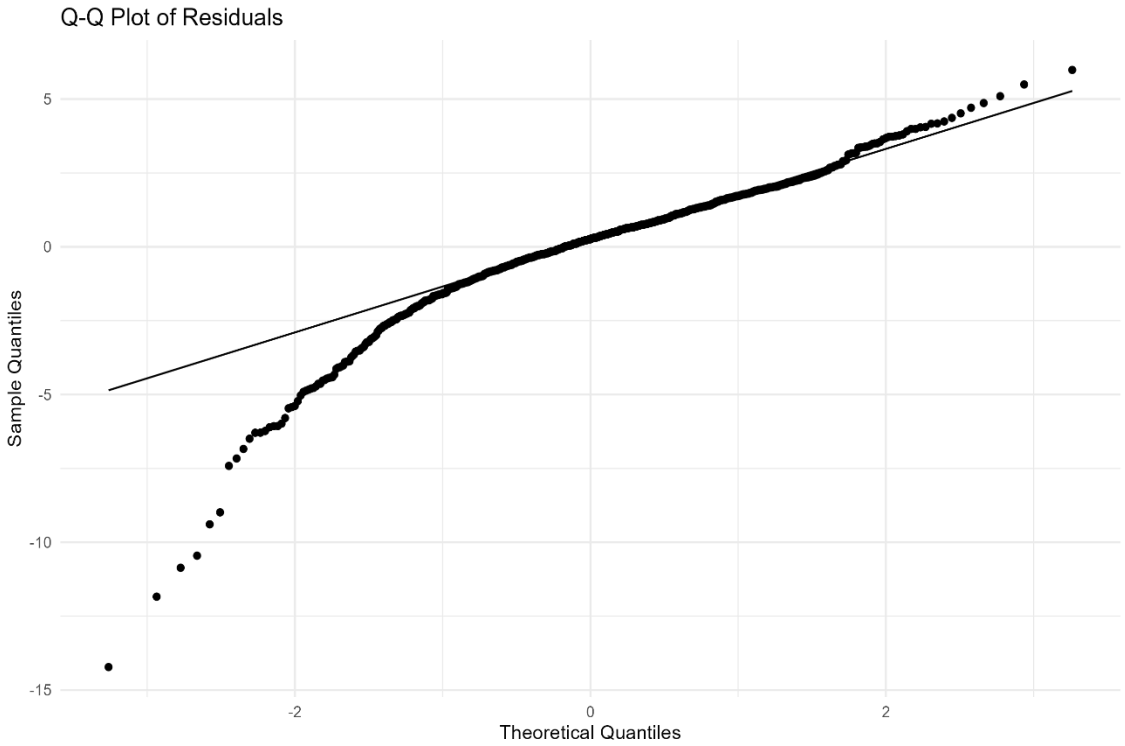


Figure D14: Residual Q-Q plot (EAL interaction model)

