



Education
Endowment
Foundation

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

Evaluation report

July 2026

Palak Roy, Lillian Flemons, Gemma Schwendel, Tara Paxman,
and Stephen Welbourne



NFER

National Foundation for
Educational Research



The Education Endowment Foundation is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, colleges, and early years settings 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 from the government, 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



Table of contents

About the evaluator 4

Executive summary 5

Introduction 7

Methods 24

Impact evaluation results 55

Implementation and process evaluation results 76

Cost.....110

Conclusion113

References119

Appendix A: EEF cost rating.....124

Appendix B: Security classification of trial findings125

Appendix C: Changes since the efficacy trial127

Appendix D: Effect size estimation128

Further appendices:129

About the evaluator

The project was independently evaluated by a team from the National Foundation for Educational Research (NFER). The trial director and principal investigator was Dr Stephen Welbourne, Research Director—quantitative evaluation, who served as Project Director with responsibility for overall design, delivery, and quality assurance of the evaluation. The evaluation team and impact evaluation were led by Palak Roy, Senior Trials Manager, who oversaw day-to-day trial delivery in line with the protocol and coordinated all strands of the evaluation. The Implementation and Process Evaluation (IPE) was led by Lillian Flemons, Research Manager, who was responsible for the design and delivery of the IPE. Gemma Schwendel served as the trial statistician, leading the quantitative analysis, with additional analytical support provided by Jose Liht. Kathryn Hurd, Research Operations Lead, oversaw all data collection and the communications strategy, supported by Jo Stringer, Senior Project and Delivery Manager, who coordinated day-to-day operations for the main trial and acted as the primary point of contact for settings, and Katherine Stoodley, Project and Delivery Manager, who led data collection and communications for the formative evaluation. Operational delivery for the main trial was further supported by Elaine McCann, Naila Hassan, Rob Green, Robert Weeks, and Andrei Illes. The RAPT marking process was overseen by David Simcock and Tara Paxman, who piloted RAPT assessments, trained markers and team leaders, developed marking materials, and ensured quality assurance and consistency. Tara Paxman also contributed as an IPE researcher, supporting qualitative data collection, analysis, and reporting. Additional contributions to the RAPT assessment development were made by Louise Bailey. Vrinder Atwal provided project coordination and administrative support throughout the evaluation.

Contact details:

Dr Stephen Welbourne
National Foundation for Educational Research (NFER)
The Mere,
Upton Park,
Slough
Berkshire, SL1 2DQ
Email: s.welbourne@nfer.ac.uk
Phone: +44(0)1753 574123

Acknowledgements

We are grateful to all the nursery settings that took part in this evaluation. In particular, we appreciate the time and insight offered by early years practitioners who contributed through interviews, focus groups, and surveys.

We would also like to acknowledge the delivery team at OxEd¹ for their continued collaboration throughout the project. Our thanks also extend to colleagues at the EEF, including Maria Rodrigues and Toby Whittaker, for their consistent guidance and support.

¹ At the start of the evaluation, the organisation was operating under its previous name, OxEd and Assessment, before later becoming known as OxEd. Throughout the document, the organisation is therefore referred to as OxEd for consistency. Note that OxEd is a spin-out of the University of Oxford.

Executive summary

The project

The NELI Preschool programme aims to improve the oral language skills of all 3–4-year-old children in nursery settings through a structured combination of whole-class language enrichment sessions and additional targeted support for those with the weakest language skills. The intended programme comprises 20 weeks of delivery, including daily whole-group sessions consisting of shared book reading, dialogical questioning, and activities to support the learning of related vocabulary (15–20 minutes), and additional targeted support for children with low language ability, in the form of three small-group sessions per week (10–15 minutes each), and one individual session per week (5–10 minutes). Whole-group sessions are delivered by the room leader or the class teacher, while small-group and individual sessions are delivered by nursery practitioners or teaching assistants, all of whom complete the programme’s online training. The programme was developed by OxEEd, drawing on prior research on early oral language development and structured shared-reading pedagogies.

The effectiveness trial involved 3,298 children across 303 nursery settings in England. The study was a two-arm cluster randomised controlled trial, meaning individual settings were randomly allocated to deliver the NELI Preschool programme or be in a control group to continue their usual practice. It was designed as an effectiveness trial to assess programme performance under real-world conditions. The primary outcome was a combined measure of oral language skills, derived from LanguageScreen and RAPT subtests. The implementation and process evaluation used a mixed methods approach, including surveys, case studies, session observations, and analysis of session delivery logs (SDL), to examine how the intervention was delivered in practice as well as to understand usual practice. The 20-week intervention was delivered over the 2024–2025 academic year, with evaluation activities also completed within that year.

As part of the Department for Education’s Early Years Recovery Programme, the Education Endowment Foundation (EEF) worked with Stronger Practice Hubs to fund Early Years settings’ access to evidence-informed programmes and study the programme’s influence on practice and children’s outcomes. This initiative aimed to support education recovery following the pandemic, whilst also developing our understanding of effective professional development in the early years. The EEF worked with the following Early Years Stronger Practice Hubs on the evaluation of NELI Preschool: East of England, South West, Northern Lights, Yorkshire and Humber Together, Kent, Derbyshire and Nottinghamshire, Thames Valley, London South, and Great North.

Table 1: Key conclusions

Key conclusions

1. Children in NELI Preschool settings made an average of two additional months’ progress in their oral language skills compared to children in control settings. This result has a high security rating.
2. Children eligible for Early Years Pupil Premium (EYPP) also made an average of two additional months’ progress in their oral language skills compared to children in control settings. These results may have lower security than the overall findings because of the smaller number of children.
3. Children receiving additional targeted support made around two additional months’ progress. This was similar to the progress of children receiving only whole-class enrichment, suggesting that targeted sessions help children with the weakest language skills benefit fully from whole-class provision.
4. Most practitioners reported that the NELI Preschool training was useful and prepared them well to deliver the programme effectively and perceived the programme to have a strong positive impact on children’s language development—including vocabulary development, confidence in speaking, and engagement.
5. Practitioners in NELI Preschool settings demonstrated improvements in their confidence, practice, and motivation to support early language development compared to practitioners in control settings.

EEF security rating

These findings have a high security rating. This trial was an effectiveness trial, which tested whether the intervention worked under everyday conditions in a large number of settings. The trial was a well-designed two-armed randomised controlled trial. The trial was well-powered; however, 15% of the children who started the trial were not included in the final analysis

because they did not complete endline testing. Nonetheless, the children included in the final analysis were similar in terms of prior attainment and other characteristics.

Additional findings

Children in NELI Preschool settings made, on average, two additional months' progress in oral language compared to children in control settings. This is our best estimate of impact, which has a high security rating. As with any study, there is always some uncertainty around the result: the possible impact of this programme ranges from no additional progress up to two months' additional progress. Results were strongest for expressive language outcomes.

The evaluation found similar positive impacts for all subgroups of children: eligible for EYPP, targeted for additional support, and with English as an Additional Language. While the trial was designed to detect an effect on children eligible for EYPP, the findings for this group are less secure than the overall findings due to the smaller sample size achieved. The findings for the targeted group suggest that the targeted sessions helped children with the weakest language skills benefit fully from whole class provision. However, this was not directly tested so we cannot be certain about the relative contributions of the targeted versus whole class elements for these children. Exploratory analysis suggests the intervention was effective in both PVI and maintained settings. In settings that followed the programme carefully, meeting all prespecified compliance criteria, the estimated impact was larger than the headline result, suggesting children benefit more where the programme is delivered as intended.

Practitioners reported that quiet spaces for small-group sessions, sufficient staffing, and child attendance at the setting were important for enabling successful programme delivery. The online training was perceived to be very useful, and practitioners perceived the programme to have had a positive impact on children's vocabulary, confidence in speaking, and engagement. Practitioners also reported increases in their own confidence, motivation, and practice for supporting early language development. However, this increase in practitioner confidence was found to explain only a small part of the overall child-outcome gains. Exposure to the sessions appears to constitute the largest driver of child outcomes, where the hypothesised mechanisms of increased opportunity to speak and language scaffolding from practitioners were observed to be borne out in practice. The strongest positive impact being found for expressive language is consistent with an approach that prioritises structured vocabulary teaching and opportunities for interaction.

Cost

The average cost of NELI Preschool for one setting was £2,682.80 over three years, which equates to around £40.65 per child per year, assuming a typical cohort of 22 children. This is considered a low cost according to the EEF cost rating, but costs vary depending on whether a setting already owns the required equipment, the number of practitioners needing training, and whether staff complete training within paid time. Start-up costs account for most of the variation, particularly where settings need to purchase a tablet, screen or monitor, or TV trolley—equipment that over half of PVI settings (Private, Voluntary, and Independent) in this trial did not have. Ongoing costs relate mainly to digital-subscription fees and staff time. Settings spent on average 112 hours on training and preparation, in addition to 55 hours on session delivery in the first year, with similar delivery-time requirements expected in subsequent years. Settings that required more staff to be trained or that had higher staff turnover may also face recurring training costs.

Impact

Table 2: Summary of impact on primary outcome(s)

Outcome/ Group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No of pupils	P-value	EEF cost rating
Oral language skills, all children	0.10 (0.02, 0.18)	2 months		3,298	0.01	£ £ £ £ £
Oral language skills, EYPP children	0.11 (-0.01, 0.24)	2 months	N/A	901	0.07	N/A

Introduction

Background

Language is the cornerstone of education and a gateway to literacy, science, and mathematics, each predominantly conveyed through language. Neuroscientific research indicates that the neural mechanisms that underpin reading largely overlap with those that support language (Welbourne *et al.*, 2011; Woollams, Halai, and Lambon Ralph, 2018). Phonological representations that support spoken language comprehension are also foundational to decoding written words, explaining why early phonological awareness robustly predicts later literacy success (Liberman, Shankweiler, and Liberman, 1989; Snowling *et al.*, 1994).

A considerable body of research supports the potential of early language interventions to shift developmental trajectories—particularly for vocabulary, narrative, and oral comprehension (Scarborough, 2009; Marulis and Neuman, 2010; Fricke *et al.*, 2012; Law *et al.*, 2017). A recent meta-analysis of early years language and communication programmes further highlights substantial positive impacts for young children (EEF, 2023a). However, in the UK, the evaluation of such programmes remains limited by methodological robustness, highlighting a pressing need for rigorous evaluations to bolster the evidence base. Hence, there exists a compelling argument for the development and evaluation of interventions that foster a rich language environment, through teaching speaking and listening skills, to ensure that all children have the opportunities they need to develop strong language skills, especially children from disadvantaged backgrounds.

Development and evaluation of interventions are also integral to the Department for Education's (DfE) Stronger Practice Hubs (SPH) policy, which aims to foster evidence-informed practice in early years education. Launched in November 2022 as part of the DfE's Early Years Education COVID-19 Recovery Package, SPHs play a crucial role in addressing the pandemic's impact on young children. They achieve this by aiding early years settings in establishing local networks and disseminating evidence-informed practices, ultimately enhancing the quality of education and care. The Education Endowment Foundation (EEF) actively supports the SPH initiative by enhancing the evidence base for early years approaches. This involves identifying evidence-informed programmes, funding evaluation-related activities, and overseeing project delivery and evaluation, ensuring ongoing progress monitoring.

The Nuffield Early Language Intervention (NELI) Preschool (formerly known as NELI Nursery or NELI-N) is an oral language enrichment programme for all children in the year before they begin school. It builds on two previous language intervention programmes, combining the focus on children's vocabulary, narrative, and active listening development in NELI and the shared reading component of PACT (Parents and Children Together). NELI is a targeted early language intervention programme for children with weak oral language skills, delivered by teaching assistants (TAs) in Reception, and has been widely evaluated. The EEF has funded two independent randomised controlled trials for NELI. The efficacy trial tested two versions of the programme, a 30-week and a 20-week intervention, both of which showed positive intervention effects on pupils' language skills. Three hundred and fifty children in 34 schools took part in the efficacy trial which showed positive impact (Sibieta, Kotecha, and Skipp, 2016; Fricke *et al.*, 2017). In the 30-week intervention, where nursery children received the programme in the final term of nursery and continued in Reception year, children made an additional four months' progress, compared to the 20-week intervention, delivered in Reception only where children made an additional two months' progress. The programme was also evaluated as an effectiveness trial with 1,071 children from 192 schools (Dimova *et al.*, 2020; West *et al.*, 2021). This trial tested the 20-week intervention in Reception, and it showed an additional three months' progress in language skills. It was subsequently scaled up nationally with funding from the DfE. The EEF funded an independent impact evaluation of the scale-up which was evaluated using a quasi-experimental Fuzzy Regression Discontinuity (FRD) design (Smith *et al.*, 2023). There were 10,759 children from 356 schools taking part in this study, which found that children who received the NELI programme made the equivalent of four additional months' progress in language skills, on average, compared to children who did not receive NELI. PACT is an early language programme for parents to deliver to children at home. The programme includes parent-child interactions that involve elements of shared book reading, explicit vocabulary teaching, and retelling aspects of the story. The efficacy trial for PACT was funded by the Nuffield Foundation and involved 208 preschool children (aged 3 years) and their parents in 22 children's centres. The trial showed improvements in children's language and narrative skills at immediate post-test (Burgoyne *et al.*, 2018).

The NELI Preschool programme was developed by Charles Hulme and his research group at the University of Oxford and is now delivered by OxEd, a University spinout company founded to translate the research into practical application. It is a 20-week oral language enrichment programme for 3- to 4-year-old children in the year before they enter formal education. The programme aims to support language development in early years settings via enrichment and targeted components. It is designed around the principles of shared book reading and guided play. The programme comprises a blend of language screening for children, online training and delivery support for practitioners, a scripted programme for in-person delivery in preschools, and supportive materials such as storybooks and digital resources. Enrichment sessions focus on whole-class activities, such as reading books, engaging in dialogical questioning, and activities to support the learning of related vocabulary, while the targeted component involves small-group and one to one sessions tailored to children at the bottom 20–25% of the class² in oral language skills, incorporating activities to support and consolidate learning, enhance narrative skills and scaffold language production.

The government's current early years strategy prioritises speech, language, and communication as a foundation for later attainment and life chances. *Giving Every Child the Best Start in Life* sets an ambition to raise outcomes at age five by providing high-quality early education and childcare, with communication and language highlighted as a core lever for improvement (DfE, 2025d). This is reinforced in the Early Years Foundation Stage statutory framework—rich, back-and-forth interactions between children and adults underpin cognitive development and shape vocabulary growth, calling for daily language-rich environments with high-quality conversations, sustained shared thinking, and frequent shared reading. Through practices such as dialogic reading, commenting on children's interests, modelling new vocabulary, storytelling, and role-play, practitioners help children embed and extend language in meaningful contexts (DfE, 2025b). Moreover, DfE's practitioner guidance (DfE, 2025e) highlights that high-quality, back-and-forth exchanges—where adults comment on children's interests, echo back and add new vocabulary, join play, adapt language using verbal and non-verbal cues, and allow turn-taking and wait time—are central to developing communication within a language-rich environment.

Since the pandemic, policy has also moved decisively towards earlier identification and targeted intervention. The Early Language Support for Every Child (ELSEC) programme funds specialist teams to spot and respond to speech and language needs earlier across early years settings and schools, embedding small-group support, and improving access for children with special educational needs (SEN) (Public Health England, 2020b, 2020c). Complementary national guidance from Public Health England (now OHID) provides the Early Language Identification Measure and Intervention (ELIM-I) and a whole-system speech, language, and communication (SLC) approach for commissioners and providers—again underscoring early screening, tailored intervention, and equity (Public Health England, 2020a).

The NELI Preschool programme aligns tightly with this policy landscape by providing (i) universal, whole-group sessions; and (ii) additional targeted small-group and one to one support for the lowest-attaining children. The programme implements these aspects through daily dialogic story sessions and guided play that script modelling of target words, prompt responsive adult-child talk and structured turn-taking, and scaffold practitioner behaviours so that children can extend vocabulary and take part in conversations.

Between 2020 and 2022, NELI Preschool was developed and then evaluated in an efficacy trial, funded by the Nuffield Foundation and conducted by Hulme's research team at the University of Oxford (West *et al.*, 2023). Sixty-five settings (64 maintained nurseries and one private setting) from seven geographical areas in England took part in the trial. All 3- to 4-year-old children attending preschool for four or more days (or half days) a week were considered eligible for the trial. In total, 1,586 preschool children across 70 preschool classrooms took part. LanguageScreen, which is an app-based assessment to measure children's oral language skills, was used as a screening test. The settings were randomised to intervention or control groups, stratified by size (dichotomised by mean number of children attending), which resulted in 33 settings in the intervention group and 32 in the waitlist control group. Following screening, the six children in each class received whole-

² This is equivalent to selecting six children from a class of 25–30 children. In the evaluations, this has been implemented by selecting six children with the lowest oral language skills in a class or per setting—see impact evaluation design for further details.

class enrichment and additional targeted support, plus four children randomly selected from the remainder of each class ($n = 726$) received individually administered in-depth testing, in order to be able to evaluate the enrichment-only element of NELI Preschool, as well as the programme received by children identified with language weaknesses receiving additional targeted support.

In-depth testing at baseline and endline included subtests from the Child Evaluation of Language Fundamentals (CELF) Preschool II UK (Semel, Wiig, and Secord, 2006) and the Renfrew Action Picture Test (RAPT) (Renfrew, 2019), and also involved assessing vocabulary knowledge taught in NELI Preschool intervention, narrative skills measured via the Renfrew Bus Story, self-regulation assessed using the HTKS-R assessment (Heads-Toes-Knees-Shoulders Revised) (Gonzales *et al.*, 2021), and children's behavioural adjustment to school from the Brief Early Skills and Support Index (BESSI) (Hughes *et al.*, 2015). The preregistered primary outcome was a latent language variable defined by loadings from eight subtests from LanguageScreen, CELF, and RAPT.

The evaluation found that the children receiving the enrichment (whole-class) component of the intervention made more progress in their language skills than control group children, with a Cohen's $d = 0.26$ (0.10, 0.43), equivalent to three months' additional progress in language development. Further analysis was undertaken for the children identified as requiring the additional targeted support (those with the lowest language scores at baseline), which showed a significant effect of the intervention on this subgroup with a Cohen's $d = 0.16$ (0.04, 0.30), equivalent to 2 months' additional progress in language development. The report highlighted this difference in the two effect sizes and suggested that the lower effect size for the targeted group may reflect both the greater severity of children's language difficulties, which additional support could not fully overcome, and the possibility that some children were not yet developmentally ready to benefit fully from targeted intervention at this stage. The report further recommended that the next evaluation should test two versions of the programme, one with and one without the targeted support element.

This evaluation was an effectiveness trial of the NELI Preschool programme, funded by the EEF and the DfE's Stronger Practice Hubs, which aimed to test whether NELI Preschool improved children's oral language skills when delivered at scale across both maintained nursery settings and PVI settings. SPHs who co-funded the trial were the East of England Early Years Stronger Practice Hub, the Early Years South West Stronger Practice Hub, the Northern Lights Early Years Stronger Practice Hub, the Yorkshire and Humber Together Early Years Stronger Practice Hub, the Kent Early Years Stronger Practice Hub, the Derbyshire and Nottinghamshire Early Years Stronger Practice Hub, the Thames Valley Early Years Stronger Practice Hub, the London South Early Years Stronger Practice Hub, and The Great North Early Years Stronger Practice Hub. This trial was designed to build on the earlier efficacy trial (West *et al.*, 2023), extending it to a larger and more diverse sample and ensuring that at least one-third of participating settings were PVI, given that the efficacy trial had included very few PVI providers. Within this broader effectiveness trial, the EEF also commissioned NFER to undertake a formative evaluation to address a key gap from the previous evaluation—namely, the limited evidence on feasibility and acceptability of programme delivery in PVI settings—and to pilot the trial's proposed data collection tools to examine feasibility in PVI settings ahead of full rollout.

NFER delivered the formative evaluation between January and July 2024. OxEd recruited 20 PVI settings (around 516 children) via regional events and expressions of interest, with participating settings varying in region, size, deprivation context, and provision type; NFER also selected ten of these settings for case studies to explore implementation in greater depth. The formative evaluation found that the programme was feasible and acceptable for PVI settings, with practitioners completing training as expected and delivering sessions with good fidelity, while valuing the programme's resources and perceived benefits for children and staff. However, low engagement with the community of practice, the intensity of the delivery schedule, and the exclusion of lower-attendance children from targeted sessions emerged as key challenges requiring adjustment. A top-level recommendation was to increase flexibility in delivery and refine implementation tools to better support equitable access and reduce burden. Overall, the trial instruments and processes functioned reasonably well in practice, though some limitations were identified. The Targeted Group Selection spreadsheet did not fully capture the rationale for children's inclusion or exclusion, and session delivery log completion—while generally acceptable—placed an additional burden on a small number of settings. In response, the tools and guidance were lightly refined, which improved

clarity for settings, eased communication, and contributed to better engagement and response rates in the effectiveness trial, without requiring substantive changes to the tools. The full report is provided in a separate Appendix.

As a brief overview, this effectiveness trial employed a setting-randomised design. We randomised 303 early years settings on a one to one basis into two groups, intervention and control, stratified by setting type (maintained nurseries or PVI settings) and setting size (settings with thirty or fewer 3 to 4 year olds vs those with more than thirty 3 to 4 year olds) to ensure equal group allocation in each strata. Intervention settings delivered the NELI Preschool programme that included both enrichment and targeted components. This was different from the efficacy trial recommendation for three reasons: (i) the difference in the effect sizes between the programme impact on children exposed solely to the enrichment component and those also receiving the additional targeted component may simply have resulted from sampling differences as the confidence intervals around these estimates overlapped considerably; (ii) we considered that offering a version without the targeted component, which aimed to make it more accessible to the lowest ability children, would raise ethical concerns; and (iii) it would have required adopting a three-armed trial which would have required substantially greater sample sizes to maintain statistical power for EYPP-eligible children, and the associated cost was judged to be prohibitive. Given these considerations, the full version of the programme was implemented in this evaluation including both enrichment and targeted components. These components were also evaluated separately, similar to the efficacy trial, as part of a secondary analysis.

All 3- to 4-year-old children (as of August 31, 2024) in the preschool classroom were eligible to take part in the trial if they were able to access LanguageScreen at baseline (see Participant selection section). These children were assessed by settings using LanguageScreen prior to randomisation which acted as a baseline measure for the primary outcome and also acted as a screening test to select children who would be targeted for additional support if the settings were randomised to receive intervention.

The primary outcome measure for this trial was a latent oral language variable generated using endline LanguageScreen and RAPT.

The secondary outcomes for the trial were the total raw score for each subtest of LanguageScreen³ and for each subtest of the RAPT at endline. There were several additional impact analyses to ascertain the effectiveness of NELI Preschool at improving the language skills of different subgroups of children. These included children in the targeted component, children who only received the whole-class (enrichment) component, disadvantaged children (as identified by Early Years Pupil Premium or EYPP), children in the PVI settings only, and children with English as an Additional Language (EAL).

The Implementation and Process Evaluation (IPE) for this trial covered several research questions that explored the extent to which NELI Preschool was successfully scaled up to a greater number and variety of settings by examining fidelity of core component implementation and potential adaptations made by settings. In particular, the effectiveness trial focused on PVI settings because these settings differ from maintained settings in their structures, staffing models, and attendance patterns, raising important questions about the feasibility and acceptability of implementing NELI Preschool in contexts that typically have more variable hours, mixed-age groups, and fewer specialist resources. The IPE also investigated how well the programme reached and supported disadvantaged children, evaluated the targeted intervention's perceived effectiveness, explored contextual and moderating factors influencing implementation, and assessed the perceived impact on settings, practitioners, and children—including disadvantaged children specifically.

There will also be a longitudinal evaluation which will test the assumption within the programme's Theory of Change that early gains—in both children's receptive and expressive language skills and practitioners' confidence and practice—

³ While the raw scores were used for all other LanguageScreen and RAPT subtests, the standard score was used for the LanguageScreen Listening Comprehension subtest, because the raw score for that test has a discontinuity depending on whether the child is younger than 4.5 years or not. Further details are given in the Outcome measures section.

translate into sustained improvements over time. To do this, the evaluation will link evaluation data with the National Pupil Database to obtain the Early Years Foundation Stage Profile (EYFSP) which will be the longitudinal outcome measure. This will be complemented by a light-touch follow-up survey with NELI Preschool Leads to explore the extent to which programme approaches have been embedded in continuous provision, any longer-term impacts observed, and any challenges encountered in sustaining programme delivery where desired. Findings from both strands of the longitudinal evaluation will be integrated into an addendum report.

Key differences between the efficacy and the effectiveness trials are presented in Appendix C.

Intervention

The NELI Preschool programme was developed by the OxEd team while at the University of Oxford. The programme was delivered for the purpose of the effectiveness trial between January and June 2025.

Rationale, theory, and/or goal of essential elements of the intervention

NELI Preschool is a scripted programme that aims to improve the receptive and expressive oral language skills of young children through screening and intervention delivered by trained early years practitioners. The programme was developed in response to evidence that language skills are a key determining factor for later attainment in both literacy (Hjetland *et al.*, 2020; Hulme *et al.*, 2015) and numeracy (Chow and Ekholm, 2019; Hornburg, Schmitt, and Purpura, 2018). Moreover, children from disadvantaged backgrounds are more likely to start school with lower levels of language abilities (Guo and Harris, 2000; Hart and Risley, 1995; Sampson, Sharkey, and Raudenbush, 2008; Hutchinson, Reader, and Akhal, 2020; Sirin, 2005). This disadvantage gap will, on average, subsequently widen over the course of their educational career (Hutchinson, Reader, and Akhal, 2020).

There is likewise evidence to indicate that early language development is influenced by the quality and quantity of spoken language children are exposed to through interactions with their caregivers (Hoff, 2003; Murray and Egan, 2014; Weisleder and Fernald, 2013; Huttenlocher *et al.*, 1991), and may mitigate the effects of disadvantage (Roulstone *et al.*, 2011; Goodman and Gregg, 2010). Dialogic exchanges in the nursery setting may consequently provide children with opportunities for language development that they may not receive in their home environment, including in cases where English is an additional language for their family. As a result, the programme seeks to provide enrichment opportunities for all children aged 3–4 years in the setting, as well as targeted intervention for children with lower abilities in an effort to close the language gap that has already emerged and would influence the child's life opportunities (Field, 2010).

Finally, the NELI Preschool programme aims to provide early years practitioners with high-quality continuous professional development (CPD) to provide more children with better support for their language development. High-quality CPD has also been found to increase satisfaction and retention in the education workforce (Fletcher-Wood and Zuccollo, 2020), which may help to address the significant recruitment and retention challenges being faced by the sector (Haux *et al.*, 2022). The developers identified the key outcomes for practitioners as increased pedagogical knowledge, confidence, and motivation around supporting early language development. Self-determined types of motivation have been found to be associated with higher job satisfaction and lower levels of both job burnout and turnover intentions in the workplace (Fernet *et al.*, 2008). Confidence has also been argued to be an 'essential element' of teacher professional performance (Nolan and Molla, 2017).

Recipients of the intervention

Recipients of the intervention are 3 to 4 year olds in maintained and PVI early years settings.

All children participate in the whole-group sessions. This includes younger children if they are present in the classroom and wish to join in.

In addition, at least six children from each setting are selected to attend additional small-group and one to one targeted sessions. These children are intended to be those with the lowest language abilities (as assessed using LanguageScreen—see below), excluding any children who are not developmentally ready to engage in small-group and/or one to one activities.⁴ In addition, OxEEd advises that only children attending the setting for 15 hours or more per week be selected for the targeted component. For the purpose of the trial, settings selected these children as per the Targeted Group Selection Guidance provided by the delivery team.⁵

Small-group sessions are conducted with children selected for the targeted component in groups of three. Each child selected for the targeted component also participates in a weekly one to one session with a practitioner.

Physical or informational materials used in the intervention

Training

All participating settings are given access to an online training platform, as well as an online Delivery Support Hub. The online training course is made up of nine modules. The first half of the course looks at why language skills are important, developing understanding of receptive and expressive language skills, how to recognise difficulties with these, and also how to support their development. The second part of the training looks at the practical considerations and preparation necessary for effectively delivering the intervention. Each section of the training also has a quiz to enable participants to monitor their progress, and there is a short assessment at the end of each of the two parts of the training course. The course is primarily text-based, with a few videos. Some of the videos are short explanations of a particular aspect or concept; others provide examples of full programme sessions being delivered in accordance with best practice. Activity prompts regularly encourage participants to post on the platform, usually to share prior experiences and reflections around the content covered.

The Delivery Support Hub follows a similar format to the online training course but is intended for use during the delivery period. It provides additional guidance around timetabling small-group sessions and supporting new learners of English and/or children with SEN, as well as additional tips and a discussion forum for each week of delivery. The Hub provides links to an online noticeboard where settings can post examples of the ways in which they have integrated the texts and vocabulary into their continuous provision and get inspiration from others.

Both the training platform and the Delivery Support Hub provide participating practitioners with access to the delivery team, experienced practitioners, and speech and language specialists who can provide support and answer any questions.

Upon completing the training, participants receive a NELI Preschool practitioner certificate, accredited by the CPD Certification Service (CPDUK).

Intervention

Settings are provided with access to the online LanguageScreen assessment for completing the child screening assessments required for selecting the group for the targeted component of the programme. The setting requires a tablet to carry out the screening process, as (at the time of the trial) LanguageScreen could only run on a tablet via an app. For the purpose of the trial, EEF-funded tablets were provided by OxEEd to settings that did not already have one available.

⁴ This may be as a result of identified or unidentified Special Educational Needs (SEN).

⁵ Implementation models are described in detail in the Trial design section.

Settings are provided with material for 20 storybooks, one for each week of the intervention. The material includes both a physical version of each book and digital slides that display the books with related questions and whole-group activities. Settings also receive a handbook with the session scripts and prompts, and physical activity materials such as vocabulary flashcards. Targeted sessions use weekly session cards, story sequencing cards, and press-out characters to facilitate small-world play. In addition, physical materials are provided to support good behaviour during the sessions, including a plush elephant puppet (NELI puppet), best listener stickers, and a listening rules poster. Additional digital materials include song files and other materials for continuous provision opportunities. Finally, practitioners are provided with guidance on selecting children for the targeted component (Targeted Group Selection Guidance), as well as the Narrative Progress Tracker template (see below for more detail) with a list of Receptive and Expressive Language Skills Checkpoints for ages 18–24, 24–36, 36–48, and 48–60 months for use with the children receiving the targeted intervention.

Settings are provided with a parent ‘home learning link’ template that can be sent out to parents for each week of the intervention to inform them of what their child is doing and to promote parental engagement in the language development process.

Settings require a large digital display such as a TV, interactive whiteboard, or projector to deliver the digital material for the whole-group sessions. For the purpose of the trial, the EEF funded OxEEd to provide settings with a screen, a wireless keyboard, and a trolley where required.

Procedures, activities, and/or processes used in the intervention

Training

All practitioners delivering the intervention must have completed the NELI Preschool online training course, which is expected to take 10–12 hours to complete. For a practitioner to be recorded as having completed the training, they need to have marked at least 90% of the course steps as complete and scored at least 70% in the two knowledge tests. For the purpose of this trial, at least two practitioners per setting were required to complete the online training course, although settings were encouraged to support as many practitioners as possible to do so.⁶ The training includes practical exercises, and participants are encouraged to participate in the forums in both the training course and the Delivery Support Hub to create online communities of practice. Settings would typically have three to four months to complete the training at their own pace before the start of delivery, but this time frame was reduced to four to five weeks for this trial due to timeline constraints.

Participants are able to revisit the training material and access the Delivery Support Hub at any time. Additional staff members can complete the training partway through the delivery period if there is any turnover of previously trained staff—although they must complete the training before they deliver any sessions.

The developer team also delivers four webinars over the course of the delivery period to provide settings with additional information about the programme and address any common queries or concerns that emerge, as well as suitable adaptations that settings could choose to make if faced with particular challenges. Practitioners are encouraged to attend as many as possible. The webinars are remote and last between 45 minutes and an hour, taking place at a range of times during working hours.

The webinars cover:

- An introduction to the programme, LanguageScreen, and selecting children for Targeted Intervention;
- Expanding learning from the programme beyond the sessions;

⁶ The EEF funded OxEEd to provide £120 per practitioner who completed training for up to five practitioners per setting.

- Tips for supporting children receiving the Targeted Intervention; and
- Dialogical reading, storytelling, and making the most of the NELI Preschool books.

The OxEd team also provide ongoing one to one support to settings throughout the set-up and delivery period, including help with scheduling, overcoming challenges, and advising on suitable adaptations.

Intervention

The intervention sessions focus on reading and dialogical questioning to support understanding of each selected book. Each book introduces four or five 'special words', which constitute the focus of a range of vocabulary learning activities over the course of the week.

The intervention involves both whole-group and small-group, and individual sessions for the six children selected for the additional targeted component. The setting must complete the child screening process with the LanguageScreen assessment to identify the targeted group before starting delivery. LanguageScreen is an online assessment that adapts based on children's ability and age, moving them on after several errors and presenting slightly different items depending on their age. The automated discontinuation rules mean that if a child gets multiple items wrong on a particular activity, the child is simply moved on to the next activity. The Listening Comprehension subtest differs by age group, as younger and older children receive different stories and question sets (see the Outcomes section for further details about the assessment). LanguageScreen must be administered one to one by a practitioner. The practitioner does *not* have to complete the training before starting the screening process. LanguageScreen is also intended to inform practitioners of each child's ability level so that they can better tailor their support to individual needs.

In the whole-group sessions, each week begins with a dialogical reading session. The practitioner displays the digital version of the book on the screen, reads it to the children and asks them questions about it. Subsequent sessions during each week focus on 'special words' from the books and engage children in activities to support vocabulary learning, including using flashcards and a wide range of digital resources. The sessions include regular consolidation of previously learned words. While a script for each session is provided, the training encourages practitioners to develop their own questions and prompts in accordance with the programme guidance around effective dialogic reading. There is, however, a prescribed structure for each session.

Targeted sessions focus on supporting children to develop their narrative skills by retelling aspects of the books engaged with in the whole-group sessions. In the first small-group session each week, each child takes on a role, and the practitioner guides the children to retell the story through role-play, following the workbook outline and using provided pictures, toys, and character cards. The individual sessions with the targeted children follow the same structure, but the child is asked to retell the story themselves. The remainder of the small-group sessions focus on vocabulary, with the practitioner using flashcards to revise the meaning of each of that week's 'special words' followed by a related activity.

At three points over the 20-week programme, practitioners use the Narrative Progress Tracker to assess how each of the children in the targeted group is progressing in relation to appropriate language development targets. In Week 1, Week 10, and Week 20, the practitioner transcribes how each child in the targeted group retells that week's story in the individual session. They record the number of words and make any notes they deem relevant. The practitioner is then asked to highlight and date evidence of the child having achieved any of the listed Receptive and Expressive Language Skills Checkpoints, either through the transcribed story or in continuous provision. The practitioner then discusses with their Special Educational Needs Co-ordinator (SENCo), class teacher, or room leader their reflections on the progress observed and how best to proceed with the tailored support.

Intervention providers/implementers

Training

The online training course was developed by the research team at the University of Oxford (now at OxEd). The training course forums and Delivery Support Hub are moderated by the OxEd team and speech and language specialists hired by OxEd. Supporting webinars are hosted by the OxEd team alongside experts and special guests to provide additional guidance and support.

Intervention

It is recommended that the Nursery class teacher (in maintained settings) or room leader (in PVI) delivers the whole-group sessions, and that an early years practitioner, teaching assistant, or nursery nurse delivers the targeted sessions. This recommended approach stems from the aim of fostering teamwork and a sharing of the additional workload across different staff in the setting. However, it is up to the individual setting to decide how the session delivery will be staffed, with the only condition being that any practitioner delivering sessions must have completed the full online training course.

For the purpose of the trial, settings were asked to nominate a NELI Preschool Lead responsible for overseeing programme implementation at a setting level and liaising with the evaluation team.

Mode of delivery

Training

All training is provided online. The training course and Delivery Support Hub are in a fully asynchronous and interactive format (see Strategies to maximise effective implementation), while the webinars are delivered live (although recordings are also provided).

Intervention

The intervention is delivered in person. Whole-group sessions are delivered to all children aged 3–4 years in the setting and any other children sharing the same space. The targeted sessions involve small-group sessions that are delivered to a maximum of three children at a time, as well as one to one individual sessions.

Location of the intervention

The intervention is delivered in maintained and PVI early years settings. It is recommended that the targeted sessions are delivered in a quieter area of the classroom, such as a hallway, book corner, or coat area.

Duration and dosage of the intervention

The intervention lasts for 20 weeks. Five whole-group sessions are delivered each week. Targeted children should attend three small-group sessions and one individual session each week. This means six small-group and individual sessions are delivered each week in total. It is recommended that whole-group sessions are delivered daily, that small-group sessions take place following the first, third, and fifth whole-group sessions, and the individual sessions following the fourth whole-group session. It is at the discretion of each setting to determine their delivery schedule.

Whole-group sessions are intended to last 15–20 minutes, small-group sessions 10–15 minutes, and individual sessions 5–10 minutes. Practitioners are encouraged to focus on the quality of interaction that takes place during the session, rather than imposing a specific duration. For example, practitioners are allowed and encouraged to strategically expand on child-led interactions relevant to the story or subject at hand.

Adaptation of the intervention

Settings can tailor delivery schedules and staffing to their own needs. In addition, sessions are scripted with flexibility for practitioners to adapt them to the cohort's ability level. In whole-group sessions, 'step-up' words and questions are provided as opportunities to further extend the vocabulary of more advanced children. 'Step-down' words and questions are likewise provided to offer children with less advanced language abilities additional support and opportunities to engage. In addition, a large number of dialogical questions are supplied for each book, with the expectation that the practitioners choose their own selection. The training includes specific additional guidance around how to support children with EAL.

Practitioners are expected to follow the overall structure of the programme, to regularly complete the activities accompanying each session and to use dialogical reading strategies throughout.

Strategies to maximise effective implementation

The online and asynchronous nature of the training course means it can be completed by practitioners as and when it is convenient for them prior to starting delivery, and as many practitioners as possible are encouraged to engage with it. Practitioners can also return to the material for a refresher at any point as required. The official CPD accreditation for completing the course is intended to further motivate practitioners to engage.

The online Delivery Support Hub provides tailored expert support to the individual needs of each practitioner throughout delivery. Participating settings also receive regular newsletters to introduce them to the next stage of the programme and remind them that they can receive support at any time via email and the Delivery Support Hub.

The programme is highly structured and scripted but with flexibility and adaptation encouraged in practice. This provides practitioners with a clear reference point when needed while allowing greater freedom where they feel confident in applying the programme principles.

The digital intervention materials are designed to be high quality and visually engaging for children. Behaviour management and listening promotion tools, such as the NELI puppet, are also embedded in the programme to encourage child engagement.

Evidence of implementation variability

The process evaluation carried out as part of the efficacy trial, which focused on maintained settings, reported that the programme had been generally well adhered to (West *et al.*, 2023). The formative evaluation (see the Further appendices), which focused on disadvantaged PVI settings, found the programme to be broadly feasible and acceptable, with strong training completion rates and high levels of fidelity in session delivery. However, it also identified challenges, including the intensity of the delivery schedule, low engagement with the community of practice, and limited access for children with fewer registered hours at the setting. These findings informed formative recommendations for greater flexibility in delivery and refinements to implementation tools. This effectiveness trial subsequently aimed to understand how well the programme could be implemented across a wider range of setting types with differing structures, resources, and staffing.

Theory of Change

The Theory of Change (ToC) for the NELI Preschool programme is shown in Figure 1. It outlines the target population of the intervention and the activities, outputs, short-term and intermediate outcomes that are ultimately intended to lead to the outcome of improving children's receptive and expressive language skills, thus providing them with a better foundation for learning in the long term. NELI Preschool involves a training programme for early years practitioners in addition to the actual sessions that are delivered to the children, so an additional row of outputs and outcomes has been included to map how practitioners' knowledge and teaching skills may develop in parallel with the children's own language skills.

The programme consists of a whole-class enrichment component and a targeted intervention. These are separated out in different coloured boxes to represent the different pathways children in the targeted and enrichment sessions are anticipated to follow. The purpose of the targeted intervention is to support children with the lowest language abilities to access and benefit from the enrichment sessions alongside their peers. This mechanism relies on the following key assumptions:

- The children that settings are instructed to select for the targeted intervention, according to guidance provided by OxEd, are those who will benefit most from this component—including in terms of being able to access the level of focus and behaviour required for the intervention.
- Practitioners are able to deliver the number of sessions required and with the right children in the context of what is often quite an unpredictable nursery environment.
- Practitioners are sufficiently skilled to adapt their teaching approaches to the needs of different levels of language abilities.
- The online training course is adequate to support effective delivery of the programme, as this depends substantially on skilled teaching, not just adherence to a script.
- All children are given the opportunity to actively participate in the sessions, not just those children who are already quite vocal.
- The skills children gain through the sessions are transferable to their broader interactions with adults, other children, and the world around them.

While a script is provided for each session, the NELI Preschool training encourages practitioners to be proactive in tailoring discussion to their specific context and participating children. It also relies on practitioners employing a range of teaching techniques, such as scaffolding and modelling, beyond the script provided.

Finally, the ToC posits that children must be actively engaged for the duration of the sessions, including by speaking when given the opportunity to do so; it is not sufficient that they are simply present.

The NELI Preschool efficacy trial (West *et al.*, 2023) provides evidence to suggest that the above holds true for the trial sample, which were overwhelmingly maintained settings. This effectiveness trial further explores these elements for both maintained and PVI settings through the surveys, interviews, and observations carried out for the IPE (see below for more detail).

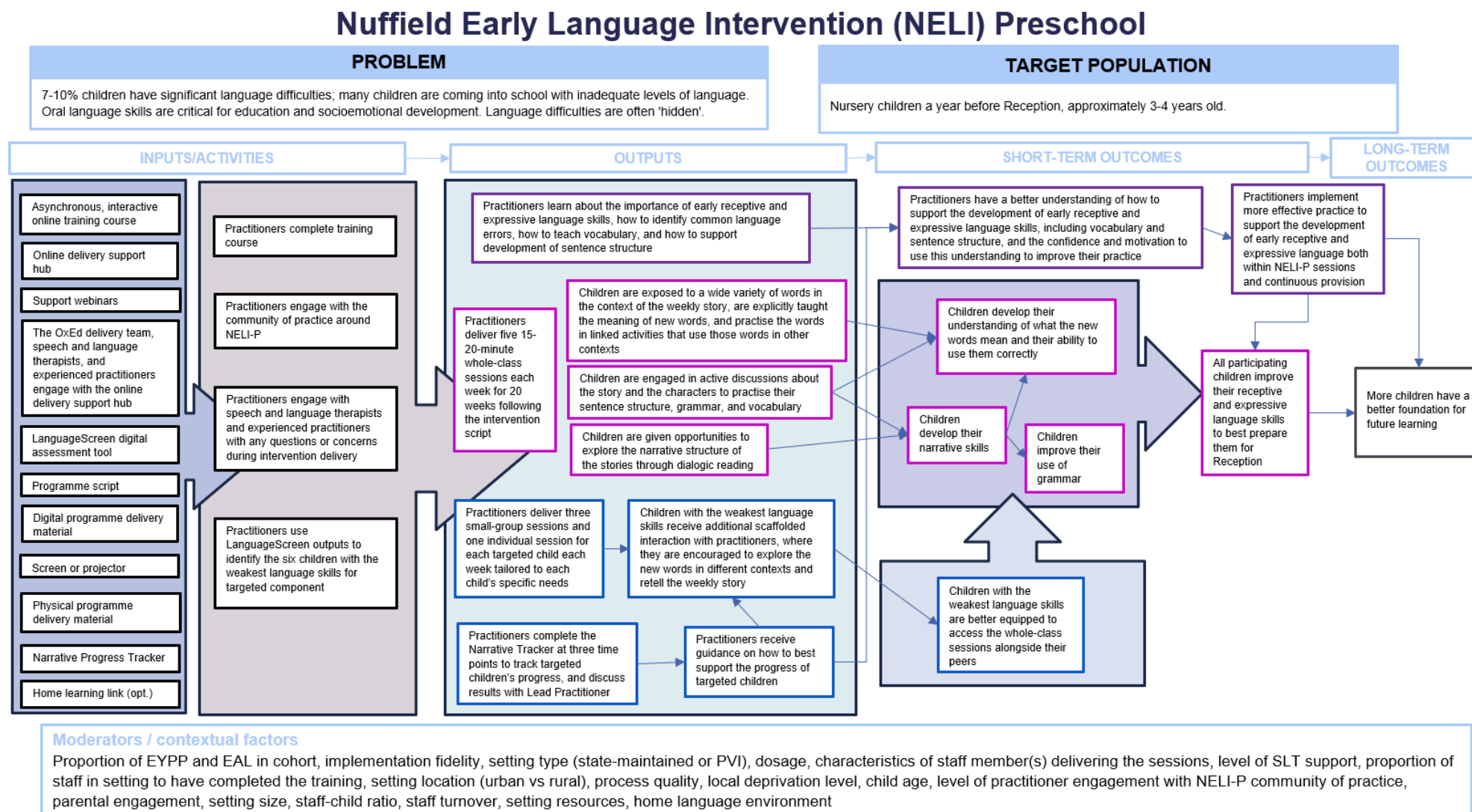
Practitioner outcomes in relation to increased understanding of supporting early language development rely on training and support provided as part of the programme to be sufficiently comprehensive and of high quality to engage the practitioners and enable this learning to occur. In addition, the potential for increased practitioner motivation and confidence rests primarily on the assumption that increased understanding breeds confidence and motivation—although positive feedback through interactions via the Delivery Support Hub and webinars may also have a direct influence on these factors.

Longer-term outcomes for both practitioners and children were not considered as part of the efficacy trial. Therefore, this effectiveness trial is crucial for determining whether the following assumptions are valid—particularly through its longitudinal component⁷:

- Increased practitioner understanding, confidence, and motivation in relation to supporting early language development leads to improved practitioner practice.
- Improvements in children's receptive and expressive language skills supports children to better engage with teaching and learning opportunities on a long-term basis, resulting in better academic outcomes.

⁷ See the Evaluation objectives section for further details.

Figure 1: NELI Preschool Theory of Change



Evaluation objectives

The aim of this effectiveness trial was to build on the findings from the efficacy trial by testing whether NELI Preschool is effective at improving children's oral language skills when delivered on a larger scale across both maintained nursery settings and PVI settings. A full evaluation protocol and statistical analysis plan (SAP) can be found on the EEF's website.⁸

The primary research question for this impact evaluation was:

RQ1: How effective is NELI Preschool at improving oral language skills of 3- to 4-year-old children in intervention settings compared to children in control group settings?

Secondary research questions for this trial were to ascertain the effectiveness of NELI Preschool on the six components of the primary assessments: LanguageScreen subtests of Expressive Vocabulary, Listening Comprehension, Receptive Vocabulary and Sentence Repetition, and RAPT subtests of Information and Grammar.

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

Additional research questions for this trial sought to ascertain the effectiveness of NELI Preschool for different subgroups of children.

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

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

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

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

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

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

The accompanying IPE investigated the following research questions:

IPE RQ1. To what extent did NELI Preschool successfully scale up to a greater (i) number and (ii) variety (i.e., both maintained and PVI) of settings?

IPE RQ2. How well did the NELI Preschool intervention reach and support disadvantaged children (those eligible for EYPP and/or those with EAL)?

IPE RQ3. To what extent can the targeted component of the intervention be seen to be fulfilling its intended purpose?

⁸The evaluation protocol can be found at https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/neli_preschool_-_evaluation_protocol_-_v.2.0.0.pdf?v=1771945634 (Roy et al., 2025b).

The SAP can be found at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/neli_preschool_-_statistical_analysis_plan_-_v.1.0.0.pdf?v=1771945634 (Roy et al., 2025a).

IPE RQ4. What are the key moderators and contextual factors that influenced how effectively the intervention was implemented in each setting?

IPE RQ5. What was the perceived impact of the intervention for participating (i) settings, (ii) practitioners, and (iii) children?

IPE RQ6. What was ‘business as usual’?

In addition to the main trial, we will carry out a longitudinal component involving administrative data analysis coupled with a light-touch IPE to assess whether any impacts of NELI Preschool are sustained over time and to examine whether its CPD element supports improvements in language development practice across the setting beyond session delivery.

Ethics and trial registration

The trial was conducted in line with NFER’s Code of Practice⁹ which sets out guiding principles relating to ethics, data protection and information security, and caring for research participants. The ethical approval for this evaluation was approved by NFER’s Code of Practice during the set-up of this trial in October 2023. All of NFER’s projects abide by its Code of Practice, which is in line with the Codes of Practice from BERA (the British Educational Research Association), MRS (the Market Research Society), the SRA (the Social Research Association), and the UK Statistics Authority, among others.

Ethical agreement for participation in the trial was provided by the Headteacher or the Nursery Manager of the setting via signing the Memorandum of Understanding (MoU) that outlines the responsibilities of all parties involved in the trial.

Before requesting settings to share their children’s data, settings were asked by NFER to distribute a parent information sheet and withdrawal form to the parents/carers of all participating children. Parents/carers who did not want their child to participate in the evaluation could inform their child’s setting using the withdrawal form. Settings were asked not to share data for the children whose parents/carers withdrew them at this stage. Parents/carers also had the opportunity to withdraw their child from the evaluation and associated data processing at any stage of the trial, using the withdrawal form.

In addition, NFER offered settings the opportunity to opt out of sharing any additional data after the trial, and we communicated this option when we contacted them at the end of the trial to complete the feedback survey. This data sharing pertains to the evaluation data collected by NFER for this trial which will be shared with OxEd following the main trial report (see the Data protection section for further details).

The trial was registered on ISRCTN registry (<https://www.isrctn.com/ISRCTN17801846>). The trial was designed, conducted, and reported to CONSORT standards (<https://www.consort-spirit.org/published-statements>).

Data protection

NFER was the data controller for this evaluation. OxEd was the data processor during the evaluation. After the evaluation report is published on the EEF website, NFER will share evaluation data with OxEd, at which point OxEd will become an independent data controller for this data. The evaluation data will also be pseudonymised and transferred to the EEF data archive, at which point the EEF will become the data controller for the archived data. More information on this data sharing and archiving is included below.

The lawful basis for processing personal data is covered by General Data Protection Regulation (GDPR) Article 6(1)(f): Legitimate interests: the processing is necessary for your (or a third party’s) legitimate interests unless there is a good reason to protect the individual’s personal data which overrides those legitimate interests. A legitimate interest assessment was undertaken which demonstrated that the trial fulfilled one of NFER’s core business purposes (undertaking research, evaluation, and information activities). It has broader societal benefits and will contribute to improving the lives of learners

⁹ https://www.nfer.ac.uk/media/cgpl42av/nfer_code_of_practice.pdf

by providing evidence about the impact of teaching techniques used in the classroom. Research cannot be done without processing personal data, but processing does not override the data subject's interests. Therefore, it was in our legitimate interest to process and analyse personal data for the administration of this trial.

NFER and OxEd signed a data processing and data sharing agreement that governed the collection and sharing of personal data for this trial. This agreement included a description of the nature of the data being collected and how it will be shared, stored, protected, and reported by each party. In addition, OxEd provided an MoU to settings (Appendix E), explaining the nature of the data being requested of settings, practitioners and children, how it will be collected, and how it will be passed to and shared with NFER. Two separate Privacy Notices were also made available: one for settings and another one for parents (Appendix F).

For the purposes of the trial, OxEd collected expressions of interest (EoI) (Appendix G) from interested settings, which also included a link to the setting information sheet (Appendix H), gathering nursery staff contact details. Upon eligibility confirmation, Headteachers/Nursery Managers signed an MoU and identified the NELI Preschool Lead in their setting. Personal data collected were: names, contact details, and job roles for Headteachers/Nursery Managers and NELI Preschool Leads. Practitioners' personal data for the trial were also collected directly from nurseries by OxEd via the MoU. OxEd shared the EoI and MoU data with NFER via a secure online portal for trial coordination. Contact details shared on the EoI or MoU were also shared with the EEF and SPHs which enabled the EEF to monitor recruitment progress and to confirm setting eligibility. Some SPHs contacted settings to inform them about support available through the Hubs.

The NELI Preschool Lead in each setting was asked to complete a survey at the beginning and at the end of the trial period. In addition, nursery practitioners were asked to complete an online survey about their confidence and motivation in supporting early language development (at the beginning and end of the trial).

As part of the evaluation, OxEd shared practitioner training attendance data with NFER. Practitioners' training attendance data at a nursery level were linked to the child-level data collected as part of this trial to understand how important practitioner training attendance is for child outcomes. Additionally, specific variables from the practitioner surveys covering self-reported confidence and motivation in supporting early language development were also linked to the child-level dataset and included in the impact analysis.

NFER collected children's personal data directly from the settings. This included full name, date of birth, class (if applicable), gender, EAL status (as perceived by nursery staff), EYPP status, and attendance patterns at the nursery. NFER shared children's personal data with OxEd via OxEd's LanguageScreen software or via NFER's secure online portal to enable LanguageScreen administration for each child. LanguageScreen assessments were administered by practitioners (at baseline) and by NFER's Test Administrators (in person at endline). NFER's Test Administrators also administered RAPT paper assessments (in person at endline). OxEd and NFER accessed children's LanguageScreen scores via OxEd's LanguageScreen software.

For intervention settings, NFER also collected whether children were selected for the NELI targeted intervention and their attendance at NELI Preschool sessions via a session delivery log, completed by nursery practitioners and shared with NFER using NFER's secure online portal.

A small number of settings were invited to facilitate a visit from an NFER researcher involving practitioner interviews and observations of NELI Preschool sessions and usual practice. Children were observed as part of this, but no further personal data were collected. Some of the training webinars in which practitioners actively participated were also observed by an NFER researcher.

Longitudinal evaluation

As part of the longitudinal evaluation, NFER will ask the NELI Preschool Lead from intervention settings to complete an online survey to give their views about the longer-term impact of the programme. This survey will be administered one year after the trial ends (May–July 2026).

NFER will share children’s personal data with the DfE in order for their data to be matched to the National Pupil Database (NPD). NFER will collect children’s Early Years Foundation Stage Profile (EYFSP) from the NPD and use it for the longitudinal analysis.

Additional data sharing between NFER and OxEd

After the main EEF report is published (currently planned for July 2026), NFER will share the evaluation data with OxEd, which will enable OxEd to link it with the data they hold and undertake their own research to publish findings from the trial in scientific journals. At this point, OxEd will become an independent data controller. The evaluation data will include:

- Children’s personal data (except the data collected from the NPD).
- Close-ended survey responses from the practitioner surveys covering self-reported confidence and motivation in supporting early language development. This will include nursery ID, and practitioner role and qualification level.
- Anonymised data from close-ended responses from the NELI Preschool Lead surveys. This will not include any personal, nursery, or contextual data that could make matching or identification possible.

Data archiving and deletion

Three months after the publication of the final evaluation report (currently planned for June 2027), pseudonymised child-level data will be added to the EEF data archive, which will include practitioners’ training attendance at a nursery level and specific variables from the practitioner surveys covering self-reported confidence and motivation in supporting early language development. At this point, the EEF becomes the data controller and is responsible for making decisions about the means and purposes of processing. The EEF archive is managed by FFT Education on behalf of the EEF and hosted by the Office for National Statistics (ONS). Further information is available in the privacy notice for the EEF data archive at <https://educationendowmentfoundation.org.uk/privacy-notice/privacy-notice-for-the-eef-data-archive>. NFER will delete any recording and digitalised transcriptions of interviews as soon as the notes are finalised and personal data collected as part of this evaluation within one year of publication of the final report (currently planned for June 2027). This deletion process is currently planned for June 2028.

Staff/teacher interview data and notes from the observations will not be shared with the EEF or archived. Open-response survey questions and data from the NELI Preschool Lead surveys will likewise not be shared or archived.

Project team

Table 3: Key members of the evaluation team (NFER and RAND) and the delivery team at OxEd

Name	Role and responsibilities
Organisation: NFER	
Stephen Welbourne	• Project Director—responsible for overall delivery and quality assurance of the evaluation
Palak Roy	• Trial Manager and impact evaluation lead—responsible for the day-to-day management of the trial, delivery of the trial as per trial protocol, and co-ordinating all strands of the evaluation
Lillian Flemons	• IPE Lead—responsible for the design and delivery of the IPE
Gemma Schwendel	• Trial Statistician—responsible for leading quantitative analysis for the main trial
Kathryn Hurd	• Research Operations Lead—responsible for overall data collection and setting communications strategy
Jo Stringer	• Senior Project and Delivery Manager—responsible for day-to-day operations for the main trial, including co-ordinating data collection and serving as point of contact for the settings
Katherine Stoodley	• Project and Delivery Manager—responsible for data collection and setting communications for the Formative Evaluation
Elaine McCann	• Operational Delivery Associate—supporting operations for the main trial including co-ordinating data collection and supporting setting communication
Naila Hassan	• Survey Support Officer—checking online templates once returned from participants, validating submitted data, following up on issues ensuring complete and accurate returns

Rob Green	• Data Manager—responsible for creating online survey and assessment instruments, and designing data collection templates as well as cleaning and validating all datasets ready for analysis
Andrei Illes	• Data Manager—responsible for creating online survey and assessment instruments, and designing data collection templates as well as cleaning and validating all datasets ready for analysis
David Simcock	• RAPT Marking Lead—piloting RAPT assessments that fed into RAPT assessment and marking material; providing oversight for RAPT marking process including training RAPT markers and team leaders, development of marking materials, and ensuring quality assurance and consistency throughout the marking process
Tara Paxman	• RAPT Marking Lead—piloting RAPT assessments that fed into RAPT assessment and marking material; providing oversight for RAPT marking process including training RAPT markers and team leaders, development of marking materials, and ensuring quality assurance and consistency throughout the marking process • IPE researcher—supporting the qualitative data collection, analysis, and reporting for the IPE
Louise Bailey	• Research Manager—piloting RAPT assessments that fed into RAPT assessment and marking material
Vrinder Atwal	• Project Co-ordinator—providing administrative assistance throughout the evaluation
Robert Weeks	• Project and Delivery Manager—managed operational delivery of the marking process, including liaison with the markers, team leaders, and the project team, and monitoring progress to ensure timely completion
Organisation: RAND	
Elena Rosa Speciani	• Consultant on the evaluation side, offering advice on IPE design, trial design, and implementation in early year settings, and contributing towards the interpretation of trial findings
Dr Marilyn Groom	• Consultant on the evaluation side, contributing to the IPE design, providing ongoing advice in working with early years settings and contributing towards the interpretation of trial findings
Organisation: OxEd	
Mariela Rios Diaz	• Project Manager of NELI Preschool Trial—responsible for the day-to-day management of the trial
Francina Clayton	• Project Manager of NELI Preschool Trial—responsible for the day-to-day management of the trial
George Ulmann	• Managing Director of OxEd—responsible for overseeing trial management
Dr Gillian West	• Director of OxEd—responsible for overseeing research aspects for the NELI Preschool trial
Prof Charles Hulme	• CEO and Founder of OxEd—responsible for overseeing research aspects of the NELI Preschool trial
Blansh Al-Awkabi	• Junior Research Officer—responsible for liaising with settings, recruitment, setting engagement, and delivery support
Hazel Carter	• Marketing Director—responsible for marketing, recruitment, and setting engagement
Joe Lowe	• Operations Director—responsible for logistical, strategic and operational planning and execution, and process improvement
Kevin Price	• Support Team Operator—responsible for tech and training support
Ina Skronska	• Support Team Operator—responsible for tech and training support
Kate Humphreys	• Research Officer—responsible for mentoring on the training
Harley Richardson	• Product Director of Technology and Technological Aspects—responsible for overseeing technology use and development, including LanguageScreen and OxEd portal
Mihaela Duta	• Data and Software Engineer—responsible for data architecture design, pipeline development, and data modelling and optimisation
Chris Speight	• Data Protection Officer—responsible for monitoring compliance and advising on data protection obligations
Jenny Mason	• Programme Delivery and Product Advisor—responsible for mentoring on the training
Sarah Hearne	• Project/Research Manager—responsible for ad hoc project and research management
Sue Lowe	• Communications Manager—responsible for developing communication strategies and co-ordinating communications for recruitment

Methods

Trial design

Table 4: Trial design

Trial design, including number of arms		Two-arm, cluster randomised controlled trial
Unit of randomisation		Setting
Stratification variable (s) (if applicable)		Setting type (maintained and PVI) and setting size (settings with thirty or fewer 3 to 4 year olds vs those with more than thirty 3 to 4 year olds)
Primary outcome	Variable	Oral language skills
	Measure (instrument, scale, source)	Latent oral language variable (range -6.20–4.70) formed from expressive and receptive subtests in LanguageScreen and the RAPT
Secondary outcome(s)	Variable(s)	1. Vocabulary knowledge 2. Literal and inferential language comprehension and expressive language skills 3. Vocabulary understanding 4. Language comprehension and production 5. Information 6. Grammar
	Measure(s) (instrument, scale, source)	1. LanguageScreen Expressive Vocabulary subtest (score range 0–24) 2. LanguageScreen Listening Comprehension subtest (score range 65–135) 3. LanguageScreen Receptive Vocabulary subtest (score range 0–23) 4. LanguageScreen Sentence Repetition subtest (score range 0–14) 5. RAPT information subtest (score range 0–41) 6. RAPT grammar subtest (score range 0–39)
Baseline for primary outcome	Variable	Oral language skills
	Measure (instrument, scale, source)	Latent oral language variable (range -3.56–4.63) formed from expressive and receptive subtests in LanguageScreen
Baseline for secondary outcome(s)	Variable	1. Vocabulary knowledge 2. Literal and inferential language comprehension and expressive language skills 3. Vocabulary understanding 4. Language comprehension and production 5. Oral language skills 6. Oral language skills
	Measure(s) (instrument, scale, source)	1. LanguageScreen Expressive Vocabulary subtest (score range 0–24) 2. LanguageScreen Listening Comprehension subtest (score range 65–135) 3. LanguageScreen Receptive Vocabulary subtest (score range 0–23) 4. LanguageScreen Sentence Repetition subtest (score range 0–14) 5. Latent oral language variable formed from expressive and receptive subtests in LanguageScreen 6. Latent oral language variable formed from expressive and receptive subtests in LanguageScreen

The trial was planned and executed as a two-arm, cluster randomised effectiveness trial with randomisation at setting level. A total of 303 nursery settings across 9 Stronger Practice Hubs were randomly allocated to either intervention (to receive the NELI Preschool programme) or control (to continue usual practice). Randomisation was stratified by setting type (maintained vs PVI) and setting size (settings with thirty or fewer 3 to 4 year olds vs those with more than thirty 3 to 4 year olds) to ensure balance between two setting types and two setting sizes. All settings had a 50:50 chance of being assigned to the intervention group within each strata group (see the Randomisation section for further details).

Practitioners from both groups of settings administered baseline LanguageScreen assessments with all eligible children (see Participant selection section for further details) prior to randomisation. All settings also selected their preferred NELI Preschool implementation model prior to randomisation, with support from the delivery team, and shared it with NFER. The delivery team developed three implementation models for the targeted component of NELI Preschool, specifically for the trial, to accommodate variation in class structures and capacity and to support settings in making informed decisions prior to randomisation. These models differed in the number of targeted intervention groups and the number of children per group.

Table 5: NELI Preschool implementation models

Model A	One targeted intervention group with six children (setting $n = 244$) This is the most common format, where settings are single form entry and have only one preschool class. The setting will run one whole-group session and will select six children for targeted intervention.
Model B	Multiple targeted intervention groups with six children per group (setting $n = 38$) This model works best in larger, multiple form entry settings where NELI Preschool can be run separately for each class. Whole-group sessions will be run for each class, and six children per class will be selected for targeted intervention. In cases where there are not enough children attending 15 hours or more per week to make up a complete targeted intervention group in each class, six children will be selected for targeted intervention across the setting.
Model C	One targeted intervention group with six or more children but split across multiple classes (setting $n = 21$) This model works best for settings with multiple classes that run more than one whole-group session (one for each class) but find the targeted intervention in Model B too intensive and unrealistic. Instead, settings ideally select three children from each class to be included in targeted intervention irrespective of the number of children per class who attend 15 hours or more per week ¹⁰ .

Settings also provided additional information for each child prior to randomisation so that NFER could select the children who would ideally be selected for the targeted component in each setting, in line with the Targeted Group Selection Guidance. In usual practice, however, practitioners themselves select the children who will receive the targeted component. This parallel selection was carried out for trial comparison purposes so that equivalent children could be identified in control settings and to compare the intervention effect on two subgroups: those selected by the practitioners and those selected by NFER (based on the Targeted Group Selection Guidance).

Settings were then randomly allocated to two groups: 151 to the intervention group and 152 to the control group. After randomisation, NFER shared baseline LanguageScreen assessment results with intervention settings, as LanguageScreen is an integral part of the programme. The data included raw and standard scores for each component as well as the overall standard score and percentile. Intervention settings also received OxEd's Targeted Group Selection Guidance consistent with their chosen NELI Preschool implementation model (selected prior to randomisation). As outlined above and in keeping with usual practice, practitioners in intervention settings used this data to select children for the targeted group intervention. After randomisation, practitioners from intervention settings also received NELI Preschool training and delivered the NELI Preschool programme to children for 20 weeks.

¹⁰ For further details on child eligibility criteria, see the Participant selection section.

Control settings continued their usual practice and did not have access to LanguageScreen assessment results or NELI Preschool training or materials during the trial.

Intervention settings received the following incentives:

- £100 after completion of all baseline data collection;
- £150 upon submission of NELI Preschool session delivery logs (SDLs);
- £150 following endline assessments;
- Paid staff time for NELI Preschool training (£120 per trainee, for up to five trainees per setting); and
- Paid staff time for administering LanguageScreen at baseline.

Additionally, 12 case study intervention settings received an extra £100 for participating in NELI Preschool session observations. Two practitioners from each of these settings who took part in interviews received £50 each.

Control settings received:

- £100 after completion of all baseline data collection;
- £150 following endline assessments;
- Paid staff time for administering LanguageScreen at baseline; and
- £1,000 at the end of the trial, intended for purchasing the NELI Preschool programme.

Both intervention and control settings received assessment feedback reports at the end of the trial. These included child-level assessment feedback, downloadable from the LanguageScreen portal. RAPT scores were not shared, as they are not suitable for interpretation in isolation and must be considered within a broader context.

Participant selection

Eligibility criteria for settings

Settings fulfilling the following criteria were eligible to participate in this trial:

- Settings were either state-maintained or PVI nurseries. Settings registered as 'childcare on domestic premises' were also eligible to take part in the trial, as these settings had more than four people who provided care and the settings were subject to the same requirements as nurseries for staffing ratios and staff qualification levels.
- Settings needed to have a minimum of fourteen 3- to 4-year-old children attending 15 or more hours per week during the 2023–2024 academic year (starting September 2023). This criterion was a proxy for the minimum number of children required for the trial in academic year 2024–2025 i.e., minimum of 12 children per setting¹¹.
- Settings must have been willing to allocate time for a minimum of two nursery staff members to complete online NELI Preschool training, encouraging as many practitioners who worked with 3 to 4 year olds as possible to complete the online training.
- Settings must have been located within the following Local Authorities from nine SPHs: Bexley, Bournemouth, Christchurch and Poole, Bracknell Forest, Brighton and Hove, Bromley, Buckinghamshire, Cambridgeshire, Cornwall, Croydon, Darlington, Derby, Derbyshire, Devon, Dorset, Durham, Ealing, East Riding of Yorkshire, East Sussex, Gateshead, Greenwich, Hammersmith and Fulham, Hampshire, Hartlepool, Hillingdon, Hounslow, Isle of Wight, Kent, Kingston upon Hull, Kingston upon Thames, Lambeth, Lewisham, Medway, Merton, Middlesbrough, Milton Keynes, Newcastle upon Tyne, Norfolk, North East Lincolnshire, North Lincolnshire, North Tyneside, North Yorkshire, Northumberland, Nottingham, Nottinghamshire, Oxfordshire, Plymouth, Portsmouth, Reading, Redcar

¹¹ A minimum of 12 children per setting were required for trial analysis such that analysis samples included six children from the enrichment component and six from the targeted component of NELI Preschool programme.

and Cleveland, Richmond upon Thames, Slough, South Tyneside, Southampton, Southwark, Stockton-on-Tees, Suffolk, Sunderland, Surrey, Sutton, Torbay, West Berkshire, West Sussex, Wandsworth, Windsor and Maidenhead, Wokingham, and York.¹²

The following exclusion criteria were also applied at recruitment:

- Settings were not eligible if they were participating in another SPH-funded programme delivered in the 2024–2025 academic year, including allocation to the control group for one of the other SPH trials (for example, Early Talk Boost, The ONE Programme, EYCP, Concept Cat, Communication Friendly Settings).
- Settings were not eligible if they took part in another EEF-funded programme or early years evaluation in the 2024–2025 academic year. Although this was identified as an eligibility criterion, there was some flexibility for the OxEd team (in agreement with NFER and the EEF) to recruit settings to the trial even if they were part of an EEF evaluation¹³. Decisions were made on a case-by-case basis and settings were deemed eligible if the programme in question was dissimilar to NELI Preschool in terms of its recipients and/or aims, or it was viewed as business as usual (BAU).
- Settings from the efficacy trial or those taking part in the formative evaluation were not eligible for this trial.

Settings were only randomised if they fulfilled baseline data collection requirements. These included: providing required data for the eligible children (see below) prior to baseline LanguageScreen assessment, completing the baseline LanguageScreen assessments with eligible children, uploading this data on the LanguageScreen portal, and completing the baseline IPE survey.

Settings that were part of a nursery chain or a nursery group were eligible to take part as an individual unit. Given that each setting was considered a separate unit for randomisation, these settings needed to agree to adhere to random group allocation. OxEd confirmed this with each setting during recruitment.

If settings were taking part in DfE's Early Years Professional Development Programme (EYPDP), they were still eligible to take part in this trial, as EYPDP was considered as BAU.

Eligibility criteria for children

All 3- to 4-year-old children as on 31st August 2024 and those who could engage with LanguageScreen assessment (for example, do not have complex special educational needs¹⁴) were assessed with LanguageScreen at baseline and formed part of the trial. This also included children who attended the setting for less than 15 hours a week, as some of these children could subsequently be discovered to be entitled to EYPP. Only the children with baseline LanguageScreen were included in the trial. Settings were asked not to assess children who were not between 3 and 4 years old on 31st August 2024 or who have complex special educational needs, even though they may be present in the preschool classroom. These children could have joined whole-group sessions (if their setting was randomised to the intervention group) but were not part of the trial. Additionally, 3- to 4-year-old children (as on 31 August 2024) who had the lowest baseline LanguageScreen scores, were attending the setting for 15 or more hours per week, were able to access small-group work, and for whom settings were able to schedule all sessions were eligible for selection for the targeted component of the intervention. Settings were instructed to select these children according to the Targeted Group Selection Guidance (Appendix I).

¹² To meet recruitment targets, settings were also recruited from two additional SPH areas: St Edmunds SPH and the Bright Futures Northwest Early Years SPH.

¹³ Some of the notable exceptions were the Maths Champion's programme, Tales Toolkit, and the ONE project.

¹⁴ Settings were given guidance about how to identify children with complex special educational needs. 'Complex special educational needs' in this guidance was described as: children who have such severe auditory, sensory, visual, or behavioural difficulties that they would be unable to access a five-minute LanguageScreen assessment. Settings were strongly encouraged to refer to their SENCo and speech and language team for further guidance on suitability.

Analysis sample selection

In order to answer the research questions outlined above, four analysis samples were defined and subsequently drawn according to Table 6 below.

Table 6: Analysis samples

S1: Ideally Targeted Children	Children who should have been selected for the targeted component of NELI Preschool if settings correctly followed OxEEd's Targeted Group Selection Guidance for their preferred implementation model. This sample was selected after NFER had received final information from settings regarding children who had left or been withdrawn so that they can be excluded from selection and includes children who are ideally targeted based on OxEEd guidance.
S2: Practitioner Targeted Children (intervention settings only)	Children in intervention settings who were actually selected for targeted intervention after randomisation. This sample reflects usual practice outside the trial. Settings were requested to select six children per class if implementing Models A or B, and three children per class if implementing Model C.
S3: Enrichment-only Sample	Children who attend the setting at least 15 hours per week and are not part of S1 or S2. NFER randomly selected six children per class from each setting. The sample includes children who were receiving whole-group sessions only.
S4: EYPP	All 3- to 4-year-old children in the setting who are entitled to EYPP. This included children who attended the settings for fewer than 15 hours a week and included some who were part of the other three samples.

These samples were selected by NFER post-randomisation after final information had been received from settings regarding attendance patterns, EYPP status, and whether children had left settings or been withdrawn from the trial. The combined samples provided the list of children who would undertake endline testing.