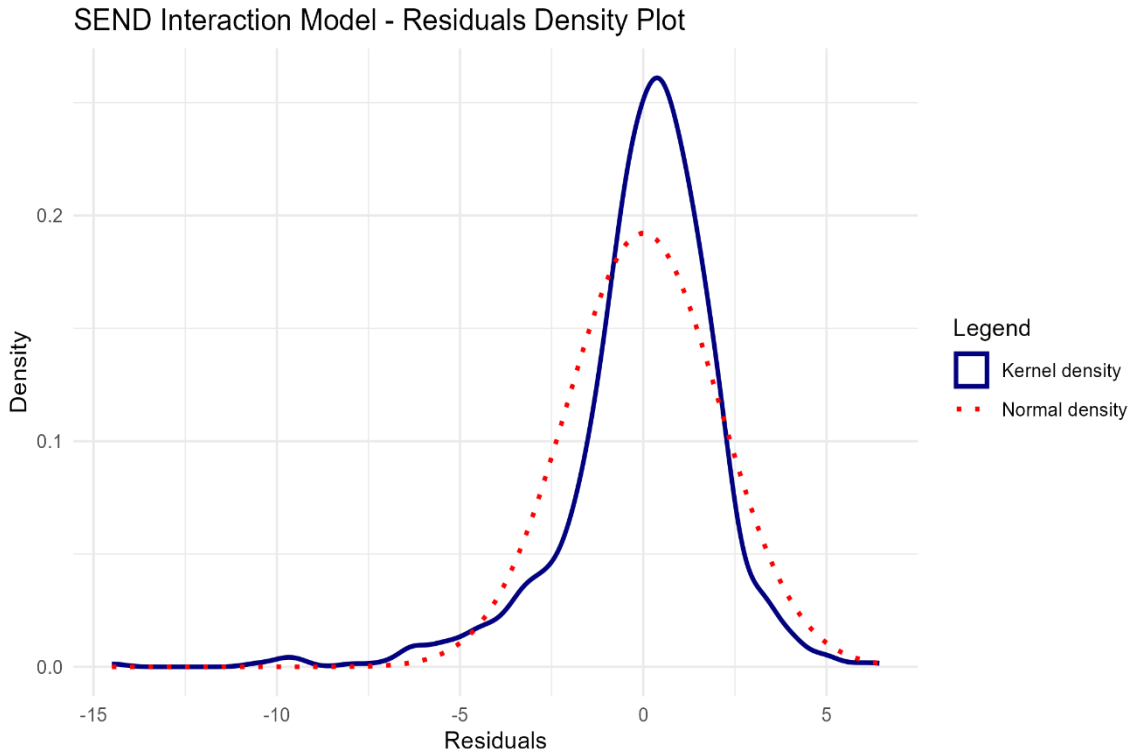


Figure D15: Residual kernel density plot (SEND interaction model)

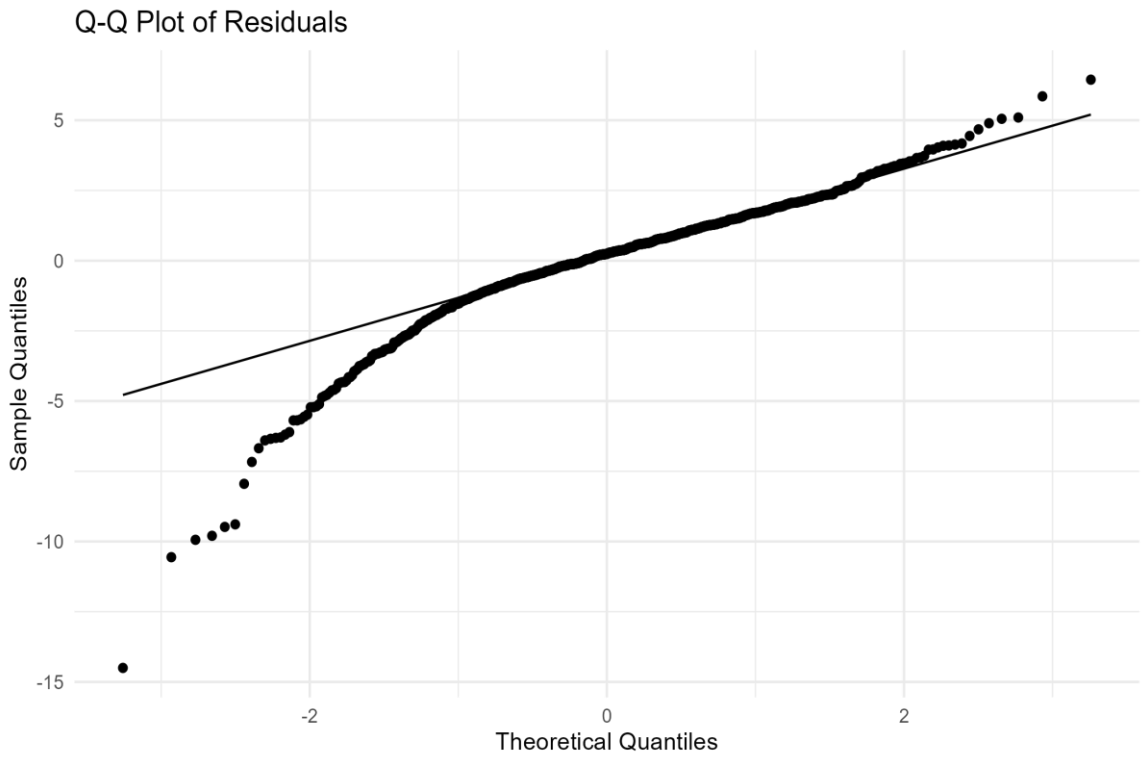


Figure D16: Residual Q-Q plot (SEND interaction model)