Outcome measures

Primary outcome

The primary outcome for this study was oral language skills, as measured by a latent oral language variable formed from the expressive and receptive subtests in LanguageScreen and the RAPT that were administered at endline. LanguageScreen is designed to evaluate oral language skills in young children. It is a language assessment tool accessed via an app on a tablet. It is administered one to one with children and takes approximately five to ten minutes to complete. It was standardised using data collected from 8,273 schools and 348,944 children aged 42 to 107 months (equivalent to 3 years and 6 months to 8 years and 11 months) between 2018 and 2020 (Hulme *et al.*, 2024). It demonstrated strong reliability for the total score with a Cronbach's alpha of 0.92 and McDonald's omega hierarchical coefficient of 0.75. The individual subscales also showed reliability ranging from 0.74 to 0.80. The test-retest reliability of LanguageScreen was 0.78 ($N = 9,778$). It is comprised of four subtests:

1. Expressive Vocabulary (24 items; raw score range 0–24) assesses the ability to name pictures.
2. Listening Comprehension (4–12 items; standardised score range 65–135) assesses children's ability to understand spoken language by asking questions about short stories being played to them. For this subtest, children's age determines the number of stories and questions they receive. Children younger than 4.5 are now played one story and are asked four questions (raw score range: 0–4) while children aged 4.5 and above are played two stories and are asked 12 questions (raw score range: 0–12). There is no overlap between the stories offered to the two different age groups.
3. Receptive Vocabulary (23 items; raw score range 0–23) assesses the ability to match spoken words to pictures by asking them to match a word they hear to one of four pictures on the screen.
4. Sentence Repetition (14 items; raw score range 0–14) assesses the ability to repeat sentences verbatim by asking children to repeat the sentences they hear.

RAPT tests the speech and language development of children who are between 3 and 8 ½ years of age by using ten picture cards, depicting a range of everyday scenarios, that stimulate children to give samples of spoken language that can be evaluated in terms of grammatical structures, sentence length, and identifying information. It has two score schemes (assessing either information or grammar), both of which use the same ten items and is scored independently using scoring guidelines:

- i Information (10 items, score range 0–41), where children are asked to describe the information shown in a set of pictures.
- ii Grammar (10 items, score range 0–39), which checks the grammar used by children, such as the use of verb tenses, while describing the information shown in a set of pictures.

Similar to LanguageScreen, RAPT is a one to one administered test. It was re-standardised in 2018 where the assessment was conducted digitally with the images presented to a child on screen. All images and scoring guidelines from digital RAPT are identical to those used in the print version, and therefore the re-standardised data can be used consistently across both print and digital versions. Over 40 Speech and Language Therapists assessed 882 children from mainstream schools in the UK and were aged 3–8.5 years. All children in the sample spoke English as their first language. For this trial, the print version of RAPT was used.

The latent oral language variable was constructed from a Principal Component Analysis (PCA) of the LanguageScreen subtest scores and RAPT Information and Grammar scores collected at endline, with the first component from this analysis used as the primary outcome measure. For both LanguageScreen and RAPT, the raw scores were used in the PCA, apart from the Listening Comprehension score, for which the standardised score was used, as children received two different versions of this subtest depending on their age. At baseline, all children were presented with 3 stories and asked 16 questions. However, at endline, the number of stories and questions varied by age: children younger than 4.5 were played one story and were asked four questions (raw score range: 0–4), while children aged 4.5 and above were played two stories and were asked 12 questions (raw score range: 0–12). Each age group received its own distinct set of stories, with no overlap between them. As the Listening Comprehension scores at endline are therefore on different scales depending on age, with children on either side of the 4.5 age threshold, the standardised score was used in all analyses instead of the raw score.

This initially raised concerns about the possible loss of information. To assess the suitability of this approach, a correlation analysis was conducted on baseline LanguageScreen data. In this analysis, two PCAs were performed. The first used the *raw* Listening Comprehension score along with the raw scores for the other LanguageScreen subtests (Expressive Vocabulary, Receptive Vocabulary, and Sentence Repetition), and the first component was extracted. The second used the *standard* Listening Comprehension score along with the raw scores for the three other LanguageScreen subtests, and again, the first component was extracted. These two components were then correlated, yielding a Pearson correlation coefficient greater than 0.9. This indicated that the proposed approach was appropriate and did not result in any appreciable loss of information. The scores were all centred and scaled as part of the analysis, and the PCA was run using the built-in *prcomp* function within R (version 4.2.1; R Core Team 2022).

Administration of LanguageScreen and RAPT at endline

At endline, both LanguageScreen and the RAPT were administered by trained independent NFER Test Administrators, rather than by practitioners within the settings who administered LanguageScreen at baseline. This approach was required to ensure that all children were assessed under standardised conditions by individuals blind to randomised groups, providing consistency in administration across intervention and control settings.

Before the endline assessment window, NFER ran a small-scale pilot to explore how best to administer LanguageScreen and the RAPT in early years settings, with the aim of identifying potential challenges and refining procedures ahead of endline data collection in the main trial. The pilot highlighted the need for clearer guidance for Test Administrators, greater consistency in administration, and stronger processes for managing testing spaces, engaging children, and accurately capturing verbatim responses. In response, the main trial incorporated enhanced Test Administrators training, a detailed handbook and one-page checklist, improved instructions on tablet and microphone positioning, and strengthened guidance on transitions between assessments and use of motivational prompts such as stickers. Clearer stopping rules were also introduced, and digital submission of RAPT responses replaced handwritten returns to improve accuracy and

legibility. These adaptations helped ensure that endline assessments were delivered consistently and to a high standard across all settings.

NFER Test Administrators conducted the assessment one to one using NFER-issued tablets, which contained the LanguageScreen app, a voice-recording app for RAPT, and SIM cards to ensure assessments could proceed even where WiFi was limited. The RAPT was administered immediately after LanguageScreen and involved Test Administrators presenting the ten picture cards, asking the scripted questions verbatim, voice-recording each child's response, and transcribing answers by hand before later submitting them electronically via a dedicated online form for each child. These recorded responses were subsequently subject to NFER's quality assurance processes to ensure transcription accuracy.

Secondary outcomes

The six subscales (four LanguageScreen subtests and two RAPT subtests) outlined were treated as secondary outcome measures, with analyses run separately for each. In five of the six analyses, the raw endline subscore was used. The analysis for the LanguageScreen Listening Comprehension subtest used the standard score for the reasons described above.

Other outcomes

Practitioner confidence was conceptualised as a short-term outcome in the programme's ToC, reflecting the assumption that NELI Preschool sought not only to improve children's language skills but also to enhance practitioners' knowledge, confidence, and motivation to support early language development more broadly. For this reason, practitioner confidence was measured both to support the planned mediation analysis and as an important secondary outcome within the IPE. Please see the trial protocol (Roy *et al.*, 2024) for further details.

For the mediation analysis, practitioner confidence was measured using a version of the second scale of the Early Math Beliefs and Confidence Survey (EM-BCS) that was adapted for assessing teachers' confidence in helping preschool children (aged 3 to 4 years) to learn maths (Chen and McCray, 2013). The measure used in this trial was further adapted to capture the shift in domain (from maths to language development) and consisted of eleven questions assessing practitioners' confidence in their knowledge of, and ability to support, four key areas of practice promoted by the programme: (1) explicitly teaching vocabulary, (2) providing opportunities to practise new words in varied contexts, (3) supporting children to explore story structure through dialogic reading, and (4) offering scaffolded support to children with the weakest language skills. Each question was answered using a 5-point Likert scale (from strongly disagree to strongly agree). The full set of questions can be found in Appendix J. The questions were included in the practitioner outcomes surveys that were administered at both baseline and endline.

As specified in the study protocol, practitioner confidence was originally expected to be defined at the setting level, with individual practitioner scores generated by summing their responses to the 11 survey items (yielding a total score between 11 and 55). A mean score would then be calculated across practitioners within each setting. However, during development of the SAP, concerns were raised that the 11 items were correlated and that this correlation structure should be accounted for. Therefore, an Exploratory Factor Analysis (EFA) was conducted using baseline practitioner confidence data to determine the most appropriate method for constructing the measure. The EFA indicated that, although the items were indeed highly correlated, the latent score derived from a one-factor Confirmatory Factor Analysis (CFA) was almost perfectly correlated with the simple summed score ($r^2 > 0.99$). On this basis, the practitioner confidence score at both baseline and endline was calculated as originally outlined in the protocol—by summing item responses for each practitioner and then averaging scores at the setting level. Results from the EFA are presented in Appendix D of the SAP (Roy *et al.*, 2025a).

Baseline measures

Only LanguageScreen was administered at baseline and used to form the baseline measure for the primary outcome. As with the primary outcome measure, a PCA was performed on the four baseline LanguageScreen subscores (using raw scores apart from Listening Comprehension, for which the standard score was used). The first component formed the baseline measure, and the PCA followed the same procedure as that used for the primary outcome measure.

The four baseline LanguageScreen subscales were used as a baseline measure for the corresponding endline subscale. As RAPT was only administered at endline, there was no corresponding RAPT Information or Grammar baseline subscales. Instead, the same baseline measure as the primary outcome, the first component obtained from the PCA performed on the four baseline LanguageScreen subscores, was used in these two analyses.

As outlined above, we calculated baseline practitioner confidence by summing the scores from the baseline survey questions for each practitioner and then computing the setting-level mean of these totals.

Sample size

Table 17 in the Impact evaluation results section sets out the parameters used to determine the minimum detectable effect sizes (MDES) at the protocol, randomisation, and analysis stages (EEF, 2022). Sample size calculations were carried out using the PowerUpR (Bulus, Dong and Spybrook, 2021) package within RStudio. The number of settings and children used in the calculations at the protocol and randomisation stages represents the samples before any attrition, with the MDES values calculated using an allowance for 10% setting-level attrition and 23% child attrition, in line with the EEF EY lessons learned (EEF, 2019). The ICC and pre-post correlations used in the protocol and randomisation calculations are taken from two prior NELI trials. The NELI effectiveness trial (Dimova *et al.*, 2020) targeted the lowest language ability children and yielded intracluster correlation coefficient (ICC) and pre-post correlation estimates of 0.349 and 0.75, respectively. By contrast, the NELI Preschool efficacy trial (West *et al.*, 2023) used all children in the class, giving a lower ICC of 0.21 (reflecting the more heterogeneous population) and a higher pre-post correlation of 0.81. The populations used in the analyses to answer RQ1 (all children), RQ5 (EYPP children), and RQ6 (children in PVI settings) were expected to be relatively heterogeneous, so the parameters used in these sample size calculations are taken from the NELI Preschool efficacy trial. The populations used in the analyses to answer RQ3 (targeted) and RQ4 (enrichment-only) were expected to be rather more homogenous within settings and hence used the parameters from the NELI effectiveness trial (Dimova *et al.*, 2020) to reflect this. Sample size calculations were carried out to ensure that the trial was sufficiently powered to detect an impact for the EYPP subgroup. This means that the number of settings and children included was large enough to reliably identify a meaningful effect, should one exist, within this specific subgroup.

Randomisation

Randomisation was undertaken at the setting level for this trial. The 303 early years settings who returned the necessary child data and had completed at least one baseline assessment were randomised on a one to one basis into two groups, intervention and control, stratified by setting type (maintained nurseries or PVI settings) and setting size (thirty or fewer 3–4-year-old vs more than thirty 3–4-year-olds) to ensure equal group allocation in each stratum.

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

The randomisation allocation sequence was generated using the sample function within R and used a simple stratified sample approach, with settings randomised into two arms (intervention and control) on a one to one basis. Randomisation was carried out by the project statistician who was not blinded to randomisation but was provided with the minimum data necessary to carry out the randomisation correctly. Then 151 settings were randomised to the intervention group with the remaining 152 randomised by default to the control group. Table 7 below sets out the results of the randomisation. The syntax used to complete the randomisation is in Appendix K.

Table 7: Randomisation results

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

Statistical analysis

As seen above, four analysis samples were drawn as per Table 6.

Primary analysis

RQ1: How effective is NELI Preschool at improving oral language skills of 3–4-year-old children in intervention settings compared to children in control group settings?

The primary analysis sought to answer the research question outlined above and has been answered using the following two-level random intercept model:

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

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the random intercept for setting j , $intervention_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention), $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, and ε_{ij} is the residual error term for child i at setting j .

Both the random intercept and residual term were assumed to be independently normally distributed:

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

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

The analysis used a combined sample comprising children from S1 (ideally targeted children) and S3 (enrichment-only sample) across both intervention and control settings, and the model was run in R (version 4.2.1) using the package 'lme4'.

Secondary analysis

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

This research question was answered by the secondary analysis, using six multilevel models. For each, the secondary outcome measure was the total score for each subtest of LanguageScreen and RAPT at endline as outlined in the outcome measures section above. Six multilevel models were run, with each secondary outcome forming the dependent variable. As per the primary analysis model, each model accounted for the clustering of children within settings as a random effect and included the same covariates. The two-level models are given by:

$$\text{ExpressiveVocabulary}_{ij} = \beta_0 + u_{0j} + \beta_1 \text{intervention}_j + \beta_2 \text{BL ExpressiveVocabulary}_{ij} + \beta_3 \text{SettingType}_j + \beta_4 \text{SettingSize}_j + \varepsilon_{ij}$$

$$\text{ListeningComprehension}_{ij} = \beta_0 + u_{0j} + \beta_1 \text{intervention}_j + \beta_2 \text{BL ListeningComprehension}_{ij} + \beta_3 \text{SettingType}_j + \beta_4 \text{SettingSize}_j + \varepsilon_{ij}$$

$$\text{ReceptiveVocabulary}_{ij} = \beta_0 + u_{0j} + \beta_1 \text{intervention}_j + \beta_2 \text{BL ReceptiveVocabulary}_{ij} + \beta_3 \text{SettingType}_j + \beta_4 \text{SettingSize}_j + \varepsilon_{ij}$$

$$\text{SentenceRepetition}_{ij} = \beta_0 + u_{0j} + \beta_1 \text{intervention}_j + \beta_2 \text{BL SentenceRepetition}_{ij} + \beta_3 \text{SettingType}_j + \beta_4 \text{SettingSize}_j + \varepsilon_{ij}$$

$$\text{Information}_{ij} = \beta_0 + u_{0j} + \beta_1 \text{intervention}_j + \beta_2 \text{BL OralLanguage}_{ij} + \beta_3 \text{SettingType}_j + \beta_4 \text{SettingSize}_j + \varepsilon_{ij}$$

$$\text{Grammar}_{ij} = \beta_0 + u_{0j} + \beta_1 \text{intervention}_j + \beta_2 \text{BL OralLanguage}_{ij} + \beta_3 \text{SettingType}_j + \beta_4 \text{SettingSize}_j + \varepsilon_{ij}$$

Where *ExpressiveVocabulary_{ij}* is the endline Expressive Vocabulary raw score for child *i* at setting *j*, *BLExpressiveVocabulary_{ij}* is the baseline Expressive Vocabulary raw score for child *i* at setting *j*, *ListeningComprehension_{ij}* is the endline Listening Comprehension standardised score for child *i* at setting *j*, *BLListeningComprehension_{ij}* is the baseline Listening Comprehension standardised score for child *i* at setting *j*, *ReceptiveVocabulary_{ij}* is the endline Receptive Vocabulary raw score for child *i* at setting *j*, *BLReceptiveVocabulary_{ij}* is the baseline Receptive Vocabulary raw score for child *i* at setting *j*, *SentenceRepetition_{ij}* is the endline Sentence Repetition raw score for child *i* at setting *j*, *BLSentenceRepetition_{ij}* is the baseline Sentence Repetition raw score for child *i* at setting *j*, *Information_{ij}* is the endline Information raw score for child *i* at setting *j*, and *Grammar_{ij}* is the endline Grammar raw score for child *i* at setting *j*. As in the primary outcome analysis, *u_{0j}* is the random intercept for setting *j*, *intervention_j* is the intervention/control dummy variable for setting *j* (0 = control; 1 = intervention), *BLOralLanguage_{ij}* is the baseline latent oral language variable for child *i* at setting *j* and is measured via the first component of a PCA run on the baseline data as outlined above, *SettingType_j* is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, *SettingSize_j* is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, and *ε_{ij}* is the residual error term for child *i* at setting *j*.

In each model, both the random intercept and the residual term were assumed to be independently normally distributed as specified for the primary outcome analysis. The analysis samples used were S1 (ideally targeted children) and S3 (enrichment-only sample) in both intervention and control settings, and as per the primary analysis, the two-level random effects models were run in R using the package ‘lme4’.

Subgroup analyses

Several subgroup analyses were undertaken to ascertain the intervention’s effect on specific subgroups as defined in Table 8 for RQ3a to RQ7. Each analysis used a two-level random-intercepts model consistent with the primary analysis (equation 1), applying the same outcome and covariates unless a covariate was not relevant for that subgroup (for example, setting type was excluded for RQ6).

In addition, interaction models were fitted to examine whether the intervention effect differed between subgroups of interest. These were run between EYPP and non-EYPP children, and for children with EAL vs no EAL, as planned. Three additional post-hoc analyses were also run; two of these provided a more complete understanding of NELI Preschool’s

impact for maintained settings (RQ6b) and whether there were notable differences between PVI and maintained settings (interaction model). A third interaction model was also run as a post-hoc analysis comparing targeted children with those receiving only the enrichment component to provide a more complete view of any differential impact between children ideally selected for the targeted component (S1) and those receiving only whole-class enrichment (S3). See below for more detail.

All subgroup analyses were run using R (version 4.2.1) using the package ‘lme4’. Table 8 sets out each research question and the corresponding analysis sample used in the model.

Table 8: Subgroup research questions and analysis samples used

Research question	Analysis samples used
RQ3a: How effective is NELI Preschool at improving the language skills of the subgroup of six children selected to receive the targeted component of the intervention?	S2 (practitioner targeted children) for intervention settings and S1 (ideally targeted children) for control settings
RQ3b: How effective is NELI Preschool at improving the language skills of the subgroup of six children who should have been selected to receive the targeted component of the intervention?	S1 analysis sample (ideally targeted children) for both intervention and control settings
RQ4: How effective is NELI Preschool at improving the language skills of children who only receive the whole-class (enrichment) component of the intervention?	S3 analysis sample (enrichment-only sample) for both intervention and control settings Interaction model (post-hoc analysis): S1 analysis sample (ideally targeted children) and S3 analysis sample (enrichment-only sample) for both intervention and control settings
RQ5: How effective is NELI Preschool at improving the language skills of disadvantaged children?	EYPP-only model: S4 analysis sample (all children eligible for EYPP) for both intervention and control settings Interaction model: S1 analysis sample (ideally targeted children), S3 (enrichment-only sample), and S4 (EYPP-eligible) for both intervention and control settings
RQ6a: Is NELI Preschool effective at improving language skills of 3- to 4-year-old children in PVI intervention settings compared to children in PVI control settings?	S1 analysis sample (ideally targeted children) and S3 (enrichment-only sample) for both intervention and control settings at PVI settings only
RQ6b: Is NELI Preschool effective at improving language skills of 3- to 4-year-old children in maintained intervention settings compared to children in maintained control settings? ¹⁵	S1 analysis sample (ideally targeted children) and S3 (enrichment-only sample) for both intervention and control settings at maintained settings only
RQ7: How effective is NELI Preschool at improving the language skills of EAL children?	EAL-only model: S1 analysis sample (ideally targeted children) and S3 (enrichment-only sample) for both intervention and control settings for EAL children only Interaction model: S1 analysis sample (ideally targeted children) and S3 (enrichment-only sample) for both intervention and control settings for all children

In order to answer question RQ5, two models were run. The first replicated the primary analysis model (equation (1)) on the subgroup of EYPP-eligible children (S4 for both intervention and control settings). The second was similar to the primary analysis model but, as well as the S1 and S3 samples, it additionally included all EYPP-eligible children (all S1 + S3 + S4 children in both intervention and control settings), as well as an EYPP eligibility covariate and interaction term between that covariate and the intervention/control dummy variable. This interaction model is defined below:

¹⁵ This was a post-hoc analysis that, in light of the RQ6a findings, was conducted to provide a more complete understanding of NELI Preschool’s impact across both PVI and maintained settings, and to explore whether any notable differences emerged between the two.

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

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the random intercept for setting j , $intervention_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention), $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, $EYPP_{ij}$ is the binary variable that shows EYPP eligibility (0 = not eligible; 1 = eligible) for child i at setting j , $intervention_j * EYPP_{ij}$ is the interaction between the intervention/control dummy variable and EYPP eligibility for child i at setting j , and ε_{ij} is the residual error term for child i at setting j .

As will be shown in the Impact evaluation results section, the analysis for RQ6a indicated that, within PVI settings, the difference in oral language skills between intervention and control children was slightly lower than the overall effect size and the confidence interval contained zero. Therefore, an additional subgroup analysis was conducted to investigate whether this overall effect was being driven mainly by children in maintained settings. This was an exploratory post-hoc analysis, not prespecified in the SAP. It was carried out to provide a more complete understanding of potential differences in outcomes by setting type, seeking to answer the new research question RQ6b: Is NELI Preschool effective at improving language skills of 3- to 4-year-old children in maintained intervention settings compared to children in maintained control settings? It used the same model definition as that of RQ6a. As with any post-hoc comparison, the results should be interpreted with caution, as they were not protected by prospective statistical planning and may be more susceptible to chance findings arising from multiple comparisons.

As these analyses focused on setting types individually, we removed the setting type covariate (as it was redundant), resulting in a model defined as:

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

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the random intercept for setting j , $intervention_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention), $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) used as a stratifier at randomisation, and ε_{ij} is the residual error term for child i at setting j .

Alongside this, a model was run that included an interaction term between setting type and interaction in order to fully assess whether the effect of the intervention differed by setting type. The model is defined as below:

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

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the random intercept for setting j , $intervention_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention), $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, $intervention_j * SettingType_j$ is the interaction between the intervention/control dummy variable and setting type for setting j , and ε_{ij} is the residual error term for child i at setting j .

In order to answer question RQ7, two models were run. The first replicated the primary analysis model (equation (1)) on the subgroup children identified as having EAL. The second was similar to the primary analysis model but included EAL as a covariate as well as an additional interaction term between the EAL covariate and the intervention/control dummy variable. This interaction model is defined below:

$$\begin{aligned} OralLanguage_{ij} = & \beta_0 + u_{0j} + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j \\ & + \beta_4 SettingSize_j + \beta_5 EAL_{ij} + \beta_6 intervention_j * EAL_{ij} + \varepsilon_{ij} \end{aligned} \quad (13)$$

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the random intercept for setting j , $intervention_j$ is the intervention/control dummy variable for setting j (0 = control; 1 = intervention), $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, EAL_{ij} is the binary variable that shows EAL status (0 = not EAL; 1 = EAL) for child i at setting j , $intervention_j * EAL_{ij}$ is the interaction between the intervention/control dummy variable and EAL status for child i at setting j , and ε_{ij} is the residual error term for child i at setting j .

For all models run for the subgroup analyses, both the random intercept and residual term were assumed to be independently normally distributed and are defined as per those for the primary outcome analysis.

Analysis in the presence of non-compliance

Compliance of intervention settings was measured at the setting level and defined as completion of NELI Preschool training and delivery of a minimum proportion of NELI Preschool whole-group, small-group, and individual sessions. It was measured using a binary compliance variable, with settings being considered compliant (compliance = 1) if they fulfilled **all** the following criteria:

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

Data on the number of sessions delivered was recorded on session delivery logs (SDLs) completed by settings for each week and returned to NFER in five-week batches. As per the analysis plan, SDLs that covered Weeks 1–15 of NELI Preschool delivery were used to derive the proportions for criteria 2–4 and only settings that returned the complete set of SDLs for Weeks 1–15 were included. This was subject to the correlation between the data from Weeks 1–15 SDLs and from Weeks 16–20 indicating that this is an appropriate strategy, which it did—the correlation coefficient indicated a strong positive relationship between the two ($R = 0.768$; $p < 0.001$).

Whole-group and small-group sessions were deemed to have been delivered if at least one child was in attendance in that session for each setting. For individual sessions, the total expected number of sessions for each setting was calculated by multiplying the number of children in the S1 sample within a setting by the number of expected weekly individual sessions (i.e., 3) and the number of weeks that the intervention was delivered for. The numerator for the percentage calculations was the actual number of sessions that settings delivered within Weeks 1–15 and was derived from the SDLs. Whilst most settings completed the SDLs correctly, a number of settings completed them to state that selected children attended sessions, including small-group and one to one sessions, every day. These have been interpreted as children having attended sessions as expected, and therefore weekly small-group and one to one session attendance has been capped at 3 and 1, respectively. Additionally, one setting reported that all children listed attended every session across all weeks, regardless of whether they had been selected for the targeted component. This issue was only identified during the analysis stage, as quality assurance at the point of data collection focused on ensuring that SDLs were not returned empty, rather than verifying whether sessions had been recorded accurately. By the time this discrepancy was identified, too much time had elapsed since delivery to allow for meaningful follow-up with the setting to confirm the true attendance patterns for the

targeted component. As a result, and given the inability to determine the correct attendance data, this setting was excluded from the compliance (dosage) analysis.

This approach means that the compliance analysis only included a subset of the settings used in the primary outcome analysis (113 of the 151 intervention settings). Therefore, the primary outcome analysis was repeated using only these settings in order to check the robustness of the primary outcome analysis results.

A complier average causal effect (CACE) estimate was obtained using instrumental variable modelling (IV) as prescribed by the EEF statistical analysis guidance. The IV regression was then run via a two-stage least squares model with random group allocation as the instrumental variable for compliance and is given by:

Stage One:

$$Compliance_{ij} = \beta_0 + \beta_1 intervention_j + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij}$$

Stage Two:

$$OralLanguage_{ij} = \beta_0 + \beta_1 \widehat{Compliance}_{ij} + \beta_2 BL\ OralLanguage_{ij} + \beta_3 SettingType_j + \beta_4 SettingSize_j + \varepsilon_{ij}$$

Where $\widehat{Compliance}_{ij}$ is the estimated compliance measure for child i in setting j obtained from the first stage of the two-stage regression, $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, $intervention_j$ is the intervention/control dummy variable for setting j , $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, and $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation. The residuals in both stages are assumed to be independently normally distributed:

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

The model was fit using the function `ivreg` from the R package 'ivreg'¹⁶ and the estimation of causal effects was done using the functions contained in the R package 'ivpack'¹⁷. The analyses were, as before, run in R (version 4.2.1).

Missing data analysis

Approximately 13% of children selected for endline assessment failed to complete both of the assessments and were thus not included in the final analysis sample. As this is greater than the 5% threshold, as per the EEF's statistical analysis guidance (EEF, 2022), further missing data analysis was undertaken. By design, only children who were assessed at baseline were eligible to be selected for endline testing. Therefore, missingness was investigated only in relation to the primary outcome measure. A substantive model was used to investigate patterns of missingness. This was a two-level (child and setting) logistic regression model, with baseline oral language skills, intervention/control dummy variable, setting type and setting size indicators included as covariates as per the primary outcome analysis, along with other covariates that were considered might be indicative of missingness. This list includes:

¹⁶ John Fox, Christian Kleibler, Achim Zeileis, and Nikolas Kuschknig (2024). `ivreg`: Instrumental-Variables Regression by '2SLS', '2SM', or '2SMM', with Diagnostics. R package version 0.6–3. <https://CRAN.R-project.org/package=ivreg>

¹⁷ Yang Jiang and Dylan Small (2014). `ivpack`: Instrumental Variable Estimation. R package version 1.2. <https://cran.r-project.org/src/contrib/Archive/ivpack/>

Child level:

- EYPP eligibility
- EAL status
- Gender
- Age at baseline in months.

Setting level:

- Proportion of 3- to 4-year-old children eligible for EYPP within the setting
- Setting's Income Deprivation Affecting Children Indices (IDACI) decile
- Implementation model
- Latest Ofsted rating
- Setting's SPH.

To run this analysis, settings were linked to the relevant information contained on the most up-to-date childcare providers and inspections statistics published by Ofsted (Ofsted, 2024). This dataset was used to access demographic information for PVI settings that isn't held on DfE's Get Information about Schools (GIAS) register. However, this meant that we were unable to include any published data about the overall proportion of EYPP-eligible children at each setting and whether they are in urban or rural areas, as this information is not available on publicly accessible data sources.

Covariates that were found to be predictive of missingness (EAL and SPH) in the logistic regression model were then included in the primary analysis model, which was rerun, and the results were compared to those obtained from the primary outcome model. The inclusion of these covariates made no material difference to the estimated intervention effect. This stability in the intervention effect indicates that the relationship between missingness and these covariates does not introduce bias into the outcome model and is compatible with the assumption that the data are Missing at Random (MAR) conditional on the included covariates. If the data were Missing Not at Random (MNAR), adjusting for predictors of missingness would typically be expected to alter the intervention effect estimates, because unobserved mechanisms would continue to bias the model. The absence of such changes, therefore, reduces concern about MNAR mechanisms. On this basis, no further sensitivity analyses were undertaken. Both the logistic regression model for missingness and the adjusted primary outcome model were run using the 'lme4' model in R (version 4.2.1).

Additional analyses and robustness checks**Dosage analysis**

The ToC hypotheses children's oral language skills are improved through exposure to a variety of outputs, including attendance at both whole-group and targeted sessions, experience of a wide variety of words, and opportunities to engage in activities intended to improve their language skills. A dosage analysis was undertaken to determine whether higher exposure to the various NELI Preschool inputs leads to better oral language outcomes. As part of this, the analysis explored whether attendance at both the enrichment and targeted components of NELI Preschool leads to higher oral language skills than attendance at the enrichment component only. This analysis focused on what was delivered rather than what was planned, using data received from the intervention settings about children's attendance at each type of NELI Preschool session via SDLs. These were completed by settings for each week and returned to NFER in five-week batches. In each, practitioners recorded the number of whole-group sessions and targeted sessions each child attended each week during the programme delivery. Only settings that completed all SDLs for at least Weeks 1–15 were included in the dosage analysis. Where data for Weeks 16–20 was missing, this was imputed using the coefficients from a linear model for each of the three session types individually. As a sense-check for this, a correlation analysis between data from SDLs in Weeks 1–15 and data from SDLs in Weeks 16–20 was carried out. The Pearson correlation coefficient ($R = 0.768$; $p < 0.001$) indicated that this was an appropriate approach to undertake.

The dosage analysis included all children who received the intervention and for whom we had endline assessment data and information about session attendance from SDLs, regardless of which analysis sample(s) they were in. Whilst this mostly corresponded to analysis samples S2 and S3 in intervention settings, it did include a number of children in S1 who weren't selected by practitioners for targeted intervention but still participated in whole-group enrichment sessions, and EYPP-eligible children (S4) who attended fewer than 15 hours per week.

Two separate analyses were conducted, with each taking the form of a linear model that set the primary outcome measure as the dependent variable and included the appropriate dosage measure(s) as (a) covariate(s). The models also included all other covariates used in the primary analysis model. The first model analysed the impact of the enrichment component only and included all children who attended enrichment sessions only. The two-level random intercepts model for this analysis is given by:

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

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the random intercept for setting j , $whole_{ij}$ is the number of whole-group sessions attended by child i at setting j , $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, and ε_{ij} is the residual error term for child i at setting j .

The second analysis analysed the impact of both the enrichment and targeted components and included all children who attended both enrichment and targeted sessions. Three dosage measures were derived: the number of whole-group sessions attended, the number of small-group sessions attended, and the number of individual sessions attended by each child. The two-level random intercepts model for this analysis is given by:

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

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the intercept for setting j , $whole_{ij}$ is the number of whole-group sessions attended by child i at setting j , $small_{ij}$ is the number of small-group sessions attended by child i at setting j , $onetoone_{ij}$ is the number of individual sessions attended by child i at setting j , $BL\ OralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, and ε_{ij} is the residual error term for child i at setting j .

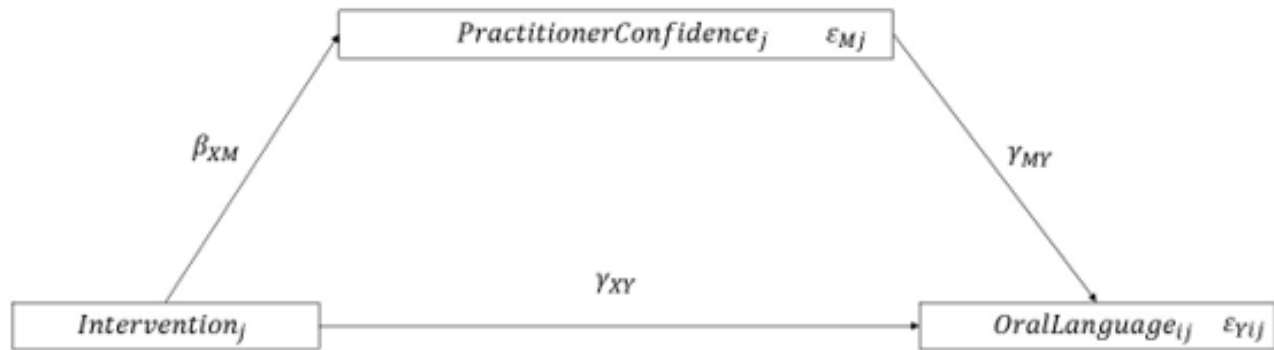
For both analyses, both the random intercept and residual term were assumed to be independently normally distributed as per the primary outcome analysis. As before, the two-level random effects models were run in R (version 4.2.1) using the package 'lme4'.

Mediation analysis

The ToC proposed that increased practitioner confidence would act as a short-term outcome of the intervention and act as a causal mechanism for improving children's language ability. To examine this, a mediation analysis was conducted to assess whether the effect of NELI Preschool on children's oral language skills was mediated through improvements in practitioner confidence.

A basic mediation analysis was assumed as per Figure 2 below.

Figure 2: Mediation model



It used a Structural Equation Model (SEM) approach, using the primary outcome as defined in the primary outcome analysis and setting-level practitioner confidence at both baseline and endline as outlined in the Measures section above. The models used are outlined below, with data assumed to be at two levels (child and setting), with clustering at the setting level. The first equation modelled oral language skills as a function of the mediator, with γ_{MY} measuring the direct effect of practitioner confidence on children's oral language skills and γ_{XY} measuring the direct effect of NELI Preschool on children's oral language skills:

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

Where $OralLanguage_{ij}$ is the latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the endline data as outlined above, u_{0j} is the random intercept for setting j , $intervention_j$ is the intervention/control dummy variable (0 = control; 1 = intervention) for setting j , $BLOralLanguage_{ij}$ is the baseline latent oral language variable for child i at setting j and is measured via the first component of a PCA run on the baseline data as outlined above, $PractitionerConfidence_j$ is the mean endline practitioner confidence score for setting j , $\beta_5 BLPractitionerConfidence_j$ is the equivalent setting-level baseline practitioner score, $SettingType_j$ is the setting type dummy variable (0 = Maintained; 1 = PVI) used as a stratifier at randomisation, $SettingSize_j$ is the setting size dummy variable (0 = thirty or fewer 3 to 4 year olds; 1 = more than thirty 3 to 4 year olds) also used as a stratifier at randomisation, and ε_{Yij} is the residual error term for child i at setting j .

The second equation modelled how the intervention was assumed to affect practitioner confidence. β_{XM} represents how much practitioner confidence changes for intervention practitioners compared to control practitioners:

$$PractitionerConfidence_j = \beta_0 + \beta_{XM}intervention_j + \varepsilon_{Mj}$$

Where, $PractitionerConfidence_j$ is the mean endline practitioner confidence score for setting j , $intervention_j$ is the intervention/control dummy variable (0 = control; 1 = intervention) for setting j , and ε_{Mj} is the residual error term for setting j .

The random effect u_{0j} and two residuals ε_{Mj} and ε_{Yij} were assumed to be independently normally distributed:

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

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

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

It was assumed that the total effect of NELI Preschool (i.e., the effect measured by the primary analysis model) is the sum of the direct effect of NELI Preschool on children's oral language skills and the indirect effect of NELI Preschool mediated through practitioner confidence. This assumption is consistent with standard causal mediation frameworks and additionally assumes that there are no unmeasured confounders affecting both practitioner confidence and the children's language skills. In this study, practitioner confidence was measured post-intervention but prior to endline child assessments, and randomisation at the setting level reduces the risk of confounding in the treatment–mediator and treatment–outcome pathways. Under these conditions, treating the total effect as the sum of direct and indirect effects is appropriate and aligns with the programme's ToC. The direct effect is measured by γ_{XY} whilst the indirect effect is measured by the product $\beta_{XM} * \gamma_{MY}$ (and hence the total effect is given by $\gamma_{XY} + (\beta_{XM} * \gamma_{MY})$). The direct effect represents the effect that NELI Preschool had directly on children's language skills, controlling for practitioner confidence, whilst the indirect effect measures how much of the effect of NELI Preschool is transmitted through practitioner confidence. In both instances, a positive effect indicates that NELI Preschool improves a child's language skills. The extent of the mediation (i.e., the proportion of the effect of NELI Preschool that is mediated through practitioner confidence) has been calculated using the formula:

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

The mediation analysis was run in R (version 4.2.1) using the 'lavaan' package.

Post-hoc sensitivity analyses

Two sensitivity analyses were undertaken to explore potential threats to internal validity identified during the peer review process. The first analysis examined whether the uptake of language-development programmes among control settings introduced bias into the analysis. The second analysis addressed evidence that a slightly higher proportion of practitioners in control settings had prior experience with NELI or NELI Preschool, again examining if this introduced bias. Although in both cases the imbalance was relatively small and thus not explicitly controlled for in the primary analysis, it was considered a potential threat to internal validity according to the EEF guidance on classifying the security of findings and hypothesised to attenuate the estimated treatment effect, meaning that the primary analysis may represent a conservative estimate of impact.

For the first sensitivity analysis, NELI Preschool Lead survey responses from control settings at both baseline and endline were compared. Any setting that reported no use of externally produced language-development programmes for 3 to 4 year olds at baseline but subsequently reported using such a programme during 2025 was removed from the control group for this analysis. In addition, control settings that did not return both baseline and endline NELI Preschool Lead surveys were excluded, ensuring that only settings with a confirmed absence of change in practice were retained. The intervention group remained unchanged.

For the second sensitivity analysis, only settings in which at least one practitioner completed the practitioner endline survey were eligible for inclusion. Within this group, any setting where at least one respondent reported previous experience with NELI or NELI Preschool training was removed from both trial arms. This approach ensured that the comparison was restricted to settings with no prior programme experience in either group, allowing us to determine whether the imbalance in practitioner backgrounds could have biased the estimated treatment effect.

Both analyses used a two-level random-intercepts model consistent with the primary analysis (equation 1), applying the same outcome, covariates, and assumptions around the random intercept and residual term and being run in the same statistical software and package.

Estimation of effect sizes

We have followed the EEF's statistical analysis guidance (EEF, 2022) to calculate appropriate effect sizes for each analysis. For analyses that have a continuous dependent (outcome) variable, the following equation has been used to calculate effect size, expressed as Hedges' g :

$$g = \frac{M_1 - M_2}{SD_{pooled}}$$

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

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

Where n_1 and n_2 are the sample sizes of the first and second groups, respectively, and SD_1 and SD_2 are the standard deviations of the first and second groups, respectively.

The numerator for the effect size calculation is the coefficient of the intervention/control dummy variable from the multilevel model for each analysis. Effect sizes are calculated using the total variance without covariates as the denominator (i.e., equivalent to Hedges' g). Confidence intervals for each effect size have been derived by multiplying the standard error of the intervention/control dummy variable coefficient by 1.96. These have then been converted to effect size confidence intervals using the same formula as the effect size itself.

For analyses that have a binary dependent variable (i.e., the missing data substantive model), effect sizes are reported as an odds ratio with a corresponding 95% CI (confidence interval).

Estimation of ICC

For the primary outcome analysis, ICC has been calculated using:

$$ICC_{setting} = \frac{\sigma_j^2}{\sigma_j^2 + \alpha}$$

Where σ_j^2 is the between-setting variance and α is the between-child variance. Pre-test ICCs have been computed using two-level random intercept models with no covariates, and post-test ICCs have been derived from the primary outcome analysis model described above.

Implementation and process evaluation

Research methods

The evaluation team drew on a range of data collection methods to answer the IPE research questions (set out in the Evaluation objectives section earlier). An overview of the IPE methods is presented in Table 9.

Table 9: IPE methods overview

Research methods	Data collection methods	Participants/ data sources	Data analysis methods	Research questions addressed	Implementation/ logic model relevance
Qualitative	Semi-structured observations	Webinars	Thematic analysis	1 ^a	Fidelity
		Case study whole-group and small-group sessions		1, 2, 3, 4 ^b	Fidelity, Responsiveness, Context and Moderators
		Case study continuous provision (process quality)		4	Context and Moderators
	Structured reviews	Online training course		1	Fidelity, Context and Moderators
		Delivery Support Hub		1, 4	
	Semi-structured interviews	Case study practitioners		1, 2, 3, 4, 5, 6 ^c	Fidelity, Responsiveness, Context and Moderators, Perceived impact, Programme differentiation
Quantitative	Online surveys	Baseline BAU survey—all nominated NELI Preschool Leads in control and intervention settings	Descriptive statistics	6	Programme differentiation
		Endline BAU survey—nominated NELI Preschool Leads in control settings			
		Endline NELI Preschool Lead survey—all NELI Preschool Leads in intervention settings		1, 2, 3, 4, 5, 6	Fidelity, Responsiveness, Context and Moderators, Perceived impact, Programme differentiation
		Baseline and Endline Practitioner Outcomes surveys—all nominated practitioners in intervention and control settings	Descriptive statistics and pre-post comparison between intervention and control	1, 5, 6	Fidelity, Perceived impact, Programme differentiation
	Training completion data	All intervention practitioners	Descriptive statistics	1, 4	Fidelity, Context and Moderators
	Webinar attendance data	All webinars			
	SDLs	All intervention children		1, 2, 3, 4	Fidelity, Responsiveness, Context and Moderators
	Targeted group selection data (including setting attendance patterns) ^d	All children			
	Case study cost pro formas ^e	Case study settings		1	Fidelity

^a It was anticipated at protocol stage that the webinar observations would offer insight into Context and Moderators through input from the settings. However, as the webinars were not interactive, this data source was ultimately not used to address RQ4.

^b It was anticipated at protocol stage that the case study session observations would inform our understanding of perceived impact. However, as it was not possible to do any kind of pre-post assessment of child or practitioner behaviour, this data source ultimately did not feed into our response to RQ5.

^c It was not anticipated at protocol stage that the semi-structured interviews would inform the findings for IPE RQ6. However, information was collected in the interviews about how the programme compared to usual practice, which provided useful data for responding to IPE RQ6.

^d It was anticipated at protocol stage that of the data provided by settings to enable the NFER team to select the S1 sample, only the children's attendance patterns would be used. However, it became apparent that other variables including the ability to access small-group work was also relevant for the IPE analysis. Furthermore, this information became relevant for answering RQ2, as the attendance patterns of disadvantaged children presented challenges for them accessing the programme. This data was also relevant for answering RQ3, as it provided information on the factors that resulted in children who may have benefitted from the targeted group being excluded from it.

^e Initially anticipated to inform only the cost evaluation, data from the cost pro formas was ultimately used for some of the IPE analysis as well.

Data collection

The intended and actual sample for each of the research methods is presented in Table 10. More details about the data collection process, including reasons for the actual sample not aligning with the intended sample, are presented below in chronological order of the data collection activities.

Table 10: Sample size for each data collection activity

Data collection activity	Intended sample	Actual sample
Reviews of the training course and Delivery Support Hub	NA	NA
Webinar observations	Two (out of four)	As intended
Case studies	12 settings—range of setting types, locations, sizes, and local deprivation levels (using the IDACI) 12 whole-group session observations 12 small-group session observations 12 process quality observations 12–24 practitioner interviews 12 cost pro formas	Setting sample as intended, although some variation in the data collected: 12 whole-group session observations 10 small-group session observations 2 individual session observations 12 process quality observations - Interviews with 22 practitioners from 12 settings 7 cost pro formas
Baseline BAU survey	303 settings	284 settings (94%)—140 intervention, 144 control
Endline BAU survey	151 control settings	123 control settings (81%)
NELI Preschool Lead survey	152 NELI Preschool Leads in intervention settings	109 NELI Preschool Leads (72%)
Practitioner Outcomes survey	303 settings, all practitioners nominated to complete the training	Baseline: 292 settings (96%); 146 intervention, 146 control (642 practitioners) Endline: 237 settings (78%); 121 intervention, 116 control (380 practitioners) Completed at both timepoints: 230 settings (76%); 113 intervention, 117 control (365 practitioners)
Training completion data	152 intervention settings	152 intervention settings (100%)
Webinar attendance data	All webinars	All webinars
SDLs	152 intervention settings	122 intervention settings (80%)
Targeted group selection data	All children who completed baseline testing (6,614)	6,614 children (100%) Data was incomplete for 4 of these children

Semi-structured review of the online training course

An NFER researcher completed a semi-structured review of the online training course in January 2025, using a framework developed with reference to the programme description and ToC. The researcher took notes relating to the format and content of each module, as well as the number and nature of any interactive activities, and the number and nature of any comments or questions from participants. The review was completed in parallel to training completion by the practitioners and informed development of the instruments for subsequent data collection activities—particularly for the webinar and session observations.

Semi-structured observations of webinars

An NFER researcher attended and observed two of the four webinars that OxEd delivered to settings over the course of the delivery period, in November 2024 and March 2025. The selected webinars covered LanguageScreen, the Narrative Progress Tracker, Targeted Group Selection, and reviewing implementation and child progress midway through the delivery period. The observation included a framework of the key programme elements based on the TIDieR (Template for Intervention Description and Replication) and ToC. The researcher took notes on the format of the session, the level of practitioner participation, the key points covered in relation to each of the programme elements, and any questions or feedback from participants. Particular attention was paid to how the webinars complemented the online training material in terms of both additional support and consistent messaging.

It was planned at protocol stage to look at levels of practitioner engagement in the webinars and the nature of the questions they asked. However, during the trial period, it became clear that the format of the webinars did not allow for any practitioner interaction and facilitated active participation only through a small number of polls. As a result, practitioner engagement was judged to no longer be a relevant area of focus, and the webinar observations constituted a less significant data source than initially anticipated. While this reduced the number and range of practitioners from whom we were able to collect qualitative data, the combination of the high survey response rates and the detailed case study data mitigated against any significant data gaps emerging as a result.

BAU surveys

Two online BAU surveys were implemented: one in September 2024 (baseline) to inform our understanding of BAU in all settings prior to programme delivery, and one in June–July 2025 (endline) to understand what BAU looked like in control settings over the trial period. The nominated NELI Preschool Lead at all participating settings was asked to complete the baseline survey, but only those in control settings were asked to complete it at endline. Intervention settings provided information about similar activities during the trial period via the NELI Preschool Lead survey (see below).

The BAU surveys asked about practice around oral language development in the setting in terms of formalised screening, programmes, and/or interventions as well as part of continuous provision. The surveys also asked about any related training participating practitioners may have completed in the past. The information from these surveys was to help the NFER team understand the extent to which the programme differs to standard practice in settings, and whether this varied by setting type, as well as the extent to which the settings asked *not* to deliver the intervention could legitimately be understood to represent a ‘control’. This understanding subsequently informed our interpretation of the findings from the impact analysis.

Response rates were high at baseline (94%) and slightly lower at endline (81%), but still within the data collection target. Completion of the endline survey was not a requirement for control settings to receive their incentive payment.

All surveys for the trial were hosted via Questback. Each respondent received a unique survey link wherever possible. Survey routing meant only relevant questions were shown to each respondent so as to minimise completion burden.

Practitioner outcomes survey

All practitioners nominated to complete the training were invited to complete a survey in September 2024 (baseline). Those who completed the survey, across both intervention and control settings, were invited to complete a similar survey in June–July 2025 (endline).¹⁸ These surveys asked practitioners about their confidence in their knowledge and abilities around supporting early language development, as well as their motivation to engage in certain practices known to support it.

¹⁸ Settings where no practitioner completed the survey at baseline were still invited to complete the survey at endline.

Practitioner outcomes are a key component of the ToC, as the programme is conceived to achieve broader and long-term impacts by acting as CPD for transferable skills around supporting early language development.

The confidence questions were adapted from the second scale of the Early Math Beliefs and Confidence Survey (EM-BCS), which asks about teachers' confidence in helping preschool children (aged 3–4) learn maths (Chen and McCray, 2013). Please see Appendix J for these questions. The motivation questions were based on the Work Tasks Motivation Scale for Teachers (WTMST) (Fernet *et al.*, 2008), looking at both the degree and type (extrinsic or intrinsic) of motivation. Please see the Outcome measures section for further details on Practitioner confidence.

At endline, all practitioners were also asked about their level of experience with the training and/or delivery of NELI Preschool, whether as part of or outside of the trial, as well as any prior experience with NELI for Reception. This was to inform our understanding of intervention practitioner participation in the programme, as well as to gain an understanding of any contamination present among control practitioners.

There was a very high response rate of practitioners from 292 settings (96%) at baseline. At the practitioner level, 51% were from settings subsequently randomised to control and 49% from settings randomised to intervention. There was a 50:50 split between baseline and endline at setting level. At endline, responses were received from 237 settings (78%), with 51% of these from control settings and 49% from intervention settings at both setting and practitioner level.

Case studies

Case study activities were successfully completed with 12 settings at staggered intervals between February and June 2025. Settings were categorised as 'PVI' or 'maintained', 'large' or 'small'¹⁹, and 'high', 'medium', or 'low' deprivation²⁰. This enabled all settings to be classified as one of the twelve categories below (Table 11). One setting was then randomly selected from each category and invited to participate as a case study setting. In cases where a setting declined or did not respond, another setting from the same category was randomly selected as a replacement. A total of 35 settings were contacted with the request to act as a case study.

Table 11 Categories for case study sampling

Category	Type	Setting size	Local deprivation levels
1	PVI	Large	High
2	PVI	Small	High
3	PVI	Large	Medium
4	PVI	Small	Medium
5	PVI	Large	Low
6	PVI	Small	Low
7	Maintained	Large	High
8	Maintained	Small	High
9	Maintained	Large	Medium
10	Maintained	Small	Medium

¹⁹ Based on the number of eligible children reported by the setting at baseline. 'Big' and 'small' settings were defined by the median number of 23 children per setting. This approach was taken with the aim of covering the range of setting sizes represented in the trial sample itself.

²⁰ Levels of deprivation determined by the IDACI decile of settings' postcode: 'high' = 1–3, 'medium' = 4–6, 'low' = 7–10.

11	Maintained	Large	Low
12	Maintained	Small	Low

Each case study setting received a visit from an NFER researcher, who observed one whole-group session and one small-group session. In two cases, an individual session was observed instead of a small-group session due to settings not scheduling a small-group session for the day of the visit. While no individual session observations had been planned due to concerns that it would be overly disruptive to both the child and practitioner, in these cases the observer took care to sit out of the child's line of sight and sufficiently far away to prevent them from feeling pressured. These observations were consequently reported to be no more disruptive than those of the small-group sessions.

The researcher used a semi-structured observation tool to take notes on:

- Context—including day and time, location and environment, session duration, and number of children
- Fidelity—in relation to use of the materials, nature of language support provided, and session structure and content
- Child engagement²¹—including for children eligible for EYPP, children with EAL, and children in the targeted group specifically, as well as measures to support it²².

As NELI Preschool was designed as a fully independent programme from NELI for Reception, it was important that a new observation tool was developed tailored to the specifics of the NELI Preschool programme. The tool was developed based on the programme materials and ToC to ensure the core components of fidelity were captured. Researchers focused on gathering qualitative descriptions of what took place in these sessions and how the children responded. The tool was initially developed as part of the formative evaluation, which means it had already been piloted ahead of its use in the main trial (see the NELI Preschool Formative Evaluation Report in a separate Appendix).

Following the observations, one to two practitioners per setting participated in an in-person or remote (via Microsoft Teams) interview. In most cases, the practitioners were those who had delivered the sessions observed. The practitioners often chose to participate in a joint interview, but some were interviewed separately—primarily for logistical reasons. Interviews lasted between 15 minutes and an hour depending on the mode and number of participants. Interviews were recorded wherever possible. Each interview covered:

- Background information—staff roles, involvement in the programme, and number of children in the sessions
- Training and support—including their experience with and perceptions of the online training course, webinars, Delivery Support Hub, and the support provided by the OxEd team
- Selecting children for the targeted intervention—including their experience with and perceptions of using LanguageScreen, process for selecting the children, and perceptions of the suitability of the children selected
- Implementation—including delivery logistics, child participation in the sessions, adaptations, communication with families, and challenges and facilitators

²¹ This refers to the physical act of being present at the session (not playing elsewhere), and the extent to which the child appears to be listening (not distracted or talking on unrelated topics) and actively contributing (responding to questions, raising their hand, etc.).

²² The names of these children were identified in advance using the child data shared by each setting, however it was not always possible to identify the children corresponding to the names in practice.

- Perceived impact—for children (including disadvantaged groups) and practitioners, as well as any drawbacks.

The interviews also picked up on any points of interest that emerged from the observations.

As with the observation tool, the interview schedule was piloted and refined as part of the formative evaluation.

Finally, the researcher carried out a 5- to 30-minute observation of continuous provision in each case study setting to gauge process quality. Qualitative observations were noted in relation to three relevant subscales from the Sustained Shared Thinking and Emotional Wellbeing (SSTEW) scale: (1) building trust, confidence, and self-regulation, (2) supporting and extending language and communication, and (3) supporting learning and critical thinking (Howard *et al.*, 2020).²³

In addition to the data collection visit, case study settings were asked to complete a cost pro forma (see Data collection under Costs below).

The case study visits were completed by two trained and DBS-checked IPE researchers. Case study settings received a payment of £100, and practitioners who participated in the interviews personally received a voucher worth £50.

Endline NELI Preschool Lead survey

All the NELI Preschool Leads in intervention settings were invited to complete an online survey in June–July 2025 about their experience of implementing the programme in their setting, including their perceptions of the online training, support and materials, targeted selection, delivery challenges and facilitators, session length, child engagement, adaptations, integration into continuous provision, and perceived impact for children and for practitioners, as well as any intentions to continue with the programme, and information about any other programmes delivered or CPD completed during the trial period. The purpose of this survey was to provide a better sense of the prevalence of different experiences, including those that emerged through the case study interviews.

The response rate was lower than for the BAU surveys (71%).

Semi-structured review of Delivery Support Hub activity

An NFER researcher carried out a semi-structured review of the Delivery Support Hub in July 2025, once the delivery period was complete. This allowed the researcher to gauge levels of participation in the community of practice. The researcher took note both of the quantity of engagement (for example, number of posts) and the nature of this input in relation to various key components of the programme.

The purpose of this data collection activity was to inform understanding of fidelity to programme intent, as well as the potential impact of moderators to programme implementation, as the level of engagement with the community of practice was identified in the IDEA workshop as a potential moderator of the fidelity and quality of delivery.

Training course completion data

The OxEd team shared with NFER the number of practitioners who registered to complete the online training course in each setting, as well as the number of practitioners who completed it—both before the start of delivery and by the end of the delivery period. This data then informed the fidelity and compliance analysis. For a practitioner to have completed the

²³ In usual practice, the SSTEW scale is completed as an assessment, using ratings against a number of indicators, on the basis of an multi-hour observation of the setting.

training, they must have marked each page of the online course as ‘complete’ and passed both of the in-course assessments.

Webinar attendance

The OxEd team shared with NFER attendance numbers for each of the four webinars, as well as the number of views of the webinar recordings. This data informed our understanding of fidelity and levels of engagement with this programme element.

Session attendance

NFER asked all intervention settings to complete an online SDL over the session delivery period. The log was in the form of an Excel spreadsheet with space to record the date for each session delivered and which children attended it, for all three session types. The Logs were pre-populated with the names of the children participating in the trial and shared with settings using NFER’s secure online portal, with each spreadsheet covering five weeks of delivery. The settings were asked to return the completed template every five weeks using the same portal and download the next template for completion over the following five weeks. Data from the SDLs was used to inform the fidelity (and compliance) analysis.

Targeted group selection data

As part of the initial child data collection, NFER asked settings to provide the necessary information for whether each child would be eligible for the targeted intervention. This covered attendance hours, whether they were able to work in small groups (for example, no complex SEN), and whether it would be possible to schedule three small-group sessions per week with that child, given their attendance patterns. This data was used to understand whether attendance patterns and/or other factors disproportionately excluded certain children, particularly disadvantaged children, from receiving the targeted intervention.

Analysis

Qualitative data—observations, interviews, and reviews

A detailed qualitative analysis plan was developed by the IPE Lead in advance. Interview and observation notes were written up as intelligent verbatim transcripts²⁴ and uploaded to the qualitative data analysis software MAXQDA. The data was then analysed using thematic analysis to identify and interpret patterns or themes (Naeem *et al.*, 2023). High-level deductive coding (approaching data with a pre-established framework for interpreting it) was used to sort the data into relevant themes. Detailed inductive coding (identifying patterns of meaning present in the data) was then used to draw out the key findings under each of these themes. The two IPE researchers who conducted the data collection also completed the analysis. Each researcher started by independently coding one interview, then met to discuss their approach and develop greater inter-coder reliability. The researchers then met at regular intervals to sense-check their interpretations of the data. All deductive coding was carried out based on a coding frame that was developed in advance based on the programme ToC and IPE research questions to ensure relevance and minimise bias.

Quantitative analysis—surveys and other quantitative data

A detailed quantitative analysis plan was developed by the IPE Lead in advance, in consultation with the Project Lead, Project Director, and Project Statistician. The plan outlined both the descriptive and inferential analyses required to answer the IPE research questions. Inferential statistics were used to understand the results of the confidence and motivation

²⁴ Intelligent verbatim transcription excludes fillers and redundancies that do not add meaning to the content to make the text more ‘readable’ (McMullin, 2021).

measures in the practitioner outcomes surveys, but otherwise, only descriptive statistics were used, with a focus on frequencies and/or means. Cross-tabulations in relation to group allocation, setting type, and/or local deprivation levels were calculated for selected variables in the survey, session delivery log, and training completion data, where this was deemed likely to provide further insight into settings' experiences. Means, maximums, and minimums were primarily used for analysing the setting cost pro forma data. The sample size was too small for cross-tabulations in this case.

All quantitative analysis was conducted using SPSS.²⁵ All syntax, as well as the outputs, were reviewed and checked following NFER's analysis quality assurance procedure. Survey data was downloaded from Questback, and the data quality was assured via NFER's data quality assurance (DQA) process prior to its analysis.

Survey data on self-reported practitioner outcomes (i.e., practitioner confidence) was statistically tested for differences between a small number of single items at baseline and endline (selected based on their relevance to the ToC), comparing intervention and control groups using Pearson chi-squared tests, and conducted at the practitioner level. This allowed us to determine whether the change from baseline to endline for practitioners in intervention settings differed significantly from the same change for practitioners in control settings.

Triangulation of qualitative and quantitative data

The IPE Lead developed an analysis framework mapping out how each of the data sources fed into analysis and reporting for each of the IPE research questions. Findings from each of these data sources were then examined in tandem to ensure their integration when responding to each research question in the final report. The IPE researchers collated and triangulated all data sources through an analysis workshop to ensure they were providing a comprehensive assessment of the implementation effectiveness and perceived outcomes of the NELI Preschool programme. An emerging findings workshop with the full team then allowed for the IPE and impact evaluation results to be brought together in a context where the team could reflect on how the IPE findings might inform our understanding and interpretation of the impact evaluation results.

Costs

Data collection

Data collection for the cost evaluation was embedded within the IPE activities to minimise the burden on settings while enabling triangulation from various sources. The surveys and SDLs were used to collect information on the number and average length of each of the session types delivered, as well as time spent on the training and number of webinars attended. Data from the case study interviews provided additional context to the cost data collected.

Case study settings were asked to complete a detailed cost pro forma to provide information about the amount of time each participating practitioner spent on the training, webinars, Delivery Support Hub, screening, set-up, session preparation, and session delivery, as well as any additional costs and the nature of these. Practitioners were also asked to indicate the scenario in which they had completed the majority of the training—whether this was during their normal working hours or not, and whether or not the time was paid. This information fed into both the IPE and costs analysis (see below). A blank copy of the cost pro forma was shared with settings as an Excel spreadsheet via email and then via a secure link. Only 7 of the 12 case study settings returned a cost pro forma, despite several reminder emails. This resulted in a smaller dataset than intended, which may have reduced the reliability and/or representativeness of the cost evaluation calculations that relied on this data. As with the endline survey, this is likely due to competing priorities and data collection 'fatigue', as well as the fact that it was likely more time-consuming for settings to collate the information requested compared to the surveys for example.

²⁵ It was anticipated at protocol stage that this analysis would be done in R. This change has been made due to a change in the team, bringing in different expertise.

To minimise the burden on settings, data for the cost evaluation was collected directly from the OxEEd team wherever possible, via a cost pro forma. This was the case for the length of webinars, the number of settings that required a tablet and/or screen for the trial, and the market price of the programme.

Analysis

We conducted the cost evaluation analysis in line with the EEF's latest cost evaluation guidance (EEF, 2023b). When calculating the staff time spent on the programme, we used a range of sources and assumptions, as outlined in Table 12. As the number of staff members involved in each activity differed per setting, we have calculated the mean number of staff, as well as the min and max involved in each activity, based on the sampled settings. However, because of the variation in staffing arrangements, time spent on each activity is presented in the findings at setting level rather than practitioner level.

Case study settings were requested to include all staff involved in the training and/or programme delivery in the cost pro forma. However, it is possible that not all relevant staff were represented for each setting, which could result in the overall time for some activities representing an underestimate. This is particularly likely in the case of the LanguageScreen administration, as this did not need to be completed by someone otherwise involved in programme delivery. Details of the setting cost pro forma can be found in Appendix L in the further appendices.

Table 12: Sources, methods, and assumptions for calculating staff time spent on the programme

Activity	What is being measured	Source(s)		Notes (including assumptions)
		Number of staff involved per setting	Time spent per setting	
Online training course	The amount of time spent on the training course per setting.	OxEd training completion data	Case study cost pro formas	The average (mean) time spent on the training per setting was calculated by adding up the time each practitioner reported spending on it and dividing this figure by the number of settings for which we had this data. Given the high rates of turnover per setting and the aim for as many practitioners as possible in the setting to complete the training, we have assumed that one additional staff member will need to complete the training course in Years 2 and 3 of delivery. As an average of 1 in 7 PVI staff members and 1 in 14 maintained setting staff leave their setting each year (DfE, 2025a), accounting for an additional one staff member trained each year is still a conservative estimate.
Webinars	The amount of time spent attending webinars per setting.	Case study cost pro formas	Number of webinars attended: Case study cost pro formas Average length of a webinar: OxEd cost pro forma	The average (mean) time spent is the average length of a webinar multiplied by the mean number of webinars attended at setting level. This meant that where two practitioners attended the same webinar, this was counted as two webinars. Where a practitioner reported attending a webinar, we have assumed that they attended for the whole thing. The minimum time spent is the lower value resulting from the following calculations: • Average number of webinars attended multiplied by the minimum reported webinar length • Minimum number of webinars attended multiplied by the average reported webinar length. The equivalent calculation was made for the maximum value. We have assumed that no staff members would attend the webinars in Years 2 and 3, as the webinars each year would cover the same content, and they would usually already have been attended by at least one practitioner in the setting.
Delivery Support Hub	The amount of time spent engaging with the Delivery Support Hub per setting.	Case study cost pro formas		The average (mean) time spent on the Delivery Support Hub per setting was calculated by adding up the time each practitioner reported spending on it and dividing this figure by the number of settings for which we had this data. We have assumed that equivalent practice will continue across Years 2 and 3 of delivery.
LanguageScreen	The amount of time spent administering LanguageScreen per setting.	Case study cost pro formas		The average (mean) time spent on LanguageScreen administration per setting was calculated by adding up the time each practitioner reported spending on it and dividing this figure by the number of settings for which we had this data. Case study settings were asked to include all staff involved in the programme training and/or delivery in the cost pro formas. However, it is possible that additional staff members were involved in administering LanguageScreen without otherwise being involved in the programme, which could result in them not being represented in the data. One case study setting for example, indicated zero time spent on this activity among the practitioners listed, suggesting another staff member not involved in training or delivery of the programme completed this. This setting has been excluded from the analysis of time spent on LanguageScreen, but it is

Activity	What is being measured	Source(s)		Notes (including assumptions)
		Number of staff involved per setting	Time spent per setting	
				possible that additional time was spent on this activity in other settings as well by practitioners not recorded in the cost pro forma. Another setting did not make it clear how many practitioners were involved in administering LanguageScreen. We have made assumptions based on the interview data in this case.
Set-up	The amount of time spent on programme set-up per setting.	Case study cost pro formas		The average (mean) time spent on programme set-up per setting was calculated by adding up the time each practitioner reported spending on it and dividing this figure by the number of settings for which we had this data. The mean assumes 5.5 hours for the one respondent that selected 'More than 5 hours', as the scale went up in half-hour increments, and it is unlikely that they spent significantly more time on this than their peers. However, it is still possible that the estimated mean and maximum times spent on set-up represent slight underestimates of the 'true' values.
Session preparation time	The total amount of time spent on session preparation across the whole delivery period per setting.	Case study cost pro formas	Time spent per week: Case study cost pro formas Number of weeks of delivery: SDLs	The total average (mean) time spent on session preparation per setting was calculated by adding up the time each practitioner reported spending on it per week and dividing this figure by the number of settings for which we had this data, then multiplying this by the average number of weeks of delivery. Only those settings that returned all of their SDLs were included in this analysis. Minimum time spent is the lower value resulting from the following calculations: • Average amount of time spent per week multiplied by the minimum number of weeks of delivery • Minimum amount of time spent per week multiplied by the average number of weeks of delivery. The equivalent calculation was made for the maximum value.
Time for session delivery	The total amount of time spent on session delivery across the whole delivery period per setting.	Case study cost pro formas	Total number of sessions delivered per setting: SDLs Length of sessions delivered: Endline NELI Preschool Lead Survey	The average (mean) amount of time spent on session delivery per setting was calculated by multiplying the mean number of each type of session delivered per setting by the mean length of each session type. Only those settings that returned all of their SDLs were included in this analysis. Minimum time spent is the lower value resulting from the following calculations: • Mean session length multiplied by the minimum approximate number of each type of session delivered per setting • Minimum session length multiplied by the average approximate number of each type of session delivered per setting. The equivalent calculation was made for the maximum value.

We also calculated the average (mean) cost of programme delivery per setting per year, over a projected three-year period. Having established a cost per setting per year, we divided this figure by the average number of children per setting (according to baseline data from the trial) to estimate the cost per child per year. We have only included in the primary cost analysis any costs that were reported by at least half of the settings for which the relevant data was available. However, we carried out a sensitivity analysis for any other costs reported by settings, including the prerequisite of the screen and optional resources such as a TV trolley and wireless TV keyboard. Sources, methods, and assumptions for these calculations are outlined in Table 13.

Table 13: Sources, methods, and assumptions for calculating the cost of the programme

Nature of cost	Source(s)	Methods and assumptions
Additional staff time for completing the training	Case study cost pro formas	We calculated the average (mean) number of hours of training spent on the training per setting. We then multiplied this number by the median hourly pay of early years workers in 2023/2024 (£12.63) (Flemons and Worth, 2025). For Years 2 and 3, we assumed that one additional practitioner would complete the training per setting. To account for this, we calculated the mean amount of time spent on the training per practitioner and multiplied this by the median hourly pay (see above), adding this amount to the Year 1 total for Year 2 and to the Year 2 total for Year 3.
Tablet	OxEd delivery team	All settings, including those that already had a tablet for the trial, were included in the calculations. We have assumed one tablet per setting, as this is what was required for trial participation. The price per tablet used for these calculations was the higher of the two prices OxEd reported for the tablets purchased as part of the trial.
Programme materials	OxEd delivery team	NA
Annual digital resources subscription	OxEd delivery team	NA
Other—extra copies of the texts, storage for the materials, additional props	Case study cost pro formas	Mean calculated from all sampled settings, including those that did not report any of these costs.
Other—printing, photocopying, internet	Case study cost pro formas	Mean calculated from all sampled settings, including those that did not report any of these costs.
Screen, TV trolley, and wireless TV keyboard	OxEd delivery team	These calculations use the prices of those items purchased by OxEd to provide to settings as part of this trial.

Timeline

Table 14: Timeline of activities for pre-trial delivery and the formative evaluation

Dates	Activity (staff responsible / leading)
Jan–Feb 2024	Recruitment of PVI settings (OxEd)
Feb–March 2024	Completion of training (Settings) Administration of LanguageScreen (Settings) LanguageScreen data shared with NFER (OxEd) IPE sample selection and onboarding (NFER) Pre-populated SDL shared with IPE settings (NFER) Child attendance patterns and targeted group selection collected from settings (NFER) NELI Preschool Programme delivery begins (staggered) (Settings)
Apr–Jun 2024	IPE activities (interviews and observations) (NFER) SDLs returned at 5-week intervals (Settings) Informal formative feedback via email and meetings (NFER)
Jul–Aug 2024	Analysis and reporting (NFER and RAND) Presentation of findings to the EEF and OxEd (NFER and RAND) Summary of findings and financial incentives shared with participating settings (NFER and OxEd)

Table 15: Timeline of activities for the effectiveness trial

Dates	Activity (organisations responsible/leading)
August 2023–February 2024	Project set-up (NFER, OxEd, and EEF) Develop and finalise recruitment documents (NFER, OxEd, and EEF)
January 2024–July 2024	Setting recruitment via EoI and MoUs (OxEd) Contact settings with instructions for baseline data collection (NFER)
September 2024	Baseline data collection activities (NFER) - Phase 1 baseline data collection—BaU survey and Practitioner Outcomes survey - Distribute parent information sheet and withdrawal letters via settings - Phase 2 baseline data collection—Child Data - Upload child data on LanguageScreen and share QR code with settings
October–November 2024	- LanguageScreen Assessment completed in settings (administration by practitioners) - Setting randomisation and share group allocation with settings upon LanguageScreen completion (NFER) - Intervention settings receive baseline LanguageScreen data (NFER) - Intervention settings receive Targeted Group Selection Guidance (OxEd) - Selection of targeted intervention children (intervention settings)
Mid-November–December 2024	NELI Preschool Training in settings (OxEd and settings)
January 2025	- NELI Preschool Programme delivery begins (settings with support from OxEd) - Intervention and control settings receive payment to cover staff time for completing baseline LanguageScreen assessments (OxEd) - Intervention settings receive payment to cover staff time to complete NELI Preschool online training (OxEd)
January–June 2025	- NELI Preschool Programme delivery (settings with support from OxEd) - Weekly NELI Preschool SDLs (Settings shared these with NFER every five weeks) - Selection of case study sample and ongoing IPE activities (NFER)
June–July 2025	- NELI Preschool Lead survey, BaU survey (control group), and Practitioner Outcomes survey (NFER shares these for settings to complete) - Endline assessments for LanguageScreen and RAPT (NFER Test Administrators)
September 2025	- LanguageScreen access to all settings (OxEd) - All settings receive payment for completing relevant trial activities and liaise with control settings on final payment or order NELI Preschool (OxEd)
September–January 2026	Analysis and report writing (NFER)
February 2026	First draft report (NFER)
February–July 2026	Report revisions and final publication of the main report (NFER and EEF)
April 2026	NPD application submitted for EYFSP (Longitudinal evaluation) (NFER)
May–July 2026	Light-touch IPE survey (Longitudinal evaluation) (NFER)
February 2027	First draft addendum report (NFER)
March–June 2027	Report revisions and final publication of the addendum report (NFER and EEF)
July–September 2027	Archive data in the EEF data archive (NFER)

Impact evaluation results

Participant flow including losses and exclusions

The participant flow is presented in Figure 3. As seen, OxEd directly contacted approximately 3,000 settings. In addition, OxEd's network-based recruitment approach meant that between 20,000 and 46,000 settings were exposed to NELI

Preschool trial recruitment in some way through cascading communications. Of these, 337 signed the Memorandum of Understanding. NFER invited these settings to take part in evaluation data collection (child data collection and baseline assessments). Of 337 settings, 27 settings did not complete baseline activities, and 7 more settings were ineligible to take part in the trial for other reasons (four settings had insufficient eligible children and three settings were taking part in other EEF trials). As a result, 303 settings were randomised with 7,393 children from these settings who were registered to complete baseline assessments. Of these, 46 children were withdrawn from the trial by their parents, and 733 children did not complete LanguageScreen assessments. The main reasons for not completing the assessments were absence during the assessment period, leaving the setting, or being unable to access or complete the full assessment. As completing baseline assessments was a prerequisite for taking part in the trial, these children were removed, resulting in 303 settings and 6,614 children at randomisation.

As mentioned above, children were selected to be followed up at endline, with different selection criteria depending on the analysis sample in question (see the section on Analysis sample selection in Participant selection). There were 4,390 children (2,320 from the intervention and 2,070 from the control group) from 303 settings selected or sampled to be followed up across the four analysis samples:

Table 16: Analysis samples as selected for endline testing

	Intervention	Control	Total
S1: Ideally Targeted Children (Primary Analysis sample)	1,036	993	2,029
S2: Practitioner Targeted Children (intervention settings only)	*1,015	993	1,015
S3: Enrichment-only Sample (Primary Analysis Sample)	938	902	1,840
S4: EYPP**	553 (173)	538 (175)	1,091 (348)

* The numbers presented for the S2 control group repeat those for the S1 control group.

** Numbers shown in brackets for S4 are those children who were ONLY in S4.

There were 348 S4 children that were not included in the other samples as either they were not amongst the lowest achieving scores at baseline, they were not randomly selected for S3, and/or they failed to meet at least one of the eligibility criteria for the targeted intervention (and were excluded from the ideally selected sample): 32 were deemed not being able to access small-group activities, 30 were not attending at least 15 hours per week, and for 60 children it was not feasible to schedule sessions.

There was a degree of overlap between samples (notably S1 and S2, and S1, S2, and S3 all with S4), which is why the totals do not sum to those shown in the participant flow diagram below.

Only children who completed both assessments were classified as having post-test data collected. Of the children selected for endline testing, 636 were lost to follow-up (341 intervention; 295 control), as they did not complete both of the endline measures. Of these, 68 children were from three control settings who did not respond to NFER communication to take part in endline assessments, and therefore, they could not be followed up. In total, 3,788 children completed the LanguageScreen assessment and 3,814 completed the RAPT assessment, with 3,754 completing both (1,979 intervention; 1,775 control).