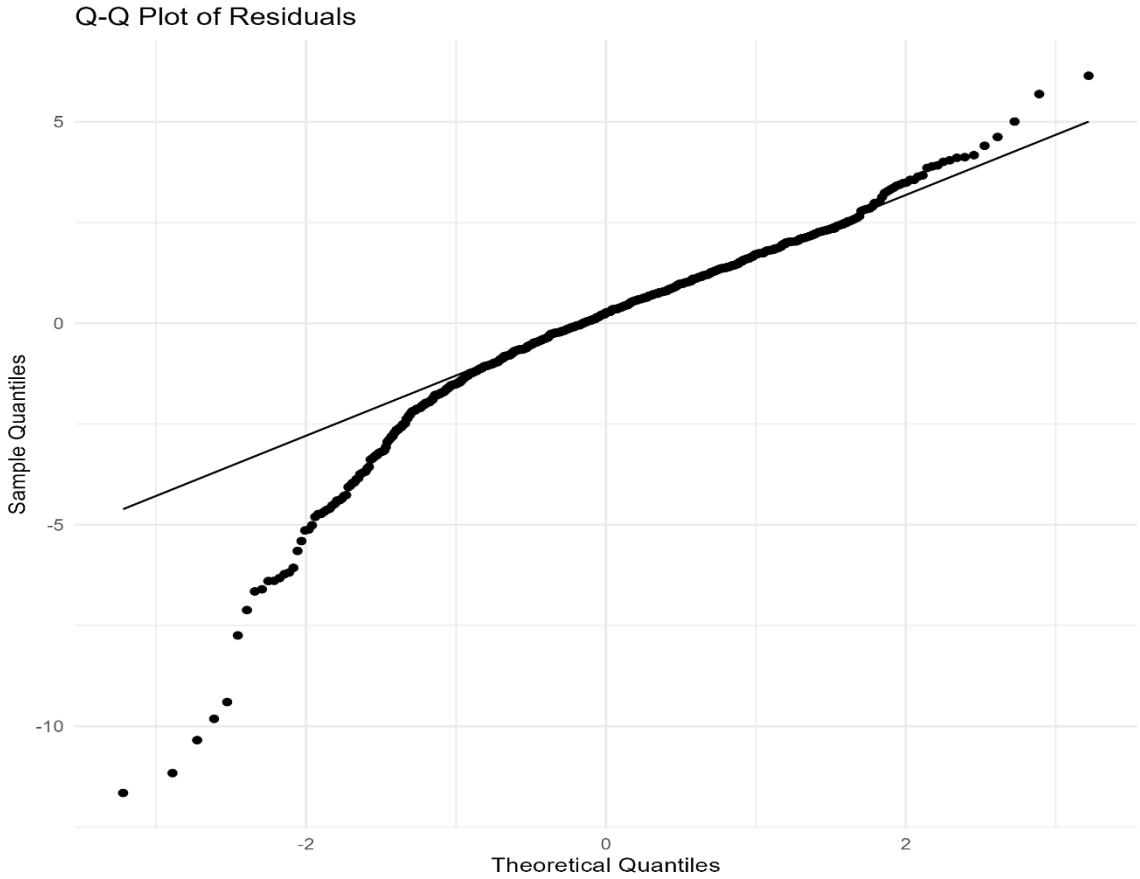


Figure D17: CACE second stage model residuals Q-Q plot

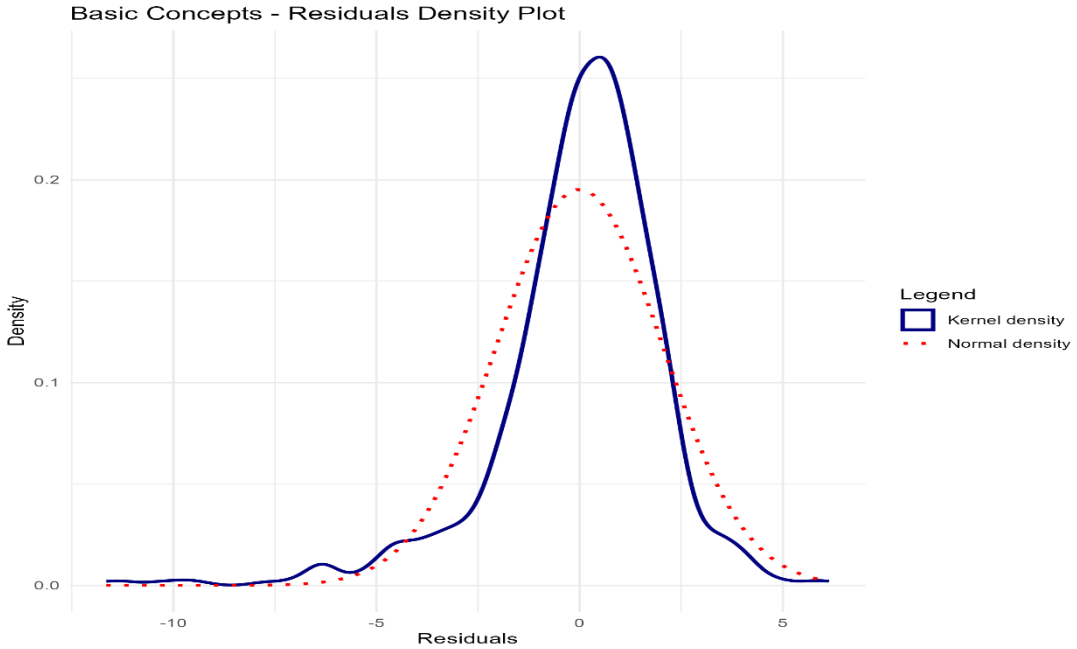


Figure D18: CACE second stage model residuals kernel density plot

Appendix E: Missing data logistic regression output

Table E1: Endline missingness – logistic regression model 1

	Estimate.	Standard error	P-value
Intercept	-1.11	0.34	0.00
Treatment	-0.08	0.20	0.69
CELF Preschool-2 UK 'Basic Concepts' baseline scores	-0.04	0.02	0.06
Region: Everton	-0.10	0.28	0.72
Region: Hertford North	-0.08	0.30	0.78
Region: Hertford South	0.00	0.29	0.99
PVI	-0.41	0.23	0.08

Pseudo R^2 (fixed and random effects): 0.05

Table E2: Endline missingness – logistic regression model 2

	Estimate.	Standard error	P-value
Intercept	-1.24	0.36	0.00
Treatment	-0.02	0.20	0.93
CELF Preschool-2 UK 'Basic Concepts' baseline scores	-0.03	0.02	0.15
Region: Everton	-0.01	0.28	0.72
Region: Hertford North	-0.15	0.30	0.60
Region: Hertford South	0.04	0.30	0.90
PVI	-0.55	0.25	0.02
EAL	0.38	0.24	0.10
EYPP	0.38	0.24	0.11
SEND	-0.33	0.41	0.42

Pseudo R^2 (fixed and random effects): 0.05

Table E3: Endline missingness – logistic regression model 3

	Estimate	Standard error	P-value
Intercept	-1.09	0.38	0.00
Treatment	-0.02	0.20	0.93
CELF Preschool-2 UK 'Basic Concepts' baseline scores	-0.04	0.02	0.12
Region: Everton	-0.10	0.28	0.72
Region: Hertford North	-0.16	0.30	0.59
Region: Hertford South	0.01	0.30	0.97
PVI	-0.56	0.25	0.02
EAL	0.38	0.24	0.11
EYPP	0.36	0.24	0.13
SEND	-0.29	0.41	0.48
Gender (male)	-0.23	0.19	0.23

Pseudo R^2 (fixed and random effects): 0.06

Appendix F: Subgroup interaction model results

Table F1: EYPP interaction models raw output