A further 22 (7 intervention; 15 control) children who had post-test data collected were subsequently excluded from analysis due to concerns about their ability to engage fully with the LanguageScreen assessment. Therefore, across all four analysis samples, data from 3,732 children were analysed (1,972 intervention; 1,760 control). Of these, 3,298 children were included in the primary analysis (1,692 intervention; 1,606 control).

Figure 3: Participant flow diagram (2 arms)

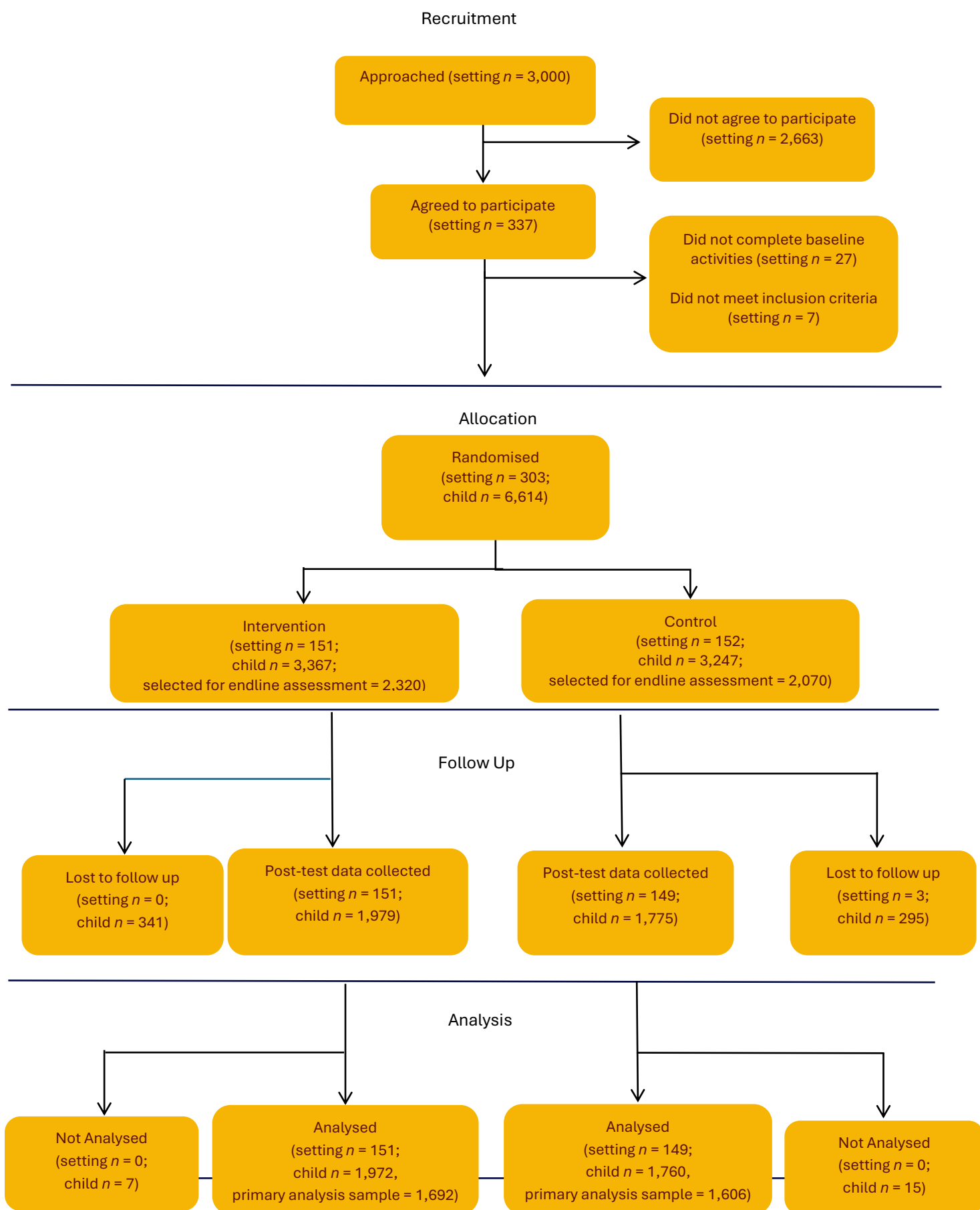


Table 17 summarises the minimum detectable effect sizes (MDES) at different stages of the trial. For both the full sample and the EYPP subgroup, the MDES is lower than originally expected, driven by several factors. Child-level attrition was just under 15%, substantially lower than the anticipated 23%, and setting-level attrition was also far lower than expected (<1% compared with the forecast 10%). In addition, the analysis incorporated a level 2 pre-post correlation that had not been assumed in earlier calculations. Together, these factors increased the effective sample size and therefore reduced the MDES.

For the EYPP subgroup, the achieved number of EYPP children at analysis stage (901) was lower than anticipated post-randomisation before assuming 23% child-level attrition (949), but higher than the comparable post-randomisation number after assuming attrition (731). It is also worth noting that 93 of the 300 settings had no EYPP-eligible 3 to 4 year olds. However, this was offset by a higher-than-expected average number of EYPP children in the remaining settings, combined with the lower-than-expected attrition at both setting and child levels. These elements collectively contributed to the lower MDES observed for the EYPP subgroup.

Table 17: Minimum detectable effect size at different stages

		Protocol		Randomisation		Analysis	
		Overall	EYPP	Overall	EYPP	Overall	EYPP
MDES*		0.163	0.200	0.165	0.200	0.110	0.179
Pre-test/post-test correlations	Level 1 (child)	0.81	0.81	0.81	0.81	0.77	0.75
	Level 2 (setting)					0.78	0.52
Intraclass correlations (ICCs)	Level 2 (setting)	0.21	0.21	0.21	0.21	0.22	0.17
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?		Two	Two	Two	Two	Two	Two
Average cluster size		12	2.3	13.7	3.1	11	4.4
Number of settings	Intervention	159	159	151	151	151	107
	Control	159	159	152	152	149	100
	Total:	318	318	303	303	300	207
Number of children**	Intervention	1,908	366	2,063	473	1,692	449
	Control	1,908	366	2,077	476	1,606	452
	Total:	3,816	732	4,140	949	3,298	901

* The MDES figures presented at the protocol and randomisation stage are adjusted for setting-level attrition of 10% and child-level attrition of 23%.

** The numbers of children presented at the protocol and randomisation stages are shown without attrition (but, as above, the MDES figures are adjusted for 10% setting-level attrition and 23% child-level attrition). The number of children presented at analysis stage are shown after attrition, which was considerably less than anticipated resulting in lower MDES despite other less favourable changes in pre-test/post-test correlations and ICCs.

Attrition

Overall, 658 children who were selected for endline testing across all four analysis samples were lost to follow-up for the reasons described in the Participant flow including losses and exclusions section above. This gives an overall attrition rate of 15%, with no difference between intervention and control, which is notably lower than the proportion assumed in earlier stages of the project (23%). Overall attrition from the primary analysis sample was similar (14.8%), though in this case a difference between intervention and control was observed, albeit relatively low (14.3% and 15.3%, respectively).

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

		Intervention	Control	Total
Number of children	Randomised and selected for EP testing	1,974	1,895	3,869
	Analysed	1,692	1,606	3,298
Child attrition (from randomisation to analysis)	Number	282	289	571
	Percentage	14.3%	15.3%	14.8%

Table 19: Number of children in each analysis sample

		Intervention	Control	Total
S1	Randomised and selected for EP testing	1,036	993	2,029
	Analysed	877	840	1,717
Child attrition (from randomisation to analysis)	Number (Percentage)	159 (15.3%)	153 (15.4%)	312 (15.4%)
S2*	Randomised and selected for EP testing	1,015	993	2,008
	Analysed	874	840	1,714
Child attrition (from randomisation to analysis)	Number (Percentage)	141 (13.9%)	153 (15.4%)	294 (14.6%)
S3	Randomised and selected for EP testing	938	902	1,840
	Analysed	815	766	1,581
Child attrition (from randomisation to analysis)	Number (Percentage)	123 (13.1%)	136 (15.1%)	259 (14.1%)
S4	Randomised and selected for EP testing	553	538	1,091
	Analysed	449	452	901
Child attrition (from randomisation to analysis)	Number (Percentage)	104 (18.8%)	86 (16.0%)	190 (17.4%)

* The numbers presented for the S2 control group repeat those for the S1 control group.

Pupil and school characteristics

Table 20 and Table 21 below set out the child and setting characteristics at randomisation and analysis, respectively. The numbers in the randomisation table (Table 20) are based on primary analysis sample (S1 and S3) for endline testing whilst the analysis table sets out the characteristics for the primary analysis sample that were subsequently analysed (i.e., not lost to follow-up) (Table 21). As the tables demonstrate, there was no evidence of baseline imbalance across any of the identified categorical characteristics. For each individual characteristic, a chi-squared test was carried out in order to determine whether any identified imbalances between randomisation and control groups were statistically significant. These tests showed that, where there are imbalances, these are not statistically significant. Furthermore, any differences between randomisation and analysis are minor and are not expected to have introduced any bias into the results. The numeric variables (proportion of EYPP-eligible children and age at baseline) are similarly balanced between control and intervention at both randomisation and analysis with negligible effect sizes. Based on this, no weighting has been applied when modelling the outcome.

Table 20: Baseline characteristics of groups as randomised and selected for endline testing

School-level (categorical)	Intervention group		Control group	
	n/N*	Percent	n/N*	Percent
Setting Type: Maintained	89/151	58.9%	89/152	58.6%
Setting Type: PVI	62/151	41.1%	63/152	41.4%
Setting Size: One class	110/151	72.8%	113/152	74.3%
Setting Size: More than one class	41/151	27.2%	39/152	25.7%
IDACI Decile: 1 (Most deprived)	13/151	8.6%	20/152	13.2%
IDACI Decile: 2	17/151	11.3%	16/152	10.5%
IDACI Decile: 3	22/151	14.6%	19/152	12.5%
IDACI Decile: 4	16/151	10.6%	15/152	9.9%
IDACI Decile: 5	12/151	7.9%	12/152	7.9%
IDACI Decile: 6	19/151	12.6%	19/152	12.5%
IDACI Decile: 7	14/151	9.3%	14/152	9.2%
IDACI Decile: 8	11/151	7.3%	14/152	9.2%
IDACI Decile: 9	12/151	7.9%	9/152	5.9%
IDACI Decile: 10 (Least deprived)	15/151	9.9%	14/152	9.2%
SPH: Bright Futures North West EYSPH	≤10/151	≤6.6%	≤10/152	≤6.6%
SPH: Derbyshire and Nottinghamshire EYSPH	16/151	10.6%	13/152	8.6%
SPH: EYE South West EYSPH	14/151	9.3%	15/152	9.9%
SPH: East of England EYSPH	16/151	10.6%	12/152	7.9%
SPH: Great North EYSPH	≤10/151	≤6.6%	12/152	7.9%

SPH: Kent EYSPH	19/151	12.6%	14/152	9.2%	
SPH: London South EYSPH	24/151	15.9%	24/152	15.8%	
SPH: Northern Lights EYSPH	≤10/151	≤6.6%	16/152	10.5%	
SPH: St Edmunds EYSPH, Yorkshire and Humber	≤10/151	≤6.6%	≤10/152	≤6.6%	
SPH: Thames Valley EYSPH	15/151	9.9%	22/152	14.4%	
SPH: Yorkshire and Humber Together EYSPH	15/151	9.9%	≤10/152	≤6.6%	
Ofsted: Outstanding	16/151	10.6%	13/152	8.6%	
Ofsted: Good	120/151	79.5%	122/152	80.3%	
Ofsted: Requires Improvement or Not Judged**	15/151	9.9%	17/152	11.2%	
Delivery Model: A	121/151	80.1%	123/152	80.9%	
Delivery Model: B	22/151	14.6%	16/152	10.5%	
Delivery Model: C	8/151	5.3%	13/152	8.6%	
Setting-level (continuous)	<i>n/N*</i>	Mean (SD)	<i>n/N*</i>	Mean (SD)	Effect size
Proportion of EYPP children***	151	16.5% (16.2%)	152	17.7% (19.2%)	-0.068
Child-level (categorical)	<i>n/N*</i>	Percent	<i>n/N*</i>	Percent	
Sex: Female	957/1,974	48.5%	954/1,895	50.3%	
Sex: Male	1,017/1,974	51.5%	941/1,895	49.7%	
EYPP: No	1,634/1,974	82.8%	1,532/1,895	80.8%	
EYPP: Yes	340/1,974	17.2%	363/1,895	19.2%	
EAL****: No	1,386/1,974	70.2%	1,354/1,895	71.5%	
EAL****: Yes	588/1,974	29.8%	541/1,895	28.5%	
Child-level (continuous)	<i>n/N*</i>	Mean (SD)	<i>n/N*</i>	Mean (SD)	Effect size
Age (months at baseline)	1,974	43.20 (3.63)	1,895	43.22 (3.59)	-0.004

* Missing numbers are not shown because, in some cases, doing so could risk potential disclosure. All proportions are calculated using the full denominator, including cases with missing data.

** Ofsted ratings Requires improvement and Not Judged have been combined due to low cell counts. The split between the two categories is consistent across both intervention and control groups.

*** Around 1 in 7 children in nursery registered for the universal entitlement were in receipt of EYPP (DfE, 2025c).

**** EAL is a practitioner-perceived measure.

Table 21: Baseline characteristics of groups as analysed (primary analysis sample)

School-level (categorical)		Intervention group		Control group		
		<i>n/N*</i>	Percent	<i>n/N*</i>	Percent	

Setting Type: Maintained		89/151	58.9%	87/149	58.4%	
Setting Type: PVI		62/151	41.1%	62/149	41.6%	
Setting Size: One class		114/151	75.5%	112/149	75.2%	
Setting Size: More than one class		37/151	24.5%	37/149	24.8%	
IDACI Decile: 1 (Most deprived)		13/151	8.6%	20/149	13.4%	
IDACI Decile: 2		17/151	11.3%	15/149	10.1%	
IDACI Decile: 3		22/151	14.6%	19/149	12.8%	
IDACI Decile: 4		16/151	10.6%	15/149	10.1%	
IDACI Decile: 5		12/151	7.9%	12/149	8.1%	
IDACI Decile: 6		19/151	12.6%	18/149	12.1%	
IDACI Decile: 7		14/151	9.3%	13/149	8.7%	
IDACI Decile: 8		11/151	7.3%	14/149	9.4%	
IDACI Decile: 9		12/151	7.9%	9/149	6.0%	
IDACI Decile: 10 (Least deprived)		15/151	9.9%	14/149	9.4%	
SPH: Bright Futures North West EYSPH		≤10/151	≤6.6%	≤10/149	≤6.7%	
SPH: Derbyshire and Nottinghamshire EYSPH		16/151	10.6%	13/149	8.7%	
SPH: EYE South West EYSPH		14/151	9.3%	15/149	10.1%	
SPH: East of England EYSPH		16/151	10.6%	12/149	8.1%	
SPH: Great North EYSPH		≤10/151	≤6.6%	≤10/149	≤6.7%	
SPH: Kent EYSPH		19/151	12.6%	14/149	9.4%	
SPH: London South EYSPH		24/151	15.9%	22/149	14.8%	
SPH: Northern Lights EYSPH		≤10/151	≤6.6%	16/149	10.7%	
SPH: St Edmunds EYSPH, Yorkshire and Humber		≤10/151	≤6.6%	≤10/149	≤6.7%	
SPH: Thames Valley EYSPH		15/151	9.9%	22/149	14.8%	
SPH: Yorkshire and Humber Together EYSPH		15/151	9.9%	≤10/148	≤6.7%	
Ofsted: Outstanding		16/151	10.6%	13/149	8.7%	
Ofsted: Good		120/151	79.5%	119/149	79.9%	

Ofsted: Requires Improvement or Not Judged**		15/151	9.9%	17/149	11.4%	
Delivery Model: A		121/151	80.1%	120/149	80.5%	
Delivery Model: B		22/151	14.6%	16/149	10.7%	
Delivery Model: C		8/151	5.3%	13/149	8.7%	
Setting-level (continuous)		<i>n/N*</i>	Mean (<i>SD</i>)	<i>n/N*</i>	Mean (<i>SD</i>)	Effect size
Proportion of EYPP children***		151	16.5% (16.2%)	149	17.7% (19.4%)	-0.07
Child-level (categorical)		<i>n/N*</i>	Percent	<i>n/N*</i>	Percent	
Sex: Female		826/1,692	48.8%	813/1,606	50.6%	
Sex: Male		866/1,692	51.2%	793/1,606	49.4%	
EYPP: No		1409/1,692	83.3%	1308/1,606	81.4%	
EYPP: Yes		283/1,692	16.7%	298/1,606	18.6%	
EAL****: No		1209/1,692	71.5%	1165/1,606	72.5%	
EAL****: Yes		483/1,692	28.5%	441/1,606	27.5%	
Child-level (continuous)		<i>n/N*</i>	Mean (<i>SD</i>)	<i>n/N*</i>	Mean (<i>SD</i>)	Effect size
Age (months at baseline)		1,692	43.3 (3.6)	1,606	43.3 (3.6)	-0.0003

* Missing numbers are not shown because, in a small number of cases, doing so could risk potential disclosure. All proportions are calculated using the full denominator, including cases with missing data.

** Ofsted ratings Requires improvement and Not Judged have been combined due to low cell counts. The split between the two categories is consistent across both intervention and control groups.

*** Around 1 in 7 (14.3%) children in nursery registered for the universal entitlement were in receipt of EYPP (DfE, 2025c).

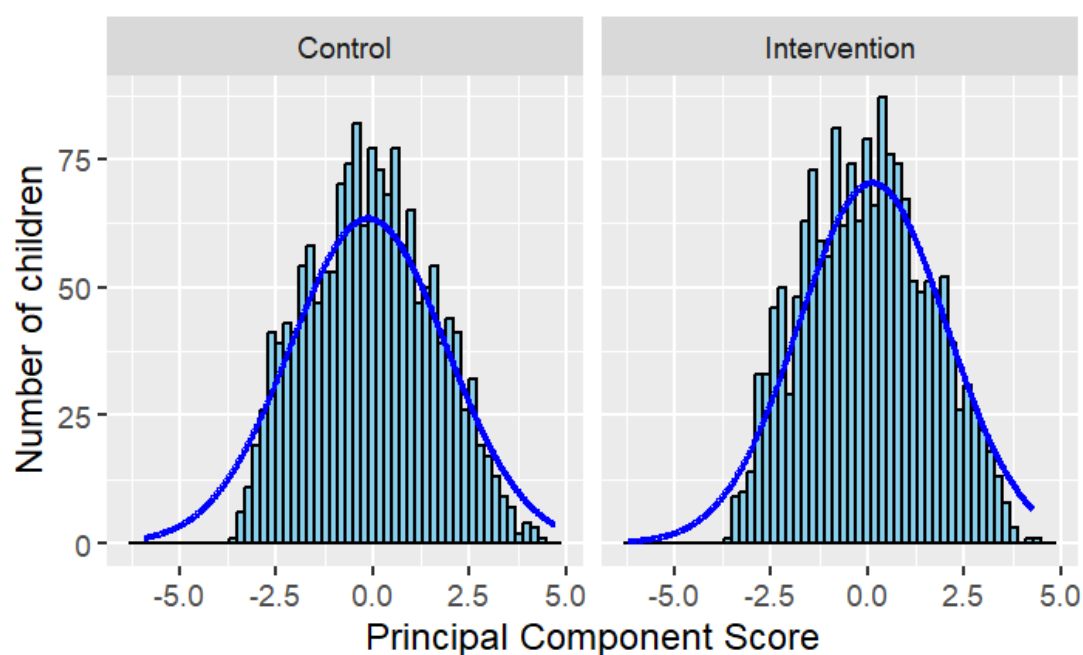
**** EAL is a practitioner-perceived measure.

Further to this, the effect size of the baseline latent oral language variable (-0.020 (-0.089, 0.045)) for the primary analysis sample suggests that there was negligible bias between the two groups immediately after randomisation, with the confidence interval spanning zero and indicating no meaningful difference at baseline. This is reflected in the distribution of the baseline latent oral language variable split by randomisation group (shown in Figure 4), which is very similar in both the intervention and control groups. The distribution of these scores closely approximates a normal distribution in both the intervention and control groups. This pattern suggests minimal baseline bias and provides confidence that applying a multilevel regression model in the primary analysis is appropriate, as the underlying normality assumption is not violated.

Table 22: Pre-test at baseline (primary analysis sample)

	<i>N</i> (intervention/control)	Hedges' <i>g</i> (95% CI)
Randomised and Selected for EP testing (primary analysis sample)	3,298 (1,602, 1,696)	-0.02 (-0.09, 0.05)

Figure 4: Histogram of baseline latent oral language variable at baseline, split by group (primary analysis sample)



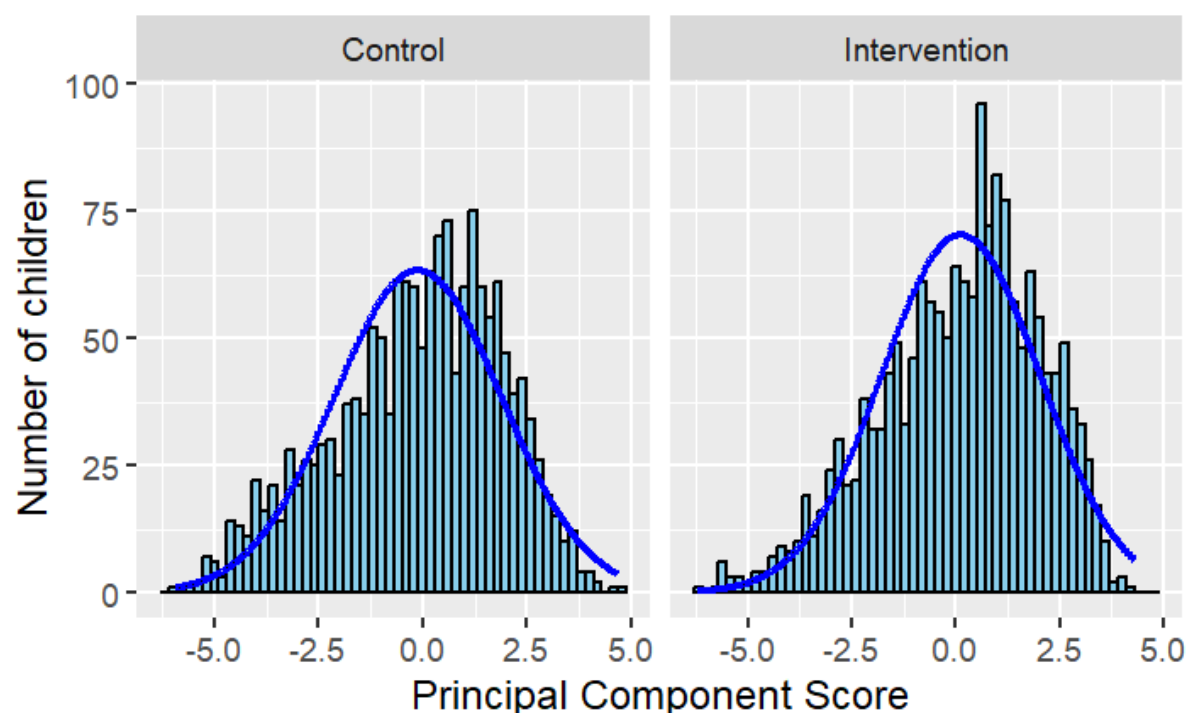
Outcomes and analysis

Primary analysis

RQ1: How effective is NELI Preschool at improving oral language skills of 3- to 4-year-old children in intervention settings compared to children in control group settings?

Figure 5 shows the distribution of latent oral language skills, derived from the first Principal Component of LanguageScreen and RAPT, split by randomisation group. The two groups appear broadly similar, and both approximately follow a normal distribution; however, the distribution for the intervention group is shifted slightly to the right, indicating that their scores tend to be higher overall which is reflected in the positive impact estimate set out below. This pattern suggests that the measure is functioning as expected and provides a suitable indicator of latent oral language skills. The clear separation between groups also supports the interpretability of subsequent analyses, as any observed differences in outcomes can be more confidently linked to meaningful variation in the underlying construct.

Figure 5: Distribution of latent oral language skills by Randomisation Group (primary analysis sample)



The results from the primary analysis are presented in Table 23. They indicate that the mean difference in latent language skills between children from the intervention group compared to those in the control group is 0.24 (intervention = 0.12 (0.03, 0.21); control = -0.12 (-0.22, -0.02)), corresponding to an effect size of 0.10 (0.02, 0.18) with a p -value of 0.01. The positive effect size, along with a confidence interval that is above zero, suggests that NELI Preschool is linked to improvements in children's oral language skills.

Table 23: Primary outcome analysis results

Outcome	Unadjusted means				Effect size			
	Intervention group		Control group		Total N (intervention; control)	ICC (full)	(95% CI)	p - value
	N (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)				
Oral Language Skills	1,692 (282)	0.12 (0.03, 0.21)	1,606 (289)	-0.12 (-0.22, -0.02)	3,298 (1,692, 1,606)	0.22	0.10 (0.02, 0.18)	0.01

Secondary analyses

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

A series of subscale analyses was conducted to establish whether NELI Preschool had a particular effect on any specific aspect of early language skill development. As shown in Table 24, NELI Preschool has a positive impact on all but one of the oral language skills subscales, particularly for Expressive Vocabulary (ES = 0.11 (0.05, 0.18)), Information (ES = 0.14 (0.05, 0.22)), and Grammar (ES = 0.15 (0.07, 0.24)). Listening Comprehension (ES = 0.10 (-0.00, 0.20)) and Receptive Vocabulary (ES = 0.05 (-0.03, 0.12)) both show small positive impacts, although the confidence intervals include zero, so the evidence does not strongly support a consistent positive effect. The effect size for Sentence Repetition is similarly low with a confidence interval that comfortably includes zero, suggesting that the intervention has no effect on this particular aspect of language. Overall, it seems that the intervention is most effective for expressive language skills (Expressive Vocabulary, RAPT Information, and RAPT Grammar) compared with receptive ones. While the ToC did not predict differences in impact

across expressive and receptive domains, these results indicate that the programme's active ingredients—especially its focus on dialogic reading, modelling, and structured opportunities for children to produce language—may be more strongly aligned with supporting expressive rather than receptive language development.

Table 24: RQ2 subscale analyses results

Outcome	Unadjusted means				Effect size			
	Intervention group		Control group					
	<i>N</i> (missing)	Mean (95% CI)	<i>n</i> (missing)	Mean (95% CI)	Total <i>N</i> (intervention; control)	ICC (full)	Hedges' <i>g</i> (95% CI)	<i>p</i> - value
Expressive Vocabulary	1,692 (282)	11.09 (10.88, 11.31)	1,606 (289)	10.61(10.38, 10.831)	3,298 (1,692, 1,606)	0.14	0.11 (0.05, 0.18)	<0.01
Listening Comprehension	1,692 (282)	96.86 (96.21, 97.52)	1,606 (289)	95.32 (94.66, 95.98)	3,298 (1,692, 1,606)	0.14	0.10 (-0.00, 0.20)	0.061
Receptive Vocabulary	1,692 (282)	14.95 (14.76, 15.13)	1,606 (289)	14.82 (14.63, 15.01)	3,298 (1,692, 1,606)	0.08	0.05 (-0.03, 0.12)	0.22
Sentence Repetition	1,692 (282)	7.87 (7.69, 8.05)	1,606 (289)	7.89 (7.69, 8.08)	3,298 (1,692, 1,606)	0.23	-0.04 (-0.14, 0.06)	0.46
RAPT Information	1,692 (282)	23.07 (22.72, 23.42)	1,606 (289)	21.89 (21.50, 22.28)	3,298 (1,692, 1,606)	0.14	0.14 (0.05, 0.22)	<0.01
RAPT Grammar	1,692 (282)	18.38 (18.04, 18.72)	1,606 (289)	17.19 (16.83, 17.56)	3,298 (1,692, 1,606)	0.13	0.15 (0.07, 0.24)	<0.01

Subgroup analyses

Additionally, a number of subgroup analyses were conducted, seeking to determine the effect of NELI on:

1. Those who were selected to receive the targeted component of the intervention versus those who should have been selected to receive the targeted component of the intervention (RQ3a and RQ3b)
2. Those who only received the enrichment component of NELI Preschool (RQ4)
3. The EYPP subgroup (RQ5)
4. Those who were attending PVI settings (RQ6a)
5. Those who were attending maintained settings (RQ6b)
6. Children with EAL (RQ7)

The results of each of these are set out in Table 25: Subgroup analysis results below. Overall, we find that effect sizes for targeted children are similar to, or slightly higher than, the overall effect size and that of the enrichment-only children. The effect sizes for both those who were selected for the targeted component and those who should have been selected for the targeted component are of a very similar positive size (0.12 and 0.14 respectively). The fact that the effect size for the ideally targeted children is similar to—and if anything, slightly higher than—the effect size for the children who were actually targeted suggests that practitioners' decisions to override the recommended targeting did not have any adverse effects.

This is because the only difference between the ideally targeted and actually targeted children is the subset of children who have been removed from the targeting. The fact that the effect size for the ideally targeted children is slightly higher than for the actually targeted children suggests that these children are doing slightly better than the remaining targeted children. These decisions, therefore, do not appear to have disadvantaged children who were removed from the originally suggested target group. By contrast, the effect size for the enrichment-only subgroup (0.10 (-0.01, 0.20)) is slightly lower than that for the targeted sample, though it is still positive (and, as is shown in the next section, there was no interaction), indicating NELI Preschool has a positive impact even without the targeted sessions.

For other subgroups, the results show that NELI Preschool had a positive effect on increasing language skills, with the most robust effects (i.e., effect sizes not spanning zero) seen amongst the children at maintained settings (ES = 0.12 (0.03, 0.22)), and those for whom English was an additional language (ES = 0.12 (0.01, 0.24)). Whilst the effect amongst EYPP children is relatively lower (ES = 0.11 (-0.01, 0.24)), it's likely this is an underestimate of the true effect of NELI Preschool on this subgroup, as many of them did not attend enough hours a week to be able to access the targeted component.

Table 25: Subgroup analysis results

Subgroup	Analysis sample	Unadjusted means				Effect size			p-value
		Intervention group		Control group		Total N (intervention; control)	ICC (full)	Hedges' g (95% CI)	
		N (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)				
Actually Selected for targeted intervention (RQ3a)	S2 in intervention; S1 in control	874 (141)	-0.70 (-0.82, -0.58)	840 (153)	-0.95 (-1.09, -0.82)	1,714 (874, 840)	0.25	0.12 (0.02, 0.22)	0.04
Ideally Selected for targeted selection (RQ3b)	S1 in both intervention and control	877 (159)	-0.71 (-0.83, -0.59)	840 (153)	-0.95 (-1.09, -0.82)	1,717 (877, 840)	0.25	0.14 (0.045, 0.237)	<0.01
Enrichment Only (RQ4)	S3 in both intervention and control	815 (123)	1.01 (0.91, 1.12)	766 (136)	0.80 (0.68, 0.92)	1,581 (815, 766)	0.22	0.10 (-0.01, 0.20)	0.06
EYPP children (RQ5)	S4 in both intervention and control	449 (104)	-0.17 (-0.34, -0.00)	452 (86)	-0.47 (-0.65, -0.30)	901 (449, 452)	0.17	0.11 (-0.01, 0.24)	0.07
Children in PVI settings (RQ6a)	S1 and S3 in both intervention and control PVI settings	582 (113)	0.61 (0.47, 0.75)	585 (114)	0.27 (0.12, 0.43)	1,167 (582, 585)	0.21	0.08 (-0.06, 0.21)	0.26
Children in Maintained settings (RQ6b)	S1 and S3 in both intervention and control maintained settings	1,110 (169)	-0.14 (-0.25, 0.02)	1,021 (175)	-0.34 (-0.47, -0.22)	2,131 (1,110, 1,021)	0.22	0.12 (0.03, 0.22)	0.02
Children with EAL (RQ7)	S1 and S3 in both intervention and control settings	483 (105)	-1.13 (-1.31, -0.95)	441 (100)	-1.49 (-1.66, -1.32)	924 (483, 441)	0.18	0.12 (0.01, 0.24)	0.04

The estimates in Table 25 represent the differences in outcomes between each subgroup and its counterpart within the intervention arm, and these are very similar to the corresponding estimates obtained from the interaction models (i.e., the sum of the 'intervention' and 'subgroup × intervention' coefficients; see Table 26: Differential impact of NELI in the subgroups analysed). Using the interaction model, we also found no evidence that the impact of NELI Preschool differed across the analysed subgroups. Although subgroup-specific effect sizes varied slightly, these differences did not translate into meaningful contrasts between each subgroup and its counterpart: the 'subgroup × intervention' coefficients were close to zero, with confidence intervals that crossed zero, indicating no detectable differential effect.

Table 26: Differential impact of NELI in the subgroups analysed

		Intervention + Subgroup x intervention (impact in the subgroup)		Subgroup x intervention (differential impact of NELI Preschool for subgroup children)	
Subgroup	Total N (intervention, control)	Effect Size (95% CI)	p-value	Effect Size (95% CI)	p-value
EYPP	3,618 (1,858, 1,760)	0.13 (0.03, 0.24)	0.01	0.05 (-0.04, 0.15)	0.29
Setting Type	3,298 (1,692, 1,606)	0.06 (-0.06, 0.19)	0.32	-0.06 (-0.22, 0.1)	0.44
EAL	3,298 (1,692, 1,606)	0.12 (0.01, 0.22)	0.03	0.02 (-0.08, 0.12)	0.68
Targeted	3,298 (1,692, 1,606)	0.12 (0.04, 0.21)	<0.01	0.05 (-0.03, 0.13)	0.21

Analysis in the presence of non-compliance

An analysis was completed on the outcomes of children who attended settings who were deemed to have complied with the NELI Preschool programme. Compliance was assessed at the setting level as per the criteria set out in the Methods section above.

Table 27 summarises the number of settings that were compliant with each individual criterion as well as overall compliance. Whole-group and small-group sessions were counted as delivered if at least one child attended. Compliance was high across all separate criteria, with over 85% of settings meeting each individual requirement, and 81% of settings compliant overall.

In total, 81% of children in intervention settings were included in the compliance analysis, along with all children in control settings. The intervention children were drawn from 139 settings who returned a complete set of SDLs for Weeks 1–15 and were therefore eligible for inclusion. Of these, 113 settings met all three compliance criteria.

Table 27: Settings meeting compliance criteria

	Compliance Criteria				Compliant overall
	Online training	Whole-group sessions	Small-group sessions	Individual sessions	
Number of settings	134 (96%)	131 (94%)	135 (97%)	120 (86%)	113 (81%)

The results of the CACE analysis (set out in Table 28) show that there is a positive effect size (0.13; CI: 0.08, 0.19) for children attending compliant settings, with the whole 95% CI lying above zero. This is a higher effect size compared to the primary outcome analysis of 0.10, suggesting that NELI Preschool has a beneficial effect on language skills among children where the intervention is implemented as expected. The *p*-value is < 0.01, suggesting that the positive effect of NELI Preschool seen in compliant settings is unlikely to have occurred by chance alone, and that following the programme closely leads to a more positive impact on children's oral language skills.

Table 28: Analysis in the presence of non-compliance

Model stage	Total <i>n</i> (intervention, control)	Predictor	Standardised effect size (95% CI)	<i>p</i> -value
Stage 1: Compliance indicator regressed on intervention status	3,186 (1,580; 1,606)	Intervention status	2.90 (2.83, 2.30)	< 0.01
Stage 2: Language Skills first principal component regressed on compliance indicator		Compliance indicator	0.13 (0.08, 0.19)	< 0.01

As the compliance analysis was completed on the subset of 139 intervention settings that returned SDLs for at least Weeks 1–15, the primary outcome analysis was rerun using this subset as a sensitivity analysis to ensure that restricting the sample in this way did not introduce bias into the compliance analysis. As set out in Table 29 below, the results of this analysis are very similar to those obtained in the primary analysis, suggesting that negligible bias has been introduced by limiting the samples for both the dosage and compliance analyses to those intervention settings that returned SDLs for at least Weeks 1–15.