Outcome	Variable	Raw coefficient	Standard error	95% CI	P-value
CELF Preschool-2 UK 'Basic Concepts' subtest	Treatment	0.50	0.20	0.13–0.88	0.009
	EYPP status	-0.46	0.27	-0.99–0.09	0.10
	Treatment EYPP interaction	0.16	0.38	-0.60–0.89	0.71
EY Toolbox ENA	Treatment	2.40	0.84	0.84–3.99	0.003
	EYPP status	-1.81	1.32	-4.47–0.79	0.17
	Treatment EYPP interaction	-3.09	1.83	-6.79–0.70	0.11

Table F2: EYPP interaction models effect size

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	460 (57)	15.57 (15.32–15.82)	431 (81)	14.92 (14.65–15.2)	891 (460; 431)	0.23 (-0.17–0.62)	0.71
EY Toolbox ENA	444 (83)	33.81 (32.59–35.03)	405 (108)	30.88 (29.61–32.15)	849 (444; 405)	-0.05 (-0.46–0.36)	0.11

Table F3: EAL interaction models raw output

Outcome	Variable	Raw coefficient	Standard error	95% CI	P-value
CELF Preschool-2 UK 'Basic Concepts' subtest	Treatment	0.57	0.19	0.16–0.97	0.006
	EAL	-0.44	0.27	-0.97–0.07	0.093
	Treatment EAL interaction	-0.28	0.36	-1.00–0.43	0.434
EY Toolbox ENA	Treatment	2.07	0.89	0.35–3.70	0.018
	EAL	1.56	1.28	-0.95–4.09	0.222
	Treatment EAL interaction	-1.67	1.72	-4.99–1.67	0.33

Table F4: EAL interaction models effect size

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (Boot. 95% CI)	Boot. p-value
CELF Preschool-2 UK 'Basic Concepts' subtest	460 (67)	15.57 (15.32–15.82)	442 (71)	14.94 (14.67–15.21)	902 (460; 442)	0.10 (-0.29–0.49)	0.43
EY Toolbox ENA	444 (83)	33.81 (32.59–35.03)	416 (97)	30.92 (29.67–32.17)	860 (444; 416)	-0.03 (-0.36–0.26)	0.33

Further appendices

Please find the further appendices as a separate document on the project page.

Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk
@EducEndowFoundn
Facebook.com/EducEndowFoundn