Table 29: Primary outcome analysis and sensitivity analysis

			Unadjusted means			Effect size			
			Intervention group	Control group					
Model	Outcome	<i>n</i> (missing)	Mean (95% CI)	<i>n</i> (missing)	Mean (95% CI)	Total <i>n</i> (intervention; control)	ICC (full)	Hedges' <i>g</i> (95% CI)	<i>p</i> -value
Primary Analysis	LanguageScreen and RAPT First Principal Component	1,692 (282)	0.12 (0.03, 0.21)	1,606 (289)	-0.12 (-0.22, 0.02)	3,298 (1,692, 1,606)	0.22	0.10 (0.02, 0.18)	0.01
Primary Analysis on Compliance Subset		1,580 (394)	0.12 (0.03, 0.22)	1,606 (289)	-0.12 (-0.22, 0.02)	3,186 (1,580, 1,606)	0.21	0.11 (0.03, 0.18)	0.01

Missing data analysis

Due to the study design (only children that had completed baseline LanguageScreen assessment were eligible to be sampled for endline testing), there were no missing baseline LanguageScreen scores. Of the 4,390 children selected for endline testing, 3,754 attended both LanguageScreen and RAPT testing at endline. However, 22 were judged by setting practitioners not to have been able to engage with or complete the LanguageScreen assessment and were therefore not included in the final analysis. Therefore, 3,732 children were subsequently analysed. Table 30 sets out the numbers who completed each of the assessments.

Table 30: Missing at endline by sample

	Selected for Endline testing	Completed LanguageScreen Only	Completed RAPT Only	Completed Both and Analysed
All Selected for Endline Testing	4,390	636	0	3,732*
S1: Ideally Targeted Children	2,029	300	0	1,729
S2: Practitioner Targeted Children (intervention settings only)	1,015	138	0	877
S3: Enrichment-only Sample	1,840	251	0	1,589
S4: EYPP	1,091	187	0	904

* 22 children couldn't engage fully with or complete LanguageScreen, so they have not been included for analysis.

Table 31: Missing at endline for primary outcome sample (S1 and S3)

	Selected for Endline testing	Completed LanguageScreen Only	Completed RAPT Only	Completed Both and Analysed
All Selected for Endline Testing	3,869	551	0	3,298
Intervention	1,974	275	0	1,606
Control	1,895	276	0	1,692

As set out in Table 29 above, 15% of children selected for endline testing had missing primary outcome data. Missingness was not evenly distributed across trial arms, with a higher proportion of missing data in the control group (19%) compared with the intervention group (11%). As overall missingness exceeded the 5% threshold, and to account for the imbalance between groups, we followed the EEF statistical guidance (EEF, 2022) to assess and address potential biases arising from missing data.

First, we examined patterns of missingness using a logistic regression model where the outcome was 'missingness'. Covariates included baseline oral language skills, randomisation group, setting type, and setting size indicators as per the primary outcome model, as well as additional variables hypothesised to relate to missingness but not included in the primary analysis. At the setting level, these were Ofsted rating, IDACI quintile, SPH, proportion of children who were EYPP-eligible, and NELI Preschool Delivery model. At the child level, these were EYPP eligibility, gender, EAL status, and age (measured as age in months at baseline).

Table 32: Missing data analysis model results

Level*	Variable	N (%) covariate missing	Odds Ratio (95% CI)	p-value
Child	Baseline oral language skills	0 (0%)	1.164 (1.083, 1.251)	<0.001
	Age (Months at baseline)	0 (0%)	1.018 (0.991, 1.045)	0.2
	EAL: Yes	0 (0%)	0.768 (0.602, 0.982)	0.035
	EYPP: Yes	0 (0%)	0.794 (0.624, 1.01)	0.061
	Sex: Male	0 (0%)	0.938 (0.778, 1.132)	0.507
Setting	Delivery Model: B	0 (0%)	1.58 (0.966, 2.584)	0.069
	Delivery Model: C	0 (0%)	1.667 (0.92, 3.019)	0.092
	Ofsted: Not judged	0 (0%)	1.294 (0.748, 2.239)	0.357
	Ofsted: Outstanding	0 (0%)	1.472 (0.888, 2.443)	0.134
	Ofsted: Requires Improvement	0 (0%)	1.513 (0.437, 5.24)	0.514
	IDACI: Decile 10	0 (0%)	0.669 (0.328, 1.363)	0.268
	IDACI: Decile 2	0 (0%)	0.686 (0.374, 1.257)	0.223
	IDACI: Decile 3	0 (0%)	1.074 (0.57, 2.023)	0.826
	IDACI: Decile 4	0 (0%)	1.179 (0.616, 2.256)	0.62
	IDACI: Decile 5	0 (0%)	0.807 (0.407, 1.601)	0.54
	IDACI: Decile 6	0 (0%)	0.859 (0.442, 1.668)	0.653
	IDACI: Decile 7	0 (0%)	0.942 (0.457, 1.943)	0.872
	IDACI: Decile 8	0 (0%)	0.803 (0.382, 1.685)	0.562
	IDACI: Decile 9	0 (0%)	0.994 (0.46, 2.15)	0.988
	Proportion of EYPP Children	0 (0%)	0.509 (0.187, 1.388)	0.187
	Setting Size: One Class	0 (0%)	1.283 (0.866, 1.902)	0.214
	Setting Type: PVI	0 (0%)	0.657 (0.478, 0.904)	0.01
	SPH: Derbyshire and Nottinghamshire EYSPH	53 (1.2%)	0.997 (0.472, 2.105)	0.993
	SPH: East of England EYSPH	53 (1.2%)	1.11 (0.527, 2.339)	0.784
	SPH: EYE South West EYSPH	53 (1.2%)	1.208 (0.565, 2.58)	0.626
	SPH: Great North EYSPH	53 (1.2%)	0.757 (0.346, 1.656)	0.486
	SPH: Kent EYSPH	53 (1.2%)	1.002 (0.482, 2.085)	0.995
	SPH: London South EYSPH	53 (1.2%)	0.492 (0.248, 0.975)	0.042
	SPH: Northern Lights EYSPH	53 (1.2%)	0.913 (0.425, 1.958)	0.814
	SPH: St Edmunds EYSPH, Yorkshire and Humber	53 (1.2%)	0.336 (0.127, 0.888)	0.028
	SPH: Thames Valley EYSPH	53 (1.2%)	1.058 (0.511, 2.193)	0.879
	SPH: Yorkshire and Humber Together EYSPH	53 (1.2%)	0.941 (0.441, 2.009)	0.875

* Baseline categories for factor variables are: EAL = No; EYPP = No; Sex = Female; Delivery Model = A; Ofsted = Good; IDACI Decile = 1; SPH = Bright Futures North West EYSPH

Table 32 indicate that EAL status and two of the SPHs are significant predictors of missingness, based on a p -value < 0.05 (in addition to baseline oral language skills and setting type, both of which are already covariates within the primary outcome analysis). Following the EEF guidance roadmap, we conducted a sensitivity analysis by re-running the primary outcome model with these two covariates added. This analysis (see Table 33) showed minimal difference between the estimates, suggesting that missing data did not bias the primary outcome analysis, so no further sensitivity analyses were undertaken.

Table 33: Missing data sensitivity analysis

	Outcome	Unadjusted means				Effect size			<i>p</i> -value
		Intervention group		Control group		Total <i>N</i> (intervention; control)	ICC (full)	Hedges' <i>g</i> (95% CI)	
		<i>N</i> (missing)	Mean (95% CI)	<i>N</i> (missing)	Mean (95% CI)				
Primary Outcome Model	LanguageScreen and RAPT First	1,692 (282)	0.12 (0.03, 0.21)	1,606 (289) (282)	-0.12 (-0.22, 0.02)	3,298 (1,692, 1,606)	0.22	0.10 (0.02, 0.18)	0.01
Missing Data sensitivity analysis	Principal Component	1,692 (282)	0.12 (0.03, 0.21)	1,606 (289) (282)	-0.12 (-0.22, 0.02)	3,298 (1,692, 1,606)	0.21	0.11 (0.03, 0.18)	0.01

Additional analyses and robustness checks

Mediation analysis

Practitioner confidence was conceptualised as a short-term outcome in the ToC, and a mediation analysis was conducted to examine the extent to which improvements in children's oral language skills were mediated by increased practitioner confidence.

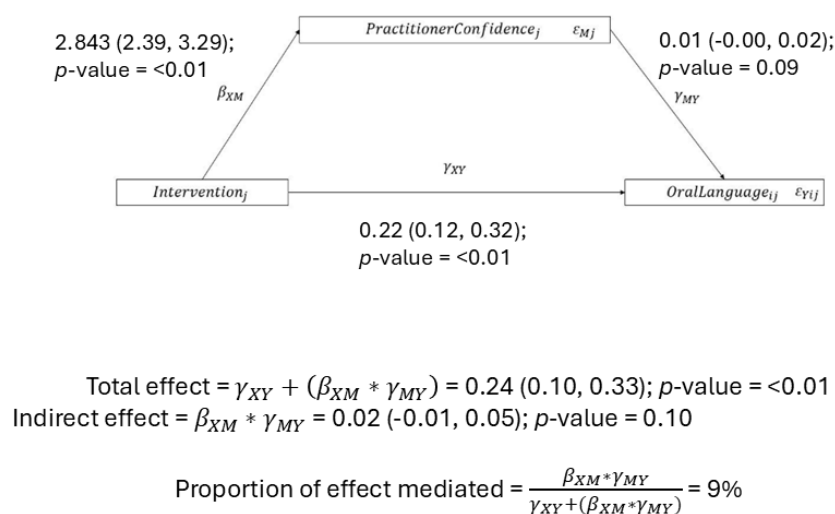
This was calculated as the summed score of confidence-related survey items completed at baseline and endline. Setting-level means were then derived at each timepoint to reflect change within settings, acknowledging that multiple staff could deliver NELI Preschool in each setting.

The results in Table 34 below show that NELI Preschool had a strong positive effect on practitioner confidence (2.84; 2.39, 3.29). However, confidence was not significantly associated with oral language skills after adjusting for the intervention (0.01; -0.00, 0.02). Reflecting the results of the primary analysis, the total effect of NELI Preschool on oral language skills was strongly positive (0.24; 0.14, 0.33), with most of this being a direct effect (0.22; 0.12, 0.32). The indirect effect through practitioner confidence was small (0.02; -0.01, 0.05), accounting for only 9% of the total effect. These findings suggest that NELI Preschool improves oral language skills primarily through direct pathways rather than via changes in practitioner confidence, and that even if practitioner confidence doesn't change, oral language skills may.

Table 34: Mediation analysis results

Path	Estimate (95% CI)	p -value
Intervention > Confidence	2.84 (2.39, 3.29)	<0.01
Confidence > Oral Language Skills	0.01 (-0.00, 0.02)	0.09
Total Effect	0.24 (0.14, 0.33)	<0.01
Direct Effect	0.22 (0.12, 0.32)	<0.01
Indirect Effect	0.02 (-0.01, 0.05)	0.10
Proportion Mediated	9%	

Figure 6: Estimated direct and indirect effects from mediation analysis



Dosage analysis

The key statistics for the dosage analysis are set out in Table 35. On average, children received 76 whole-group sessions, 38 small-group sessions, and 12 one to one sessions over the delivery period. These numbers fall below the programme's expected dosage of 100 whole-group, 60 small-group, and 20 one to one sessions. There was also considerable variation in the number of small-group sessions received ($SD = 18.80$), indicating large differences in children's exposure to this element. Overall, only a very small proportion of children met the expected dosage: 3% for small-group sessions and 1–1.5% for whole-group sessions. A higher proportion—though still modest—met the expected level of one to one sessions (10.0%)²⁶.

Table 35: Dosage attendance statistics

	Whole-group sessions— Enrichment-only Children	Whole-group sessions—Targeted Children	Small-group sessions	One to one sessions
Mean (<i>SD</i>)	75.7 (19.1)	76.8 (17.3)	38.4 (18.8)	12.1 (6.9)
Minimum	11	13	0	0
Maximum	100	100	60	20
Proportion achieving target	1.3%	1.6%	3.2%	10.0%

²⁶ The apparent discrepancy reflects differing definitions: compliance is a setting-level measure based on expected versus delivered one to one sessions over 15 weeks, while dosage is a pupil-level measure over 20 weeks (imputing weeks 16–20) that counts total sessions received and the percentage receiving at least 20.

The results of the dosage analysis models are set out in Table 36. For the enrichment-only model, the results indicate that each additional whole-group session is associated with a very small improvement in oral language skills. This pattern is mirrored in the Enrichment + Targeted model for both whole-group and one to one sessions, although the estimated effect of one to one sessions is notably larger. While the effect of small-group sessions is very slightly negative, the 95% confidence interval (-0.01, 0.01) spans zero, indicating no evidence of a meaningful effect.

This contrasts with other results that do demonstrate meaningful impacts of NELI Preschool and is likely explained by the dosage analysis being conducted within intervention children only, which provides insufficient contrast to reliably estimate the effect of different dosage levels. It is also possible that total dosage has reached a saturation point—i.e., children may already be receiving more input than is needed for additional gains, so further increases in small-group dosage produce little to no incremental benefit. These null results, therefore, do not undermine the overall conclusions about the effectiveness of NELI Preschool.

Table 36: Dosage results

Model	Variable	Mean (SD)	Expected Number (% meeting)	Dosage Effect Estimate (95% CI)	p-value
Enrichment Only (N = 860)	Whole-group session	72 (19.38)	100 (1%)	0.00 (-0.00, 0.01)	0.80
Enrichment + Targeted (N = 1,079)	Whole-group session	72 (17.75)	100 (1.5%)	0.00 (-0.02, 0.04)	0.42
	Small-group session	47 (18.92)	60 (3%)	-0.00 (-0.01, 0.01)	0.89
	One to one session	15 (6.80)	20 (10.7%)	0.01 (-0.02, 0.04)	0.56

Post-hoc sensitivity analyses

As outlined in the Methods section above, two sensitivity analyses were conducted to address potential threats to internal validity identified during the peer review process. The first analysis examined whether the uptake of language-development programmes among control settings introduced bias into the analysis. Results from this sensitivity analysis are presented in Table 37. The estimated mean difference in latent language scores between the intervention and control groups was 0.22 (intervention: 0.12 [0.03, 0.21]; control: -0.12 [-0.22, -0.02]). This corresponds to an effect size of 0.11 (0.03, 0.19) with a *p*-value of 0.01. This estimate is highly consistent with the primary analysis effect size of 0.10 (0.02, 0.18), indicating that the uptake of concurrent language-development programmes by a small number of control settings did not introduce detectable bias into the primary outcome results.

The second sensitivity analysis addressed evidence that a slightly higher proportion of practitioners in control settings had prior experience with NELI or NELI Preschool training, again examining if this introduced bias. The results from this sensitivity analysis are also presented in Table 37. The estimated mean difference in latent language scores between the intervention and control groups was 0.33 (intervention: 0.19 [0.08, 0.31]; control: -0.14 [-0.27, -0.00]). This corresponds to an effect size of 0.17 (0.06, 0.28) with a *p*-value < 0.01. This estimate is notably higher than that from the primary analysis (0.10 [0.02, 0.18]). This suggests that including settings with practitioners who had prior NELI or NELI Preschool training experience may have biased the primary results downward, leading to an underestimate of the effect of the intervention for settings with no previous exposure. We note, however, that this is an effectiveness trial and prior experience of NELI (a different intervention to NELI Preschool) was considered part of BAU, so the lower effect size of 0.10 is still the best estimate of the overall effect for the population of interest.

It is also worth noting that for all three effect sizes, the 95% confidence intervals overlap, suggesting that the observed differences between the estimates are small relative to their uncertainty and that the data provide limited evidence of meaningful differences between them.

Table 37: Post-hoc sensitivity analysis results

Model	Unadjusted means				Effect Size			
	Intervention		Control					
	<i>N</i>	Mean (95% CI)	<i>N</i>	Mean (95% CI)	Total <i>N</i> (intervention; control)	ICC (full)	Effect Size (95% CI)	<i>p</i> -value
Primary Outcome Analysis	1,692	0.12 (0.03, 0.21)	1,606	-0.12 (-0.22, -0.02)	3,298 (1,692, 1,606)	0.22	0.10 (0.02, 0.18)	0.01
Sensitivity Analysis One	1,692	0.12 (0.03, 0.21)	1,153	-0.10 (-0.22, 0.01)	2,845 (1,692, 1,153)	0.20	0.11 (0.03, 0.19)	0.01
Sensitivity Analysis Two	992	0.19 (0.08, 0.31)	875	-0.14 (-0.27, -0.00)	1,867 (992, 875)	0.23	0.17 (0.06, 0.28)	<0.01

Estimation of effect sizes

As set out in the Methods section above, effect sizes have been calculated as set out in the EEF's statistical guidance document (EEF, 2022).

For analyses that have a continuous dependent (outcome) variable (i.e., the primary outcome analysis, subgroup analyses, and CACE analysis), effect sizes are presented as Hedge's *g* using the formula set out in the Methods section. In each instance, the numerator for the effect size calculation is the coefficient of the intervention/control dummy variable from the multilevel model for each analysis. The variance figures are taken from the equivalent empty model (i.e., without covariates). The 95% confidence intervals for each effect size have been derived by multiplying the standard error of the intervention/control dummy variable coefficient by 1.96. These have then been converted to effect size confidence intervals using the same formula as the effect size itself.

For analyses that have a binary dependent variable (i.e., the missing data substantive model), effect sizes are reported as an odds ratio with a corresponding 95% CI.

Estimation of ICC

Setting-level ICCs have been calculated using the formula set out in the Methods section above. Pre-test ICCs have been computed using a two-level random intercept model with no covariates, whilst post-test ICCs have been derived from the primary outcome analysis model described above. The post-test ICC from the primary outcome analysis has been used in power calculations and to derive the MDES for the primary outcome analysis for both the primary analysis sample and the EYPP sample.

Implementation and process evaluation results

The following findings are based on the data collected as part of the IPE, including staff surveys, session delivery logs, and interviews and observations with case study settings (see the IPE methods section for more information). The case study data, in particular, represents a very small, self-selected sample, so findings that rely on this data alone should be treated with caution.

The IPE results are structured according to the IPE dimensions and research questions outlined in the Evaluation objectives section above. Each research question is divided into more specific sub-questions. The key findings for each of these sub-questions are provided as bullet points at the start of each section, and a summary of the findings in response to the overall research question is presented at the end of each section.

Compliance and fidelity

IPE RQ1²⁷: To what extent did NELI Preschool successfully scale up to a greater (i) number and (ii) variety (i.e., both maintained and PVI) of settings?

As part of this trial, the NELI Preschool programme was delivered in a larger number of settings than were seen in the efficacy trial (West *et al.*, 2023). Furthermore, a much higher proportion of these settings were PVIs compared to the sample used in the efficacy trial, and a much larger number of PVIs participated in the trial compared to the formative evaluation (see separate Appendix for further details). As a result, one of the aims of this trial was to understand how the programme scaled to both a larger number and variety of settings.

In this section, we will outline our findings in relation to:

- Compliance with the training requirements, and whether the training and support provided equipped settings to successfully deliver the programme;
- Compliance with the intended protocol for selecting children for the targeted group; and
- Delivery of the programme sessions and whether this aligned with intended programme practice.

Training and support

IPE RQ1.1.1: To what extent were practitioners engaged with training and support that equipped them to successfully deliver the intervention?

- The training course had high completion rates; the webinars less so.
 - Participation in the community of practice was low and mentor input was minimal.
 - The training course and delivery team support were perceived to be useful for supporting settings to successfully deliver the intervention, but this was less evident for the webinars and Delivery Support Hub.
-

The online training course had high completion rates, but attendance rates at the webinars were lower

Compliance with the training requirements was very high. According to data from the platform that hosted the programme training, two practitioners had completed the training before the start of delivery in 80% of settings, with an average of three practitioners per setting. By the end of the delivery period, 94% of settings had at least two trained practitioners, and there was an average of 3.6 trained practitioners per setting. However, 8% of settings still reported cases of untrained staff delivering one or more sessions. There was no meaningful difference in training completion by setting type or deprivation level. Three settings had no practitioners complete the training but were deemed by the delivery team to have completed a sufficient proportion of it to proceed with delivery. Two of these did not proceed to deliver any of the programme but still completed endline testing, while the third did neither.

Attendance rates at the four webinars were lower. There were 158 attendees at the first webinar, but this dropped to 32, 22, and 18 for the remaining three, respectively. A similar pattern was observed for views of the recording, with 196 views of the first webinar recording, dropping to 106, 108, and 66 views for the following three. The higher number of views than attendees, particularly after the initial webinar, suggests that practitioners did not see particular value in attending the webinars live, indicating that these webinars were not perceived as active opportunities for knowledge exchange. This hypothesis is supported by practitioner perspectives on the webinars shared via the surveys and interviews (see below).

²⁷ IPE sub-questions from the protocol may be worded slightly differently in this section, but the findings presented reflect the full scope of the protocol's intended RQs.

Results from the survey suggest that webinar attendance was relatively distributed amongst settings, with 81% of surveyed NELI Preschool Leads saying they attended or viewed at least one webinar. Nearly all case study staff who reported attending few or no webinars said that this was because of constraints posed by their work schedules. Only one interviewee reported that they did not attend because they did not feel they needed it.

Participation in the community of practice was low and mentor input was minimal

The programme ToC identifies practitioner engagement with the community of practice around NELI Preschool, via the interactive elements of the online training and the Delivery Support Hub, as a key activity. Limited interaction occurred via either of these digital platforms.

While there were 151 intervention settings, there were on average only 38 responses to each interactive element in the first half of the training and 11 in the second half.²⁸ There were very few instances of practitioners responding to each other or to the mentors. While most of the surveyed NELI Preschool Leads (81%) reported having used the Delivery Support Hub, levels of contribution to interactive elements of the Hub were very low, with only 26 contributors in total and an average of less than one post per week of delivery.

Across both platforms, practitioner contributions to these online platforms usually related to the specific activity prompts, with some participants also asking about potential adaptations, or sharing positive feedback about the programme. Mentors provided positive commentary on practitioner contributions, as well as responding to questions about potential adaptations.

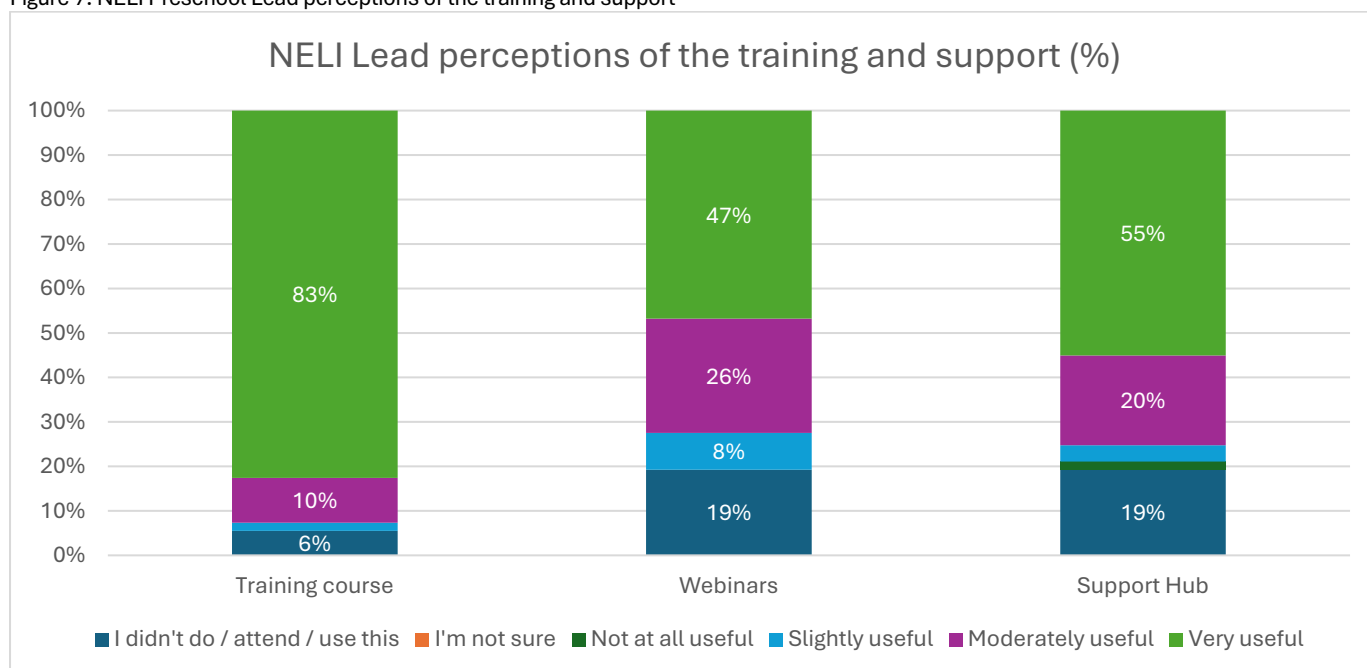
NELI Preschool Leads considered the online training course to have been very useful for supporting effective programme delivery—the webinars and Delivery Support Hub slightly less so

The large majority of surveyed NELI Preschool Leads reported finding the online training course ‘very useful’ in preparing them to deliver the programme (

²⁸ These included responses from mentors and a small number of settings from a previous evaluation.

Figure 7), while around half did so for the webinars and Delivery Support Hub. However, around one-fifth did not find the webinars or Delivery Support Hub useful. This is reflected in the declining rates of attendance at the webinars (see above) and levels of interaction with the Delivery Support Hub. These findings suggest that the webinars and Delivery Support Hub may not be core requirements for effective delivery. The lower levels of engagement observed for these elements are consequently unlikely to have compromised the programme's potential for impact.

Figure 7: NELI Preschool Lead perceptions of the training and support



Source: Question C from the NELI Lead Endline Survey: How useful did you find (1) the online training course (FutureLearn), for preparing you to deliver the programme; (2) the webinars, for supporting you to deliver the programme effectively; (3) the NELI Preschool Support Hub, for supporting you to deliver the programme effectively? (n = 109)

Note: Data labels for proportions under 5% have been removed to support visual clarity.

Most of the case study practitioners ranged from somewhat to very positive about the online training course, describing it as 'enjoyable', 'informative', and 'helpful', but also 'fine' and 'pretty standard'. The videos of session delivery, quizzes, and the ability to go back to the training at any point during delivery were noted as particularly helpful aspects of the course. However, many also noted that the training course had been time-consuming to complete, and a few mentioned that practitioners with higher qualification levels and/or more experience would already be very familiar with the content covered.

'It was alright. It was sometimes quite informative. But some of the videos you have to watch, oh my god, I'm sitting there for 20 minutes. [...] But maybe if you're not confident, watching the whole thing would give you more ideas or something.'—

Practitioner

Others felt that the training had not struck the right balance in terms of combining information on theory and on implementation, and that there could have been more guidance on how to support children who were less well behaved and/or less willing to participate. Accessibility concerns were also flagged due to the amount of reading required. Several practitioners reported finding the tips about introducing programme elements into continuous provision a particularly helpful aspect of the webinars, and one practitioner reported appreciating the chance to ask questions. On the other hand, a couple of case study settings reported finding the first webinar unhelpful, as they felt it did not provide any information that had not already been covered by the training. Some case study settings reported finding the Delivery Support Hub useful for ideas for integrating NELI Preschool into continuous provision, particularly for less experienced practitioners. However, others reported that they had not used the Hub because they did not have enough time for it and/or did not perceive the need for it. Settings reported that support from the OxEd team had met their needs. Nearly all of the surveyed NELI Preschool Leads (94%) said they were 'very' or 'extremely' satisfied with the support they received from the OxEd team, which was reported to primarily occur via phone calls and emails. Several of the case study interviewees reported having received a very high level of support, with a few particularly appreciating the high levels of responsiveness from the support team. A couple of settings mentioned that they would have appreciated the opportunity to see a session delivered in person in a similar context to their own.

Targeted group selection

IPE RQ1.1.2: To what extent did practitioners correctly select children for the targeted component using LanguageScreen and the Targeted Group Selection Guidance?

- **Most settings fully complied with the guidance when selecting children for the targeted group.**

Most settings complied with the guidance when selecting children for the targeted group

The large majority (81%) of surveyed NELI Preschool Leads indicated that they fully complied with the guidance when selecting the group of children for the targeted intervention.

One of the case study settings reported having included children in the targeted group, not in line with the initial guidance to replace children who had initially been selected but then left. A few other case study settings reported that they would have selected different children outside of the trial context, primarily due to the perception that higher-ability children had performed poorly on the screening for reasons other than language ability (such as behaviour).

While 17% of the children selected by the settings for the targeted group (sample S2) were not among those NFER selected based on the guidance (sample S1), this appears to have primarily been the result of a change in situation or understanding in the time since the relevant data was shared with NFER, rather than non-compliance with the guidance. For example, children's attendance patterns and/or practitioners' understanding of their needs had changed since they initially provided this data to NFER, resulting in some children no longer meeting the eligibility criteria, or vice versa.

Compliance with session delivery requirements

IPE RQ1.1.3: To what extent did practitioners deliver the 20-week programme in the intended format?

- Settings delivered close to the expected number of whole-group and individual sessions, but small-group session numbers were lower—potentially due to settings delivering fewer sessions with larger groups. In addition, the average number of sessions attended by children was also lower than expected.
- There was a wide range of staffing arrangements between settings, but in most cases, a range of staff members were involved.
- Whole-group sessions were usually delivered in an appropriate environment, but targeted sessions could be more challenging to find space for.
- Settings mostly delivered sessions of the recommended length, with some variation either side.
- Settings consistently used most of the programme materials, but a significant minority did not always use the digital slides, which may have resulted in children missing out on some of the content and activities.
- Use and perceived usefulness of the Narrative Progress Tracker varied between settings.

Settings delivered close to the expected number of whole-group and individual sessions, but small-group session numbers were lower—potentially due to settings delivering fewer sessions with larger groups

Settings were expected to deliver a total of 100 whole-group sessions (5 per week x 20 weeks), 120 small-group sessions (3 per group per week x 2 groups x 20 weeks), and 120 individual sessions (1 per individual per week x 6 individuals x 20 weeks) over the course of the delivery period. Compliance levels were set at 75 whole-group sessions (75%), 84 small-group sessions (70%), and 78 individual sessions (65%). As shown in Table 38: Number of NELI Preschool sessions delivered, settings that returned all their SDLs on average far exceeded compliance levels for whole-group and individual sessions, and rates of child attendance at them are high. However, the average number of small-group sessions delivered was much lower than intended, even though rates of child attendance remained high. This suggests that instead of settings delivering six small-group sessions per week, with three children in each, some settings may have been delivering fewer sessions per

week with a larger number of children in each session. This was observed in several case study settings, usually in an effort to accommodate child absence. Overall, these findings suggest that children may not have always received the level of focused attention and opportunity for participation intended for the small-group sessions.

Table 38: Number of NELI Preschool sessions delivered

	Whole-group sessions	Small-group sessions	Individual sessions
Mean number of sessions delivered (% of intended)	95 (95%)	69 (58%)	105 (88%)
Mean number of sessions attended per child (% of intended)	72 (72%)	47 (78%)	15 (75%)

Source: Session delivery logs (n = 122)

Note: The mean for child attendance at the small-group and individual sessions was calculated using only those children selected by settings for the targeted group (S2). Each child was expected to attend half of the small-group sessions and only one of the individual sessions delivered each week.

Some of the surveyed NELI Preschool Leads reported that their setting chose not to deliver any small-group and/or individual sessions—but this was a very small proportion (4% and 8%) and can be accounted for in part by settings who withdrew from programme delivery entirely. This means that children with the lowest language development levels were not able to benefit from additional targeted support in these settings, which may have impeded their access to whole-group session content as well, as predicted by the ToC.

A wide range of staff were involved in delivering the programme, as intended

Settings were advised by OxEd to staff session delivery in the manner that worked best for them. The survey and case study data both indicate that settings took a wide variety of approaches to staffing the sessions, and most had a range of staff members with different experience and qualification levels involved. However, according to the surveyed practitioners, more senior and higher qualified staff were more likely to deliver a large proportion of the whole-group sessions, and less likely to deliver a large proportion of the targeted sessions. This trend was particularly observed in PVI settings. Based on the case studies, maintained settings often had teachers delivering the whole-group sessions and TAs the targeted, although other arrangements did also occur.

Whole-group sessions were usually delivered in an appropriate environment, but this was less common for targeted sessions

While the programme is not prescriptive about where the sessions are delivered, it is advised that it be a quiet space that supports the children to remain focused on the session. The large majority of the observed whole-group sessions took place in the main room of the nursery, with the remainder taking place in a nearby room. In both these cases, the children were given a quiet environment in which to focus on the session. When the session took place in the main room, any children for whom the programme was not seen to be suitable (for example, younger children or those with complex needs) were taken by other practitioners to play outside or in a separate room. In some cases, however, all children participated—including two-year-old children sharing the same space. This occurred either as an active choice on the part of the setting or as a result of staff shortages that meant it was not possible to split the cohort into multiple spaces while respecting the required staff-child ratio.

There was, however, more variation in the suitability of space for targeted session delivery. Half of the surveyed NELI Preschool Leads (53%) reported that the space where they delivered their targeted sessions was ‘very’ or ‘extremely’ suitable, and this was the case for only 40% of Leads from PVI settings. There were some observed instances of the small-group session taking place in the main room with the other children present during free play. This created an extremely noisy and disruptive environment that could not be considered suitable for the session. This suggests that children in PVI settings may have had a more disrupted—and consequently less effective—experience of the targeted sessions.

Settings mostly delivered sessions of the recommended length

Settings were recommended that whole-group sessions should last around 15 to 20 minutes, small-group sessions 10 to 15 minutes, and individual sessions 5 to 10 minutes. However, these were intended only as guidelines, with practitioners advised to be responsive to child participation levels in determining the length of the session. Most of the surveyed NELI Preschool Leads reported targeted sessions of the intended length, but a significant minority of settings delivered shorter whole-group sessions than intended (Table 39). Clear differences also emerged by setting type, with maintained settings more often delivering shorter whole-group sessions (46%—30 out of 65 surveyed NELI Leads) compared to PVI settings (18%—7 out of 40). While the programme is not prescriptive on session length, it emphasises that there must be sufficient time for as many children as possible to speak and interact. These findings suggest that children in maintained settings may have received fewer opportunities of this kind.

Table 39: Session length reported by NELI Preschool Leads compared to intended session length

	Whole-group sessions (%)	Small-group sessions (%)	Individual sessions (%)
Shorter than intended	35	21	NA
Intended length	45	61	60
Longer than intended	20	18	36

Source: Question H from the Endline NELI Lead Survey: ‘On average, how long did each of the following types of sessions last?’ (n = 98–105)

This data aligns with what was observed in the case study settings, where most sessions lasted the recommended length, with a few lasting between 5 and 10 minutes longer. One setting did, however, report that their whole-group and targeted sessions usually lasted just 5–10 minutes and 5 minutes respectively. A couple of practitioners mentioned modifying the length of the session depending on how the children were responding to it, as OxEd intended. However, several practitioners did report that both the whole-group and small-group sessions could be too long at times.

Settings consistently used most of the programme materials, but a significant minority did not always use the digital slides, which may have resulted in children missing out on some of the content and activities

A quarter (23%) of surveyed NELI Preschool Leads reported that they had not always used a screen for the whole-group sessions. This was particularly common in PVI settings (31%—13 out of 42). OxEd does not consider this an acceptable adaptation, and they emphasised the importance of using the screen throughout the delivery period, including through the webinars. As the content of the handbook does not cover everything in the slides, children would have not received all the programme content when only the handbook was used to deliver the whole-group sessions. This was observed in one of the case study visits. One setting explained that they did not always use the slides for the dialogic reading sessions, as having the children looking at the slides all the time ‘feels like a lot’ and is not engaging for the children. These findings indicate that children did not necessarily receive all the programme content even when all sessions were delivered, and that this particularly affected children in PVI settings.

Use of the other materials for their intended purpose was, however, very consistently observed throughout the case study visits. The hard copies of the texts were always used in the small-group sessions and sometimes in the whole-group sessions as well, particularly at the start of the session to provide a recap of the story before moving into the activities. The handbook was usually, but not always, used as part of the small-group sessions, although this was recommended. Use of the press-out characters was not observed, but one case study setting noted that the children had enjoyed using them. The NELI puppet, listening rules poster, and stickers were consistently used as intended in the observed whole-group sessions and sometimes in the small-group sessions as well. Settings also regularly provided additional resources to support with the activities or to act as prompts during dialogic reading of the texts.

A small number of case study settings mentioned using the Narrative Progress Tracker extensively throughout the delivery period, as intended for example, by sticking post-it notes on the Tracker throughout the week with indications of how each child was progressing. Around half of the surveyed NELI Preschool Leads (53%) reported finding it ‘very’ useful for helping them to support the progress of the targeted children, particularly those in maintained and/or more deprived settings, with

a further 28% finding it ‘moderately’ useful. However, 11% of the surveyed NELI Preschool Leads reported not using the Narrative Progress Tracker at all.

Overall, these findings suggest that the materials provided were judged to be helpful and appropriate, but that there were some divided opinions around the desirability of using the screen for every session, which likely resulted in some children not being exposed to the full programme content and activities.

Fidelity to intended practice within the sessions

IPE RQ1.1.4: To what extent did practitioners deliver the programme in accordance with the script?

- Observed sessions broadly followed the training guidance around supporting language development, although exploration around narrative structure was less clearly implemented.
- Most settings invited most children to speak during whole-group sessions, using a range of different approaches to encourage the children to do so, although additional guidance may be required around how to facilitate discussion over single-word responses and ensure more reluctant speakers are given the opportunity to participate.

IPE RQ1.2: To what extent, and in what ways, did settings adapt the intervention to better meet their own needs?

- Most settings adapted the programme in some way, by including adding, modifying, and/or skipping particular session elements.
 - Settings varied in the extent to which they integrated NELI Preschool’s elements into their continuous provision and parental engagement was low.
-

According to the ToC, each programme session is intended to provide children with the opportunity to learn and practise new vocabulary, explore the narrative structure of stories, and engage in active discussions to practise their language skills.

Practitioners in all of the observed sessions consistently demonstrated excellent scaffolding of new vocabulary

This included clearly introducing new words, repeating the words many times and in different contexts, and reminding the children about new vocabulary from previous sessions when the opportunity arose. However, the extent to which the children used the new words themselves varied significantly between settings. There was also variation in the extent to which practitioners linked the new vocabulary with other related words. These strategies align with what the programme advised as part of the training. This suggests that children were largely provided with consistent high-quality support to develop their vocabulary through the programme sessions.

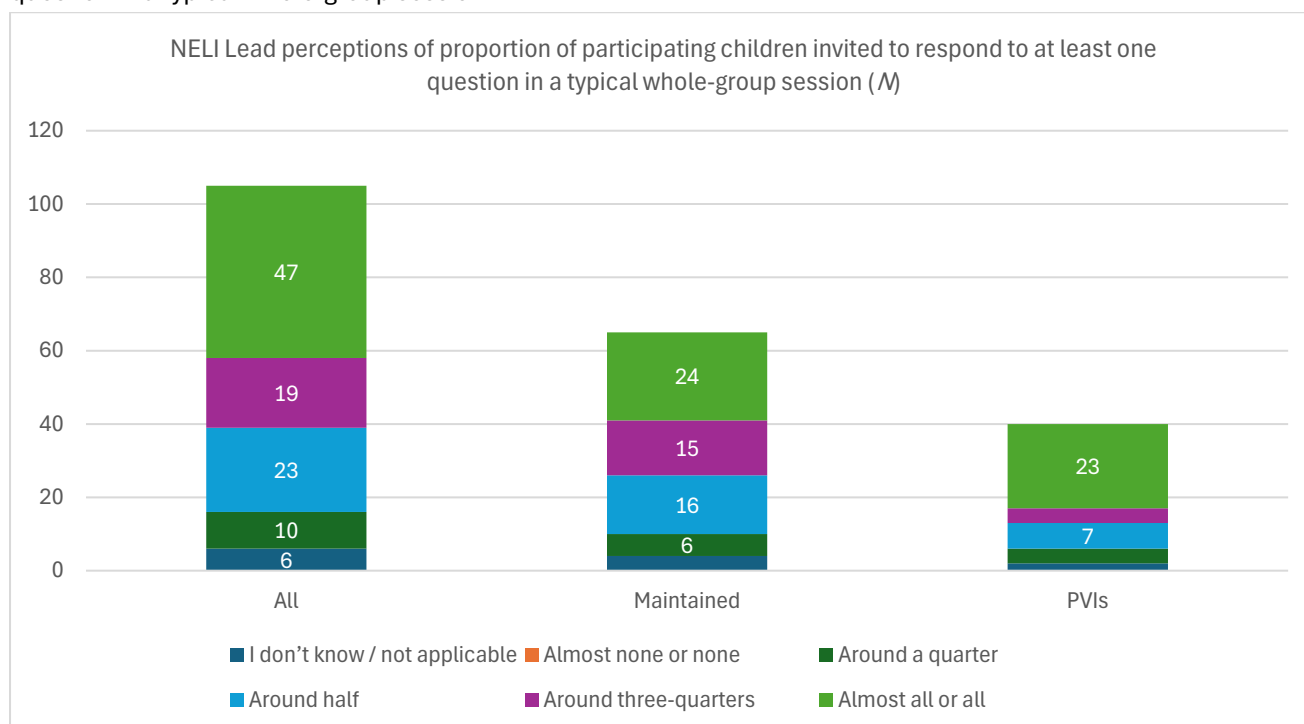
Opportunities to explore narrative structure were less clearly evidenced in the observed sessions

Many practitioners would emphasise the sequence of events when talking through or summarising the story, but in several of the observed cases, this aspect was not addressed. One case study setting reported that expecting the children to use the character and narrative cards to re-construct the story was not an age-appropriate task, and children were seen to struggle with this activity in one case. These findings indicate that children were provided with fewer opportunities to develop their narrative skills as part of the sessions, which suggests that the impact on more complex language development may have been limited.

Most children were given the opportunity to speak during the sessions, but level of discussion varied

Nearly half of the surveyed NELI Preschool Leads (45%) reported that almost all or all children were invited to respond to at least one question in a typical whole-group session (Figure 8). This was more commonly reported in PVI settings (58%—23 out of 40), compared to maintained settings (37%—24 out of 65), which suggests that children in maintained settings may have received fewer or unequal opportunities to practise their active language skills—a core element of the ToC.

Figure 8: NELI Preschool Lead perceptions of proportion of participating children invited to respond to at least one question in a typical whole-group session



Source: Question 1.2 from the Endline NELI Lead Survey: 'In a typical whole-group session, what proportion of the participating children were invited to respond to at least one question (i.e., selected to respond to a question)?' (n: all =105, maintained = 65, PVIs = 40)

Note: Percentages are not used in this chart due to the small sample size of the maintained and PVI subgroups. Data labels for numbers under five have been removed to support visual clarity.

Practitioners in nearly all of the observed sessions prompted children to speak using a range of different question types, including open-ended questions requiring recall and/or relating to the child's own life. Some practitioners asked more challenging questions that scaffolded children to respond for example, by breaking the question down into smaller steps or using a prop to make an abstract concept more concrete and specific. However, there were a few cases where the focus was primarily on inviting children to fill in blanks, which tended to elicit single-word responses rather than full sentences. While practitioners in nearly all of the observed sessions consistently evaluated and expanded on children's responses, there were only a very small number of cases where the child was then asked to repeat the expansion in turn.

A range of practices recommended by the training for encouraging children to speak were also observed, including allowing children plenty of time to respond, providing positive reinforcement when children did contribute (including non-verbally for pre-verbal children), following the child's lead and not blaming the child if the practitioner found it difficult to hear or understand them.

Some observed settings deliberately invited a range of different children to speak. However, in several cases, children were encouraged to call out answers, which usually facilitated only single-word responses, and only occasionally were specific children called on to respond. In a few cases, more confident, talkative children dominated the discussion, with quieter children not being called upon to speak. This occurred both when children were free to call out and where the practitioner asked for children to put their hands up to respond.

The nature of the small-group sessions was such that children were nearly always each asked for their input in turn, as well as given opportunities to speak through repetition and call-and-response.

These findings suggest that the programme's intention of providing opportunities for children to practise speaking as part of the sessions was largely successful in practice, and practitioners demonstrated skills in responding to the children in ways that would support further development. However, additional guidance may be required around how to facilitate discussion over single-word responses and ensure more reluctant speakers are given the opportunity to participate.

A significant minority of settings reported making adaptations to the sessions, usually to support child engagement, suggesting that there may be scope for making the programme more age appropriate

A third (30%) of surveyed NELI Preschool Leads reported skipping, modifying, or adding session activities. Many of the case study practitioners reported that they made ‘tweaks’ to some of the activities, usually in an effort to make them more suitable for the children’s ability levels. This was observed in several of the sessions, with practitioners introducing additional elements such as a recap of words or activities from previous sessions, short additional activities (such as a quiz), or physical activity breaks. These adaptations were considered acceptable by the delivery team. However, several settings also reported skipping some of the activity elements or shortening the session if they felt it was not sufficiently engaging for the children, and these adaptations the delivery team did not deem acceptable.

Some settings integrated reading of the story into the sessions more than once per week to support children in accessing it, but this had implications for child focus and recall opportunities

A quarter (25%) of NELI Preschool Leads reported reading the whole book at the start of every whole-group session. This was an acceptable adaptation and likely to have been positive for the children in terms of reinforcing learning. Some case study settings explained that they made this adaptation because otherwise children with lower attendance would miss out on the story and/or because the children struggled to remember the story after only reading it once.

One case study setting reported that the children loved reading the book multiple times per week. However, other settings reported that doing this could make the session very long, which could be challenging for the children’s focus—particularly for those who had already heard the story on previous days. It also meant that the children would often not be asked to try to recall what had happened in the story, which is the intended approach to starting each session.

The prevalence of this adaptation suggests that the programme’s assumption that the children will be able to recall the story after reading it just on the Monday may not always hold. While the webinars encouraged settings to read the book with children outside of the sessions as well, this is not a formalised expectation of the programme.

Challenges around child recall also emerged in relation to the targeted sessions, with some case study settings reporting that the targeted sessions always took place straight after the whole-group sessions for this reason. While one setting felt this worked well, the other reported that it was a challenging length of time for the children to remain focused.

Additional elements for consideration in the ToC

OxEd introduced some elements of the programme to settings as optional, and these were not included in the ToC.

Settings took advantage of the flexibility around how and to what extent they integrated the programme into their continuous provision, rather than there being prescriptive requirements

Settings were encouraged to integrate elements of the programme (such as themes, vocabulary, and activities) into their broader continuous provision so as to reinforce the learning outside of the sessions. This was an optional element open to setting interpretation, but nearly all surveyed NELI Preschool Leads reported doing so—albeit to varying extents. Around a third (35%) reported integrating programme elements into the continuous provision ‘to a great extent’, while a further 29% reported doing so ‘to a moderate extent’, and 34% ‘to a small extent’.

Most case study settings had integrated the programme quite extensively into their wider provision, with some having the explicit aim of bringing it into every aspect of their provision to make everything as connected as possible. Many had other activities, stations, and/or displays relating to the special words and/or book for each week. Reading the book in circle time or story time outside of the sessions was common, and a couple of settings had purchased a second copy of each of the books for children to look at on their own. A few settings also mentioned using particular NELI materials and/or routines, such as the NELI puppet, vocabulary flashcards, song, and/or listening rules as part of their continuous provision as well. However, there were some case study settings that did not bring NELI Preschool into their wider provision—either because they wished to maintain some variety for the children and/or because they felt they did not have the time to do so.

The fact most settings chose to integrate the programme into their continuous provision suggests this was seen to be worthwhile. Practitioner enthusiasm for talking about this aspect in the interviews also suggests that this aspect may have positively contributed towards their motivation in relation to the programme by engaging their agency and creativity. It may be worth including this component in the ToC in future while retaining the openness of exactly how it is implemented.

Parental engagement was low, reflecting that this is not a core component of the programme

Parental engagement was another optional area of activity that settings were encouraged to pursue. Nearly all of the settings mentioned sending out the parent newsletter, and a small number implemented additional parental engagement strategies, such as a NELI-themed 'stay and play' area, a display for parents with the weekly special words and activities, feeding back to parents of targeted children at pick-up and parent-teacher night, and sharing the links to the audio recordings of the books. Settings also asked questions about parental engagement on the training and Support Hub platforms, as well as the webinars, suggesting a wider interest in this element. However, most settings reported very little interest or communication from parents about the programme, although they were aware of a very small number of parents buying copies of the books for their child and/or doing some of the activities at home.

Summary of findings for IPE RQ1

Overall, this trial saw the NELI Preschool programme being successfully delivered in a larger number and variety of settings than in previous studies. Compliance with the training requirements was high, the programme sessions were mostly delivered with fidelity, and suitable children were selected for targeted support. A significant minority of settings did, however, diverge from intended practice by not always using the digital slides for the whole-class sessions, and not all children were given the opportunity to speak in every session. Fewer small-group sessions were delivered than intended, which may have been a consequence of settings including more children than intended in each session. Levels of engagement with the webinars and community of practice were low, and these were seen to be less important for effective delivery.

Child responsiveness

IPE RQ2: *How well did the NELI Preschool intervention reach and support disadvantaged children (those eligible for EYPP and/or those with English as an Additional Language [EAL])?*

Children growing up in socio-economically deprived contexts already demonstrate a significant developmental gap compared to their less deprived peers by the time they reach Reception (Department for Education, 2026). Given the focus on English language in the NELI Preschool programme, it is also possible that children with EAL may face additional barriers in benefitting from it. In this section, we consider the evidence around how the programme experience may have differed for these particular children as compared to their peers, and whether there were any particular challenges or facilitators at play in this.

IPE RQ2.1: To what extent did disadvantaged children engage with the intervention?

- Most children, regardless of EAL or EYPP status, were seen and reported to have high engagement in the programme sessions, although this did not always translate into opportunities and/or willingness to speak.
-

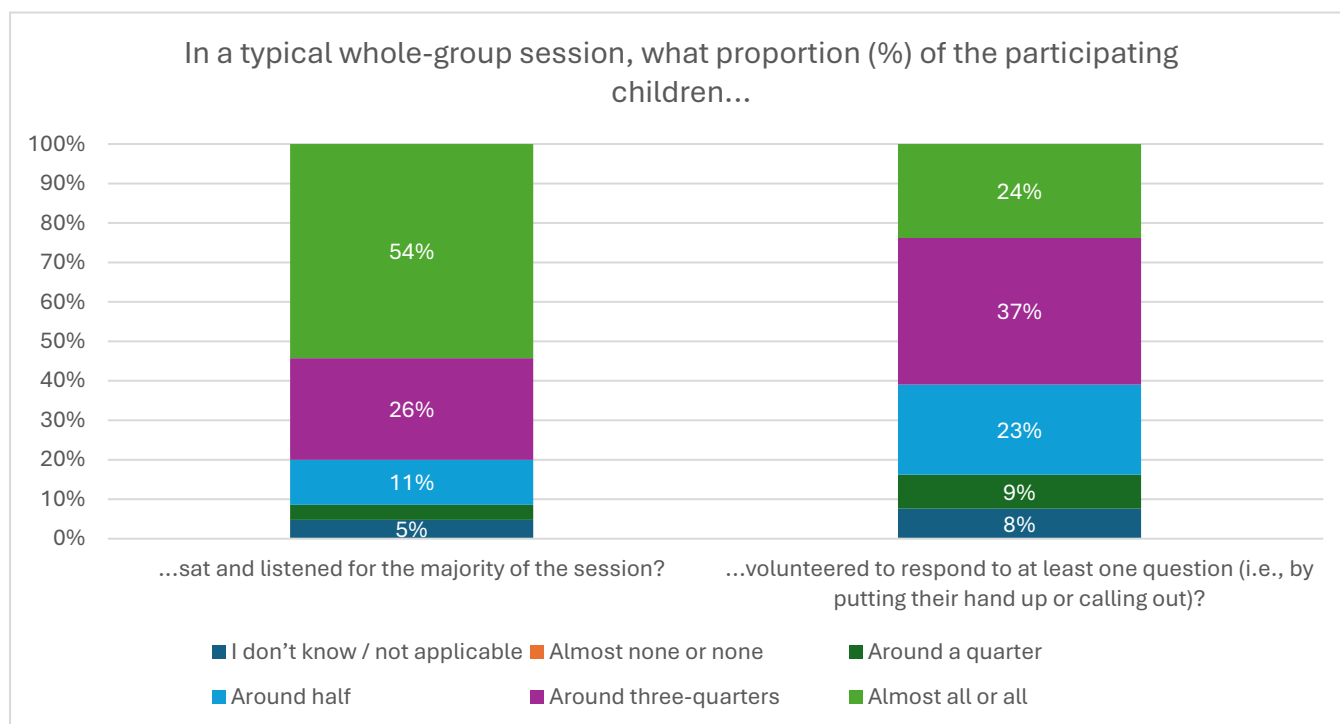
There was no clear difference in levels of engagement among disadvantaged children compared to their peers

Case study practitioners did not mention any concerns around the extent to which disadvantaged children engaged with the programme, and session observations suggest that these children tended to respond to the programme in an equivalent variety of ways to other children.

Most of the surveyed NELI Preschool Leads (80%) reported that in a typical whole-group session, most or all of the participating children sat and listened for the majority of the session, and a slightly lower proportion (60%) reported this in

relation to children volunteering to respond to at least one question (Figure 9). Similar patterns were observed across different setting types and deprivation levels. This suggests that levels of passive participation among children were very high, but levels of active participation were slightly more varied.

Figure 9: NELI Preschool Lead perceptions of child responsiveness and focus during whole-group sessions



Source: Question 1 in the Endline NELI Lead Survey: In a typical whole-group session, what proportion of the participating children (1) volunteered to respond to at least one question (i.e., by putting their hand up or calling out); (3) sat and listened for the majority of the session? (n = 105).

Note: Data labels for proportions under 5% have been removed to support visual clarity.

Many of the case study settings, and others on the Delivery Support Hub, emphasised the high levels of enthusiasm children showed towards the programme and the enjoyment they got out of it. In nearly all of the observed sessions, children were calling out or raising their hands to respond to questions, recalling parts of the story and previous special words, singing the song, calling out the listening rules, and applauding best listener. However, children did vary in the extent to which they appeared to be willing to speak during the sessions. A small number of children appeared to be distracted in several of the observed sessions, and in a small number of cases, many of the children became restless towards the end. One observed session showed very poor levels of child engagement after the first 5 to 10 minutes of a whole-group session, indicated by a high level of noise and movement within the group.

IPE RQ2.2: Were there any challenges or facilitators that applied to disadvantaged children in particular? How (if at all) did settings address these?

- Children eligible for EYPP attended fewer whole-group sessions than their peers. This appears to be the result of higher levels of absence, rather than fewer registered hours. EAL children attended more whole-group sessions than other children, and, similarly, this appears to be because of higher levels of attendance rather than more registered hours.
- Children with EAL were more likely than their peers to receive additional support via the targeted intervention.
- Children eligible for EYPP were less likely than their peers to meet the additional eligibility criteria for receiving the targeted intervention, primarily for reasons relating to attendance levels.
- While settings did not report taking any measures to support disadvantaged children to access the programme specifically, they took a flexible approach to the targeted sessions to try to ensure absence did not result in children missing out.

Case study practitioners were asked explicitly in the interviews whether they felt there were additional challenges for engaging disadvantaged children in the sessions. Neither the session observations nor these interviews provided any indication that socio-economically disadvantaged children responded differently to the programme sessions compared to their peers.

Attendance levels were lower among children eligible for EYPP, and those with the lowest levels were both less exposed to whole-group sessions and excluded from the targeted group because of their infrequent attendance—while the reverse was observed for children with EAL

Practitioners did not report any specific challenges for socio-economically disadvantaged children. However, according to the SDLs, children eligible for EYPP attended, on average (mean), a much lower proportion of whole-group sessions (60% of sessions) compared to their peers (73% of sessions). On the other hand, children with EAL were found to attend *more* whole-group sessions, on average (mean) (77%) than their peers (71%).

The lower dosage for children eligible for EYPP does not appear to be the consequence of fewer registered hours. While children eligible for EYPP were registered for 3 fewer hours per week on average (mean) compared to their peers (20 hours vs 23 hours), children with EAL were registered for 2 hours fewer than their peers but still attended more sessions.

Socio-economically disadvantaged children are more likely than their peers to have a higher level of authorised absence in primary school (DfE, 2025g). It is possible that a similar trend would be observed in early years settings and may account for these children attending fewer of the whole-group sessions. Children with EAL have, on the other hand, slightly lower rates of authorised absence in primary school compared to their peers (DfE, 2025g), which may also play a role in the differences observed.

Children selected for the targeted group (sample S2) attended a higher average (mean) number of whole-group sessions (77%), compared to all children (72%). This reflects the fact that children with lower levels of attendance and/or fewer registered hours tended to be excluded from the targeted group due to anticipated challenges around scheduling the small-group sessions, as stipulated by the Targeted Group Selection Guidance. However, the difference among children eligible for EYPP is much larger: 79% for those in the targeted group, compared to 60% of all children eligible for EYPP. This suggests that children eligible for EYPP were disproportionately impacted by the attendance criteria required for the targeted group, with 9% of all children eligible for EYPP failing to meet these criteria, compared to 4% of their peers.

There was no notable difference in the number of small-group and individual sessions attended by children eligible for EYPP or with EAL and their peers. This is likely the result of the fact that targeted sessions could be moved around to accommodate child absence, as well as the likely higher rates of attendance among those children eligible for EYPP selected for the targeted group compared to other children eligible for EYPP.

Overall, these findings indicate that attendance levels were lower among children eligible for EYPP, and those with the lowest levels were both less exposed to whole-group sessions and excluded from the targeted group. This suggests that the programme constraints may prevent it from reaching those children who would benefit the most.

Children with EAL appear to have particularly benefitted from the targeted support

On the other hand, children with EAL were disproportionately represented in the targeted group, with 31% of eligible children with EAL in the targeted group compared to 19% of those eligible and without EAL. This is likely the result of EAL children being more likely to meet the eligibility criteria (93% compared to 88% of their peers), as well as being more likely to demonstrate lower levels of language proficiency. These findings indicate that children with EAL were particularly well placed to benefit from the additional targeted support. Moreover, some case study settings reported finding the targeted sessions particularly beneficial for children with EAL, as it gave them a chance to speak if they were not confident doing so in the whole-group sessions. Children with EAL were also reported to face additional challenges in the whole-group sessions around vocabulary and length of the stories, which the targeted sessions would have helped to address.

The targeted sessions were also reported to have helped practitioners to develop a greater awareness of what they did and did not understand, as in the whole-group sessions, it could be ‘hard to tell how much they’re taking in’. However, one setting reported finding it challenging to deliver the small-group sessions when all the children in them had very limited English.

Settings introduced measures to ensure children with lower attendance did not miss out

While settings did not report taking any measures to support programme access for disadvantaged children specifically, they did report flexibility around when the targeted sessions were delivered and how many children were in each one to work around child absence. One also reported delivering the whole-group sessions in both the morning and afternoon to reach as many children as possible. In a small number of instances settings reported successfully requesting parents to alter the days of the child’s attendance to enable them to be part of the targeted group.

Summary of findings for IPE RQ2

Overall, we found that settings did not report any particular additional barriers to accessing the programme for socio-economically disadvantaged children. However, the data suggests that lower levels of attendance among children eligible for EYPP meant they were often exposed to less of the programme than their peers and were less likely to be eligible for additional targeted support. On the other hand, children with EAL appear to have received more of the programme and been particularly well placed to benefit from the targeted support.

IPE RQ3: To what extent can the targeted component of the intervention be seen to be fulfilling its intended purpose?

The purpose of the targeted intervention was to provide children with the weakest language skills with additional support to help them benefit from the programme alongside their peers. In the efficacy trial, a smaller impact was found for children in the targeted group compared to children who participated only in the enrichment element (West *et al.*, 2023). This led the OxEd team to hypothesise that the children selected for the targeted group may have been at too early a stage of their language development to be able to fully benefit from the intervention as it stands (West *et al.*, 2023). In this section, we will therefore look at whether the targeted group selection process enabled settings to identify those children who would benefit most from this additional support to receive it, and then whether the targeted component was seen as effective at supporting these children to better access the whole-group sessions. We also look at the extent to which there was a genuine distinction between children who received the targeted intervention and those who did not.

IPE RQ3.1: Did the Targeted Group Selection Guidance (including use of LanguageScreen) enable settings to identify the children who would benefit most from the targeted component?

- Eligibility criteria for the targeted group around SEN and attendance levels may have resulted in some children who would have benefitted more missing out on the targeted intervention.
 - Some settings felt that higher-ability children were sometimes included in the targeted group because they did not sufficiently engage with the LanguageScreen process.
-

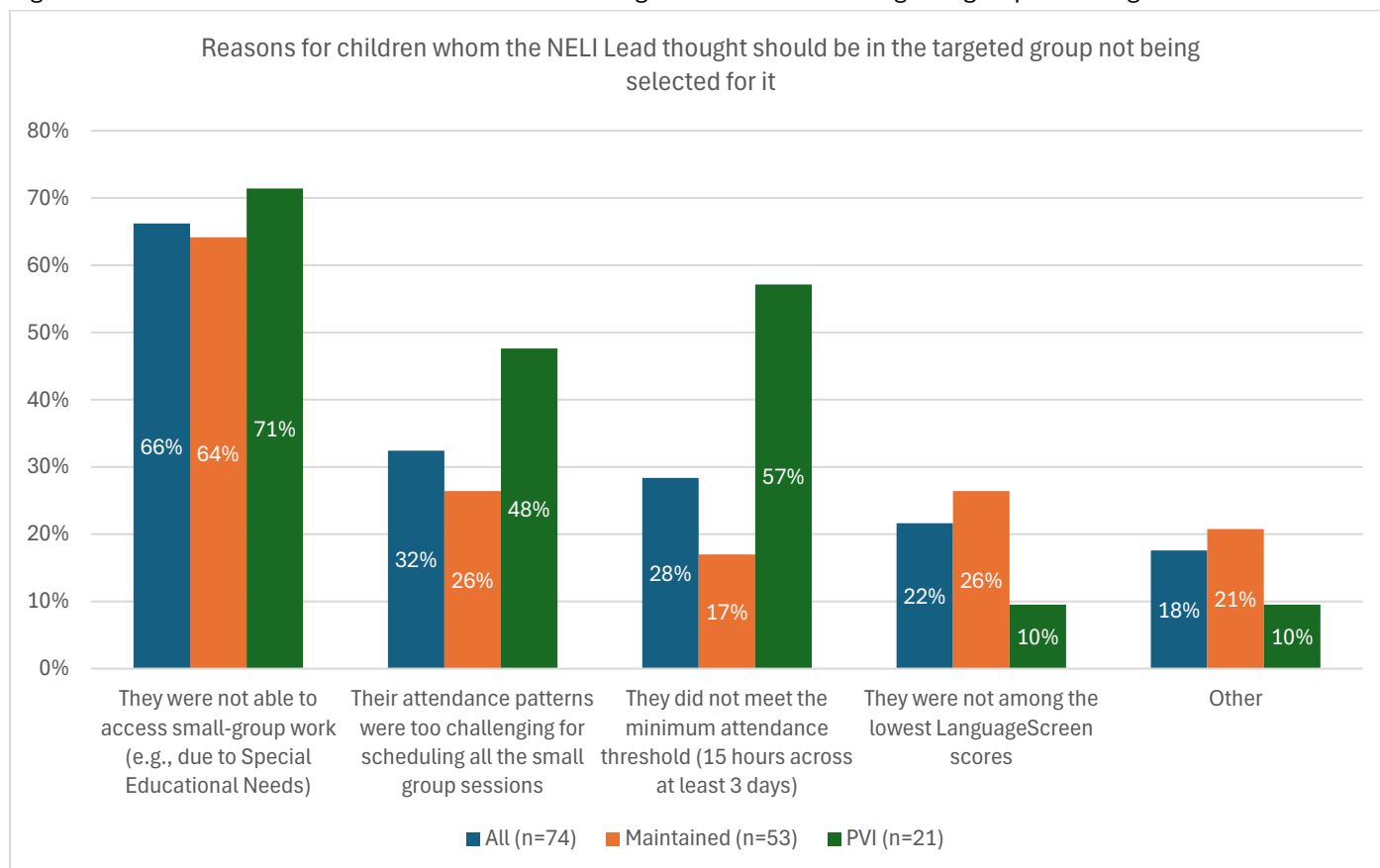
It was common for settings to feel that some children who would benefit most had missed out on the targeted intervention, primarily due to SEN and/or attendance levels—but there were also some concerns about the accuracy of LanguageScreen

Two-thirds (69%) of surveyed NELI Preschool Leads reported that there were other children in the nursery whom they thought would have benefitted more from the targeted intervention than the group selected in accordance with the guidance. This proportion was much higher in maintained settings (80%—53 out of 66) than in PVIIs (51%—22 out of 43), although further research is required to understand why this is the case.

According to NELI Preschool Leads, the most common reason (66%—49 out of 74) reported for these children not being selected for the targeted intervention was that they were not able to access small-group work (for example, due to SEN)

(Figure 10). Attendance levels and patterns were also a common issue, and much more so in PVIs than in maintained settings, as shown in Figure 10. This indicates that there were some children who would have benefitted from additional language support who were not able to access it. One case study setting reported feeling there were children who should have been in the targeted group who were struggling to get anything out of the whole-group sessions without the additional support.

Figure 10: Reasons for children who the NELI Lead thought should be in the targeted group not being selected for it



Source: Question E.1 from the Endline NELI Lead Survey ‘On the previous page, you indicated that there were children who may have benefitted from the small-group and individual sessions more than some of those selected. What were the reasons for this child or children not being selected to receive the small-group and individual sessions?’ (n = 74)

Around half of the case study settings reported feeling that the right children had been selected for the targeted group, but only one attributed this to issues around attendance levels. The case study sample is not intended to be representative of all settings in the trial, and is likely to overrepresent more engaged settings, which may explain this divergence from the survey findings. Others reflected that they had been surprised by some of the children selected, as they had not perceived any difficulties with their language abilities, while one noted that some children performed better than expected in the screening.

One in five of the surveyed NELI Preschool Leads (22%—16 out of 74) felt that there were children who had not been among the lowest LanguageScreen scores who would have benefitted more from the intervention than those who were. Some of the case study settings reported suspecting that some higher-ability children were included in the targeted group because they had not engaged with the screening process appropriately, resulting in low scores that did not reflect their true ability. They felt that other children who would have benefitted more had missed out as a result and would have changed the selection accordingly if they had not been part of the trial.

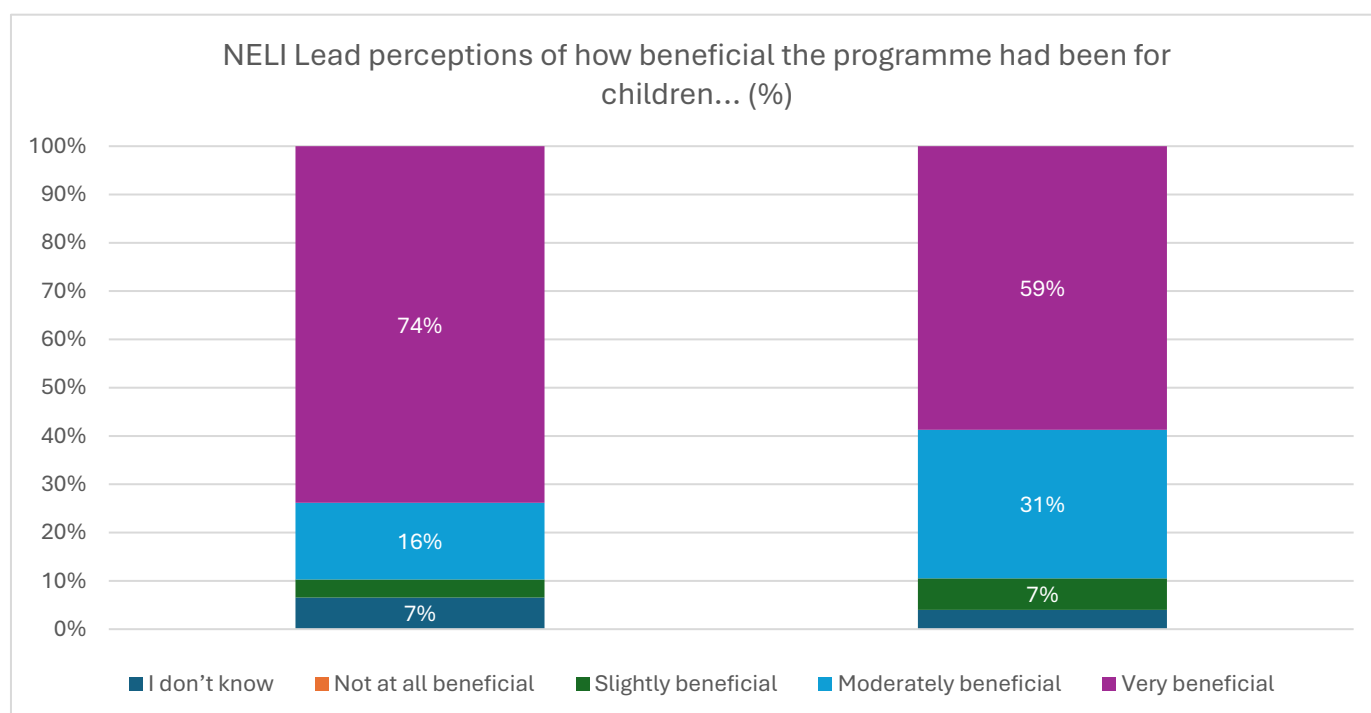
IPE RQ3.2: Is the targeted component set up to support selected children to access the whole-class element?

- The targeted intervention was seen to support participating children's access to the whole-group sessions by providing additional opportunities to engage with the content and practise the skills required.

The targeted intervention was seen to be highly effective for supporting children above and beyond what would have been possible just through the whole-class element, and support participation in the latter as well

Most (74%) of the surveyed NELI Preschool Leads reported that the programme had been 'very' beneficial for children in the targeted group (Figure 11). A smaller proportion, but still most of the surveyed NELI Preschool Leads (59%), thought the programme was 'very' beneficial for children *not* in the targeted group. These findings suggest that the targeted intervention was seen to bring additional benefit for participating children beyond that of the enrichment programme, as intended. However, the difference between the proportion who considered the programme to have been 'very beneficial' for targeted children and for non-targeted children was much larger for maintained settings (80% vs 58%), while in PVI the perceived benefit for the two groups was very similar (64% vs 60%). Maintained settings in the trial were significantly more likely to be in deprived areas compared to PVIs, which suggests they also had a higher population of socio-economically disadvantaged children, so this difference may relate more to differences in level of need between PVI and maintained settings rather than implementation differences. The impact evaluation does not support differential impact in the two setting types meaning that NELI Preschool worked similarly well in PVI and maintained settings (see the Conclusion section for further details).

Figure 11: NELI Preschool Lead perceptions of how beneficial the programme had been for children in the targeted group and not



Source: Question M in the Endline NELI Lead Survey: 'How beneficial did you feel the programme was for children who (4) participated in the small-group and individual sessions; (5) only participated in the whole-group sessions? (n = 107)

Note: Data labels for proportions under 5% have been removed to support visual clarity.

Several case study settings emphasised the benefits of the targeted sessions of providing opportunities to practise their oral language for children who did not feel confident speaking in a larger group, particularly as they would be asked questions directly and practitioners were able to focus more on using the relevant support strategies than when attending to the whole class. Practitioners perceived the children to enjoy the targeted sessions as a 'chance to be heard' and that they were particularly helpful for children who found it difficult to focus in the context of the whole-group sessions. It was both observed and reported that some children were more talkative and attentive in the small-group sessions compared to

the whole-group ones, likely due to more individualised attention and hands-on activities. However, group dynamics did still play a role, with the potential for one more confident child to dominate the session, or, conversely, for all children in the group to be reluctant to speak. Both of these cases are likely to have limited the usefulness of the sessions for some or all of the children involved.

Some practitioners also reported that targeted children were more willing to talk in the whole-group sessions because they had already covered the content in the targeted one. One practitioner also reported that changes in child language skills would be observed first in the targeted sessions before emerging in the whole-group sessions.

Several case study settings also reported finding the targeted sessions helpful for giving them a better understanding of what the children's strengths and areas of difficulty were, which may have led to these children receiving more appropriate targeted support as part of continuous provision as well.

IPE RQ3.3: To what extent does the distinction between targeted and enrichment activities and groups play out in practice?

- A significant minority of settings offered targeted sessions to children outside of the targeted group, but there is no indication that targeted children were regularly missing targeted sessions.
-

The distinction between targeted and non-targeted children mostly held, although a significant minority of settings chose to extend targeted support beyond the selected children as well

One in five (21%) of the surveyed NELI Preschool Leads reported delivering targeted sessions to children outside of the targeted group. This was more common in maintained settings (25%—16 out of 65), compared to PVLs (14%—6 out of 42), which may be a consequence of maintained settings having a stronger view on the relative usefulness of children having access to the targeted component, as discussed above.

Half of the case study settings reported delivering some kind of additional targeted NELI support to children outside the targeted group. This took the form of bringing other children into the small-group session if one of the targeted children was absent, or delivering additional targeted sessions to other children. Practitioners reported that they felt this would be beneficial for the other children without any negative impact for the targeted children. One practitioner saw it as beneficial to have non-targeted children in the group as well as they would model speaking skills for the targeted children.

While it is likely to have been beneficial for these children to have received additional and more tailored support, this may have diluted the distinction between targeted and non-targeted children in the trial analysis.

There is, however, no indication that targeted children were regularly missing targeted sessions, with dosage levels for the targeted group (sample S2) being relatively high (average [mean] of 78% attendance at small-group sessions, 75% at individual ones). Several settings reported that the children in the targeted group enjoyed the small-group sessions and were happy to leave their play to do them, although it could take a moment for them to settle down. If a child did not want to join in or could not focus, they were allowed to return to play, although this was seen and reported to happen very rarely.

Summary of findings for IPE RQ3

Overall, the targeted component of the intervention can be seen to have fulfilled its intended purpose of supporting children with the lowest language abilities to better access and consequently benefit from the enrichment component of the intervention, primarily by providing the children with more opportunities to speak and increasing their confidence through prior exposure to the material and personalised adult attention. A significant minority of children not in the targeted group also received targeted support, but this would simply benefit more children at scale and indicates that settings felt this aspect of the programme to be useful. However, the necessary eligibility criteria around being able to access small-group work and attend the setting for enough days to be part of the sessions does mean some children with low language abilities are unable to benefit from the additional support.

Context and moderators

IPE RQ4: *What are the key moderators and contextual factors that influenced how effectively the intervention was implemented in each setting?*

The programme ToC presented a number of factors that were anticipated to play a key role in determining the extent to which the programme would be effectively delivered within a given environment. In this section, we look at the role these factors played on the trial and reflect on their significance for potential future scale-up of the programme.

We had intended to look at the potential moderating role of process quality in the settings. However, a high level of process quality was observed in all case study settings, so it has not been possible to comment on its relationship with the quality of delivery.

IPE RQ4.1–3: What barriers and facilitators did settings encounter in implementing the intervention? What (if any) is the role of the structural and process quality of a setting in moderating how effectively the intervention was implemented? What (if any) is the role of other potential moderators (see the ToC) in moderating how effectively the intervention was implemented?

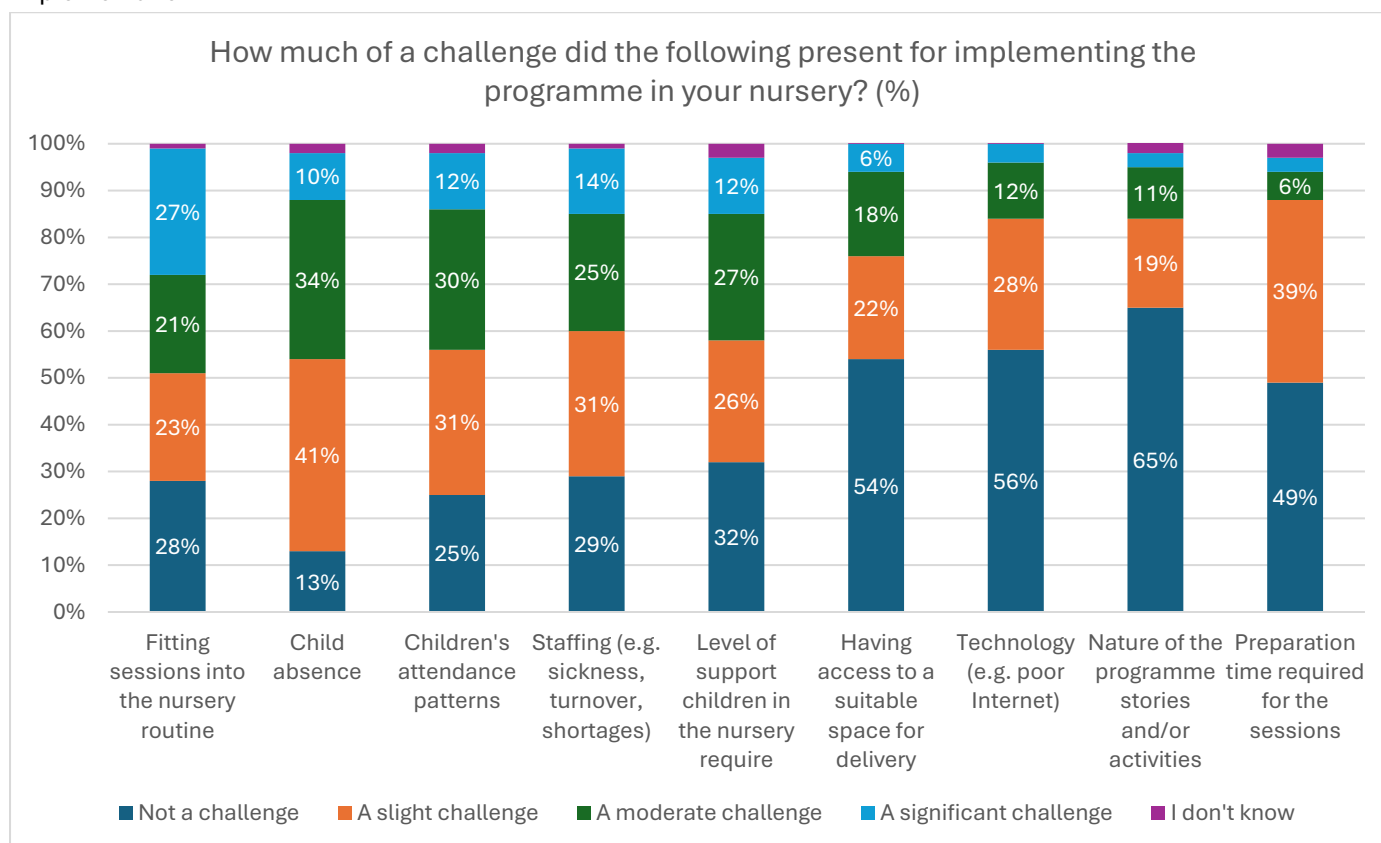
- Children’s attendance and staffing shortages were the main moderators for effective programme implementation.
 - Finding the time to complete the training and attend the webinars was a challenge, with nearly all practitioners reporting to have undertaken these activities in their own time—sometimes unpaid.
 - The level of preparation required to implement the programme was both significant and necessary for effective implementation; however, this was mostly felt to be manageable by practitioners and facilitated by the quality of programme materials.
 - Most settings found LanguageScreen straightforward to use and useful for selecting the targeted group, as well as providing additional information about children’s ability levels and areas for improvement—although the challenging nature of the assessment was also flagged.
 - The space and technology required for the intervention were not felt to be barriers for implementation; however, observations of case study settings illustrated how these may moderate delivery.
 - The cost of the programme may act as a barrier for scale-up, particularly for settings that do not already have a suitable screen.
 - High levels of process quality were observed in all case study settings, so it was not possible to comment on its relationship with the quality of delivery.
-

Challenges

Settings found it challenging to fit all the sessions into their routine , which may present a challenge for scale-up

Most of the surveyed NELI Preschool Leads (71%) perceived fitting sessions into the nursery routine to be a challenge, as shown in Figure 12, with a quarter (27%) reporting this to have been a ‘significant’ challenge. Over half (38 out of 68) of those surveyed NELI Preschool Leads who said they would not continue with NELI Preschool or only in a modified form indicated that this was at least in part due to challenges fitting the programme into their schedule. One case study setting also explained that it was not sustainable to use up so much of the nursery timetable on one programme, and others reported that children were missing out on other activities they would like to deliver. These findings suggest that the high levels of dosage required by the programme may constitute a challenge and may not be necessary for achieving the intended impact (see Dosage analysis in the Impact evaluation results section for further details).

Figure 12: NELI Preschool Lead perceptions of the extent to which various factors present a challenge for programme implementation



Source: Question G in the Endline NELI Lead Survey: 'How much of a challenge did the following present for implementing the programme in your nursery?' (n = 108)

Note: Data labels for proportions under 5% have been removed to support visual clarity.

The nature of the programme content does not appear to have caused any major difficulties: most of the surveyed NELI Preschool Leads reported that the programme was at least 'moderately' well aligned with their existing nursery curriculum (83%) and that the nature of the programme stories and activities did not present any challenges for implementation (65%) (Figure 12). However, a few case study settings reported that the story cards to support the development of narrative skills were 'not that intuitive', and that errors identified in the slides and handbook meant the settings had to check and correct the materials in advance of the sessions. A couple of settings also reported that they did not think the programme activities and examples were always age appropriate, requiring skills that the children were unlikely to have developed by this age—such as clapping with a partner, skipping, or playing a musical instrument. Some of the surveyed NELI Preschool Leads also reported that the selected texts were not always sufficiently engaging for the children.

Child absence and attendance patterns were the most widely reported challenges

Most of the surveyed NELI Preschool Leads (85%) identified child absence as at least a 'slight' challenge for programme implementation, with nearly half (44%) considering it at least a 'moderate' (Figure 12) challenge. This was slightly more of a challenge for settings in more deprived areas, likely as a result of lower attendance rates among more socio-economically disadvantaged children. Similarly, most (73%) of the surveyed NELI Preschool Leads identified children's attendance patterns as a challenge to implementation. This was more of a challenge in PVI settings than maintained ones, reflecting the more irregular attendance patterns that occur in the former case. Both of these factors determined the ease with which settings were able to schedule the targeted sessions. Where these challenges were very significant, it is likely that children with lower attendance rates would miss out on elements of the programme (as discussed earlier), and this may have disproportionately affected more disadvantaged children and children in PVIs. Where staff delivered additional sessions to ensure these children still received support, this is likely to have increased workload and taken children away from other activities. This factor is likely to constitute a key moderator for which children are able to benefit from the programme as part of any wider scale-up as well as the level of delivery burden experienced by the setting.

Staffing was a key challenge for programme implementation, both for session delivery and training completion

Most of the surveyed Neli Preschool Leads (70%) reported that staff absence due to sickness, turnover, and shortages had been a challenge for implementation (Figure 12), and this was particularly common among settings in more deprived areas. Roughly a quarter (28%—19 out of 68) of Neli Preschool Leads who reported that they would not continue with the programme, or only in a modified form, indicated that this was at least in part due to not having adequate staffing capacity. Many case study settings highlighted challenges in relation to staff shortages and identified how this impacted their ability to effectively implement the programme for example, having to include younger children in the whole-group sessions (rather than the younger children being supervised elsewhere), which could cause disruption; not being able to support the level of need in their class during whole-group sessions; having to deliver targeted sessions during continuous provision for all children, leading to more interruptions, or not being able to deliver the targeted sessions at all.

Staffing challenges also limited the proportion of practitioners within the setting able to complete the training, with most case study settings reporting difficulties in finding the time or capacity to complete training. For most settings, the staffing ratios did not allow for practitioners to do the training during their usual working hours, resulting in 24 of the 26 case study practitioners doing the training in their own time, and this time was unpaid for 10 of these practitioners. This reliance on their own time meant that it was difficult for some practitioners to fit in the training depending on their personal circumstances, such as childcare or other responsibilities. For similar reasons, most case study settings reported that it was even more difficult or not possible to find time to attend webinars, as reflected in the lower levels of attendance in comparison to training. Of the 19 case study practitioners who attended at least one webinar, 16 reported doing so in their own time, and 7 of these were unpaid. As the trial provided additional funding to enable settings to pay staff for their training time, it is likely the proportion of staff doing training unpaid in their own time would increase in the context of wider take-up. However, settings outside of the trial may also choose to spread training completion over a longer period of time, which could allow for more opportunities for staff to complete it during their working hours.

‘[Doing the training] was a bit of a strain for me because of my kids and stuff like that. And I had to do it quite late, obviously, after they’d gone to bed.’

—Practitioner

Where case study settings had sufficient staff available to complete the training, preparation and delivery could be shared, younger children could be supervised in separate areas during whole-group or small-group sessions, and children with SEN could be supported. It was observed that interruptions could be better managed and individual children better supported when the session was facilitated by a greater number of adults. In particular, children with SEN were observed occasionally joining in with sessions enabled by additional adults who supported them in exiting/rejoining sessions when they felt able to.

As the sector faces significant ongoing recruitment and retention challenges, as well as high rates of turnover, these findings suggest that the staffing demands of the programme may act as a barrier to scale-up and/or lead to lower levels of trained staff and implementation fidelity, particularly in more deprived communities.

The level of preparation required to implement the programme was both significant and necessary for effective implementation; however, this was mostly felt to be manageable

Information from seven of the case studies indicates that settings spent an average (mean) of 5.5 hours setting up the programme, and a further 3.5 hours preparing for sessions each week. Nonetheless, only half (48%) of surveyed Neli Preschool Leads reported finding preparation time a challenge (see Figure 12), and most (76%) reported that the level of preparation for whole-group sessions was ‘very’ manageable or better. A slightly lower but still majority proportion (59%) reported finding preparation for the targeted sessions manageable.

Case study settings reported that advance preparation was necessary for the smooth delivery of sessions, particularly for ensuring any materials for the session are available and for checking if any adaptations are required. One case study practitioner explained: ‘It is hard work to do it properly and do it well. But we are enthusiastic about it, so we want to spend the time’. This suggests that the amount of time spent on preparation could be flexible depending on practitioner

expectations around the quality of delivery. As a result, while preparation time may not constitute a common barrier to wider take-up, challenges in this respect may result in a lower quality of delivery.

Neither the space nor technology requirements commonly presented challenges for programme delivery

Less than half of surveyed NELI Preschool Leads (45%) reported that finding a suitable space had been a challenge for delivery, and only a quarter (24%) considered it a 'moderate' or 'significant' challenge. However, this challenge affected a larger proportion of more deprived settings, likely reflecting constraints around infrastructure and resources. Nonetheless, most (88%) NELI Preschool Leads reported that the space for the targeted sessions was at least 'moderately' suitable, suggesting that this is unlikely to be a key concern for scale-up. Where the space was not suitable, however, the level of disruption to the targeted sessions from external noise and movement was observed to be substantial.

'You have to have that quiet space because a couple of times I've been at the back of the room and done [the Targeted sessions] [...] but their mind is just not fully concentrated on it.'—Practitioner

Observations of the NELI Preschool sessions in case study settings highlighted that the utilisation of available space can vary based on competing priorities, such as the dual use of spaces for multiple interventions, sensory areas, and provision for children with SEN. Furthermore, additional space in the setting can only be capitalised on for delivery when staff are also available to supervise children.

A similar proportion (44%) of surveyed NELI Preschool Leads reported technology such as poor internet to have been a challenge for delivery, and only 16% considered it a 'moderate' or 'significant' one (Figure 12). However, this was, again, slightly more common in more deprived settings.

Several case study settings encountered minor technical difficulties, some related to internet connectivity, that disrupted whole-group session delivery. Technical difficulties were also an issue addressed in webinars, and advice was given on how support could be sought and slides set up for offline use. Overall, however, this was not seen to be a key area of concern.

The cost of the programme and required equipment may act as a barrier for scale-up

Nearly half (46%—31 out of 68) of the surveyed NELI Preschool Leads who indicated that they would not be continuing with the programme, or only in a modified form, indicated that this was at least in part due to cost. Moreover, this does not take into account the cost of the prerequisite materials for delivery. One-third (36%) of the intervention settings were provided with screens for the purpose of the trial, as they did not already have one. In addition, 60% were provided with a digital tablet, 30% with a TV trolley, and 11% with wireless TV keyboards. All four of the case study settings that reported having received a screen from OxEd indicated that they would not otherwise have been able to purchase one due to the cost. This is likely to be one of the most important barriers to scale-up.

'We would have had to buy [the screen] [...] which I think is a bit beyond what we could afford just in the current state of things. [...] I don't think the children would be as engaged if we didn't have it. [...] So I think we'd have had to save up a bit. Which is tricky when you're taking on something new, and you don't know if it's any good; it would have been a lot to shell out.'—Practitioner

Facilitators

Most settings found LanguageScreen useful and straightforward to use, which is likely to facilitate scale-up

Most of the surveyed NELI Preschool Leads (77%) reported finding LanguageScreen 'very useful' for identifying the children who would benefit the most from the targeted intervention, with a further 17% finding it 'moderately useful'.

These findings are echoed by the case study settings, many of which described LanguageScreen as being easy and straightforward to administer. Several settings reported using the LanguageScreen data for more than just the targeted

group selection process. For example, they mentioned using the data to improve their understanding of the children's ability levels and to determine areas of language development that required greater focus across the nursery as a whole.

However, several case study settings reported feeling that the assessment was too long and/or challenging for the age group. One practitioner also indicated that it would have been helpful to have had more guidance around the marking criteria. A few settings noted that they did not feel the LanguageScreen scores always accurately reflected the children's abilities, as in some cases the children provided incorrect answers because they were distracted, uncooperative, or did not understand what was required. This was likely exacerbated by the challenge of finding a suitable space, as mentioned by several settings.

'It was good. But some of the children are very literal, so they wouldn't say some stuff even though you know they could say it, so it wasn't a true reflection of the children who were doing it.'—Practitioner

Nonetheless, with most settings reporting a positive experience, use of LanguageScreen is likely to attract settings to the programme more than it will deter.

The quality of the programme materials was broadly seen to facilitate programme delivery

Case study practitioners were consistently positive about the usefulness and quality of the programme materials, describing them as 'straightforward', 'well structured', and 'accessible'. Several mentioned appreciating that the handbook was comprehensive and easy to follow, as well as that the script was also available on the slides—so that the handbook was not required during the whole-group sessions. Having the script on the slides was also seen to encourage consistency where different practitioners were involved in delivering the session.

'Everything is so structured. For me, in my first year teaching, that's very helpful because I know what I'm going to do. The previous day I prepare all the resources and everything [...], and if not, I'll just check on the side of the slide, and then I know what to ask, particularly for questions that will challenge the children.'—Practitioner

Some practitioners mentioned that the texts provided were high quality and 'well chosen', and resources such as the listening rules, NELI puppet, and press-out characters were effective for supporting the children's behaviour and engagement. Many of the case study settings also noted that the use of digital slides made the programme much more engaging for the children than other similar programmes and that the level of engagement from the children would have been much lower without them. This was seen to be in part because the screen marked the sessions as a distinct activity, but also because it allowed children to clearly see all the visuals, as compared to reading from a book. Some settings also reported that the screen and the routine-led approach to the sessions helped to make the programme more accessible to some children with SEN.

Nonetheless, several interviewees made suggestions for how the programme could be improved, including increasing the number of activities that involved physical movement to improve child engagement and providing more pictures and videos to support vocabulary learning.

Summary of findings for IPE RQ4

As outlined above, the implementation of NELI Preschool in terms of compliance and fidelity to the programme was broadly good. Nevertheless, the context of individual settings moderated how easy they found it to deliver NELI Preschool. Overall, whilst delivering the programme schedule, children's attendance, staffing, and the level of support required by the cohort may present possible barriers to implementation and scale-up, the programme materials were broadly seen to facilitate implementation. Cost-related challenges around the requirement of a screen for delivery and payment for time spent on the training may further impede scale-up.

Factors such as the space and preparation time required for NELI Preschool do not appear to have presented significant widespread challenges for delivery. However, fitting the programme into the existing routine, linked to issues around child attendance, does appear to have created challenges.

Perceived impact

IPE RQ5: *What was the perceived impact of the intervention for participating (i) settings, (ii) practitioners, and (iii) children?*

In this section, we consider the extent to which the programme was perceived to achieve its intended objectives for children, practitioners, and settings, as outlined in the ToC, as well as any unintended consequences—whether positive or negative. These findings are important for understanding any impact beyond what the impact analysis measured, as well as providing insight into practitioner perceptions of programme effectiveness, which are likely to impact take-up independently of the impact analysis results.

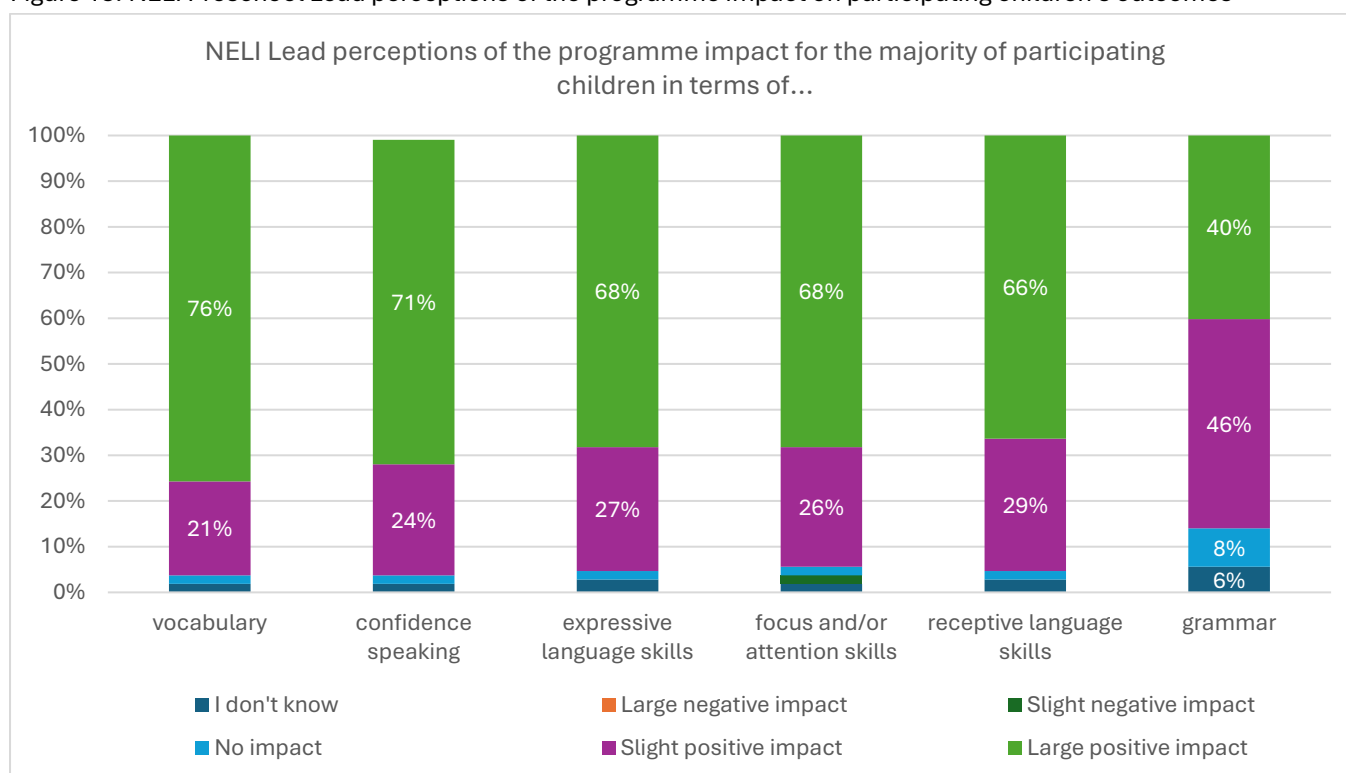
IPE RQ5 (iii): What was the perceived impact of the intervention for participating children?

- Overall, practitioners felt that NELI Preschool improved all areas of children's speech and language.
- Practitioners reported the largest impact on vocabulary, closely followed by children's increased confidence in speaking.
- Most case study settings emphasised the particular benefit of the programme for children with EAL and identified the small-group and individual sessions as the primary mechanism for the positive gains for those children with weaker language skills.

Practitioners felt that NELI Preschool improved all areas of children's speech and language, with large positive impacts perceived for children's vocabulary and confidence speaking

Overall, surveyed NELI Preschool Leads perceived the programme to have had a slight to large positive impact across a range of language outcomes for the majority of participating children (Figure 13). A large positive perceived impact was most commonly reported in relation to vocabulary (76%), which aligns with the high level of scaffolding around vocabulary that was observed during the sessions.

Figure 13: NELI Preschool Lead perceptions of the programme impact on participating children's outcomes



Source: Question L from the Endline NELI Lead Survey: 'Overall, what (if any) impact do you feel the NELI Preschool sessions have had on the majority of participating children in terms of...?' (n = 107).

Note: Data labels for proportions under 5% have been removed to support visual clarity.

These findings were reflected in case study interviews, where all settings described improvements in children's range of vocabulary and word knowledge.

'I hear the children using the language in their play which is really nice to hear. The children memorise the words, and I thought they wouldn't be able to recall them when we did the recap 5-weeks later, but they could remember every single one. We had some children who were almost mute at the start of the year; it has brought them out of their shells.'

—Practitioner

A particularly positive perceived impact was also reported for children's confidence speaking (Figure 13), and many case study settings observed that children were more willing to answer questions during stories or contribute to conversations. Most of the surveyed NELI Preschool Leads also perceived large positive impacts on children's expressive skills, focus and attention, and receptive language skills. Many case study settings likewise highlighted improvements in children's concentration and their listening skills, as well as school-readiness skills such as turn taking.

'I feel that the transition [into Reception] for these children will be so much easier because of NELI. This cohort does have low achievement, so the fact that they are so excited to learn considering the SEN and behaviour levels, is just amazing.'

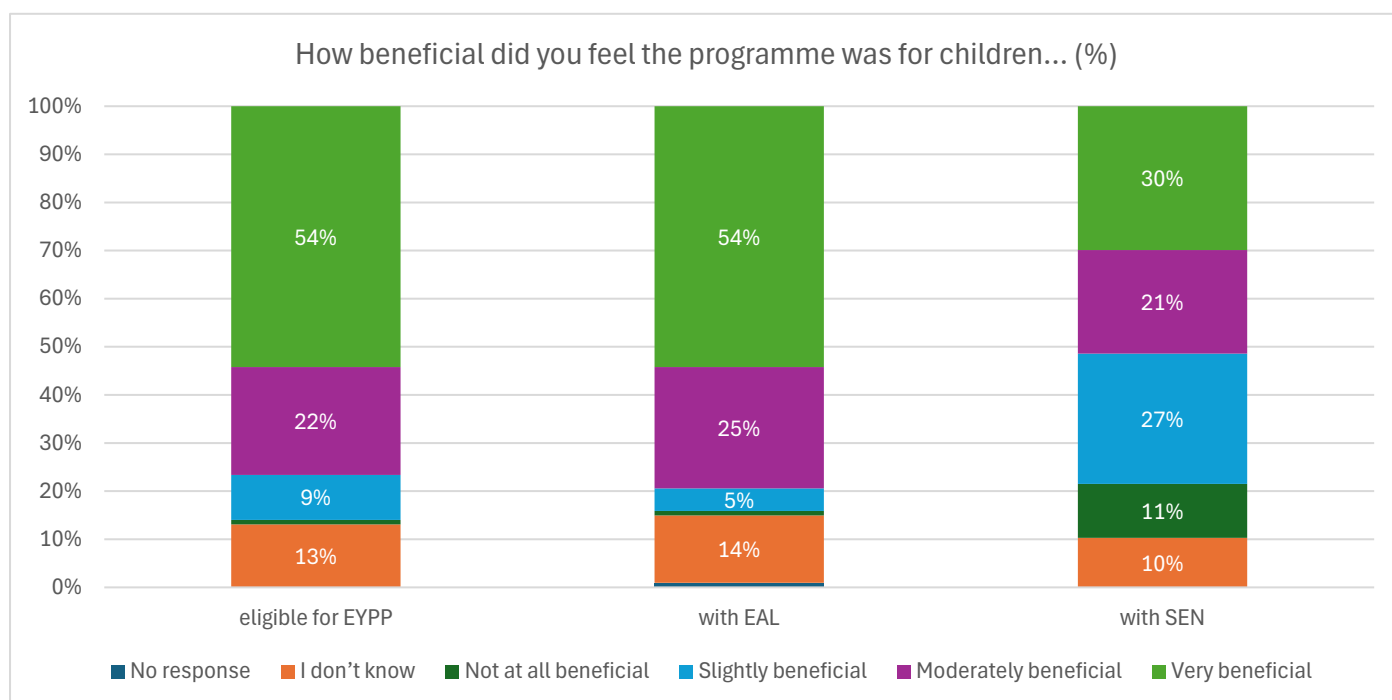
—Practitioner

Grammar was the only area where practitioners were more likely to report having perceived slight positive impacts (46%) rather than large ones (40%). This may reflect the programme's limited explicit focus on grammar. These findings are contrary to the findings in the secondary outcomes analysis which found that the programme was most effective for expressive language skills compared to receptive language skills.

Most NELI Preschool Leads perceived the programme to have been beneficial for children eligible for EYPP and/or with EAL, but views on the benefit for children with SEN were more mixed

Half of surveyed NELI Preschool Leads (54%) reported that they felt the programme had been 'very beneficial' for children facing disadvantage (for example, eligible for EYPP) (Figure 14). Very similar patterns were seen in relation to EAL. However, opinions were much more divided in relation to the extent to which the programme was seen as beneficial for children with SEN, with only a third (30%) of surveyed NELI Preschool Leads considering it to have been 'very' beneficial.

Figure 14: NELI Preschool Lead perceptions of how beneficial the programme was for particular groups of children



Source: Question M from the Endline NELI Lead survey: 'How beneficial did you feel the programme was for (1) children facing disadvantage (i.e., eligible for the Early Years Pupil Premium [EYPP]); (2) children with identified or suspected special educational needs (SEN); (3) children with English as an Additional Language (EAL)?' (N = 107)

Note: Data labels for proportions under 5% have been removed to support visual clarity.

A few settings mentioned that some children with SEN were consistently not able to engage with the whole-group sessions, so they were taken elsewhere to do a different activity during this time. One setting mentioned that they felt the whole-group sessions were too long for children with SEN. However, other settings reported strong participation among children with SEN, depending on the individual and the availability of staff to support them to access, exit, and/or rejoin the session.

RQ5 (ii): What was the perceived impact of the intervention for participating practitioners?

- Intervention practitioners demonstrated a significantly greater increase in confidence in their knowledge and ability around many areas for supporting children's language development, compared to the control group.
- Levels of motivation in relation to practices that support early language development increased significantly more among intervention practitioners compared to those in the control group.

RQ5.1: To what extent were practitioner outcomes seen to mediate child outcomes?

- Practitioners did not report perceiving their own outcomes to have been mediators for the impacts they saw on the children.

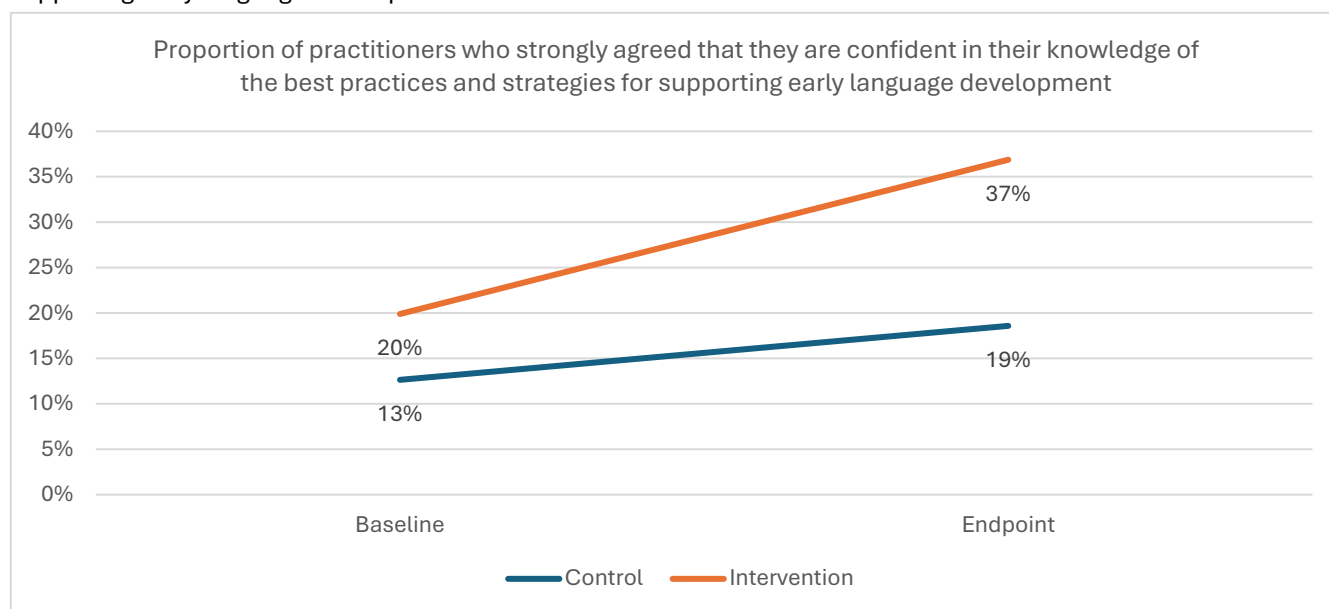
Intervention settings demonstrated larger increases in practitioner confidence in their knowledge and abilities around supporting children's speech and language compared to control

All practitioners who completed the practitioner outcomes surveys were asked at baseline and endline a number of questions about their confidence around their knowledge and abilities in relation to supporting early language development.

Compared to control practitioners, intervention practitioners demonstrated a larger increase in their confidence in relation to 'reasonable' language development goals, the best practices and strategies for supporting language development, national standards for early language skills, and assessing early language skills. This difference was statistically significant at $p < .001$. The proportion of practitioners who strongly agreed that they felt confident in these areas increased by between +1 to +6 percentage points for the control practitioners, compared to between +14 to +17 percentage points for intervention

practitioners. Figure 15 provides an example of this significant increase for the NELI Preschool group between the start of the intervention and the end in comparison to the control group in the area of confidence that saw the largest increase in the intervention group: knowledge of the best practices and strategies for supporting early language development. The only area of knowledge that did not show a statistically significant increase in confidence was in relation to the level of children's language skills when they enter their class.

Figure 15: Practitioner perceptions of the programme's impact on their knowledge of the best practices and strategies for supporting early language development



Source: Question E from the Endline Practitioner Surveys: 'I am confident in my knowledge of: The best practices and strategies for helping 3–4 year olds to develop their language skills' (Control $n = 184$ baseline, $n = 183$ endline / Intervention $n = 181$ baseline, $n = 179$ endline)

Similarly, statistically significant increases in intervention practitioners' confidence in their abilities in relation to supporting early language development were observed in comparison to the control group. This covered a number of areas, including gauging the level of children's language skills, incorporating regular opportunities to practise language skills into common preschool situations, planning activities to develop language skills, supporting language development in response to children's spontaneous comments/discoveries, helping children navigate their confusion when developing language skills, and translating assessment results into curriculum plans for groups and individuals. The proportion of practitioners who strongly agreed that they felt confident in these areas changed by between -5 and +3 percentage points for the control practitioners, compared to between +11 and +28 percentage points for intervention practitioners.

Intervention practitioners demonstrated a greater increase in motivation to support children's speech and language compared to those in the control group

Intervention practitioners were found to have increased their motivation around explicitly teaching the meaning of new vocabulary, practising new words in a range of contexts, and exploring narrative structure through dialogic reading. These are all areas that are directly targeted by the NELI Preschool programme. These reported improvements were all statistically significant when compared to the control group, with a change of between +4 and +8 percentage points in the proportion of intervention practitioners reporting to be 'very' motivated, compared to -3 and -4 percentage points for control. There was, however, no change observed in the reasons behind the intervention practitioners' motivation, with the large majority indicating at both baseline and endline that they did these things because they thought it was important for the children's intellectual development.

The proportion of intervention practitioners reporting to be 'very' motivated to provide additional scaffolded support to those children with the weakest language skills increased by 3 percentage points between baseline and endpoint, while for control practitioners the proportion decreased by 4 percentage points. This difference was not statistically significant. The smaller

change seen for intervention practitioners compared to other areas of practice may be the consequence of a higher proportion already reporting high levels of motivation at baseline (88% compared to 82–83% for the other areas of practice).

Overall, these findings align with the practitioner outcomes hypothesised in the ToC, supporting outcomes for practitioners' wider practice (not just directly in the sessions). This suggests that the programme may act as an effective form of CPD that could lead to improved child outcomes via continuous provision and not just directly through the sessions. Future research could look to understand how important delivering the sessions is for these practitioner outcomes or whether completing the training alone is sufficient.

IPE RQ5 (i): What was the perceived impact of the intervention for participating settings?

IPE RQ5.2: To what extent have the principles of and learning from the intervention become integrated into broader and/or ongoing practice within the settings?

- Most settings felt that NELI Preschool had improved their support for early speech and language.
 - Several case study settings welcomed the inclusion of LanguageScreen in their practice for gaining a better picture of achievement.
 - Almost all settings reported that they would use NELI Preschool again; however, two-thirds would opt to modify it by for example, delivering fewer sessions and/or different activities, spreading the session out across more weeks, having more children doing the targeted sessions and/or reordering the books to fit better with their curriculum, times of the year, or external events. Plans to modify were more common in settings with lower levels of deprivation.
 - For those who would modify the programme or not continue with it, the top three reasons given were challenges in fitting NELI Preschool into their schedule, the cost, or their staffing capacity.
-

Most settings felt that the NELI Preschool programme had improved their support for early speech and language

Most (78%) of the surveyed NELI Preschool Leads reported that the programme training had helped staff to improve their practice around supporting early speech and language development by 'quite a bit' or more, with a further 19% saying it had done so 'a little'. Most of the case study settings described improvements in practitioners' understanding and skills, including around developing narrative and language skills, understanding of children's needs, reflection on their practice, and their presentation skills. Several case study settings also mentioned that using LanguageScreen had provided them with a better picture of achievement by understanding children's strengths and areas of difficulty.

'[NELI Preschool has] given me the language and communication teaching that I wanted to provide for the children. I knew something was missing last year, but I wasn't sure what it was. We have never really had a programme for nursery before, bits here and bits there. I knew it needed to be something more, and NELI gave us that.'—Practitioner

Some of the practitioners explained that if it was not possible to continue with the programme, they would still use their learning to implement a similar structured approach.

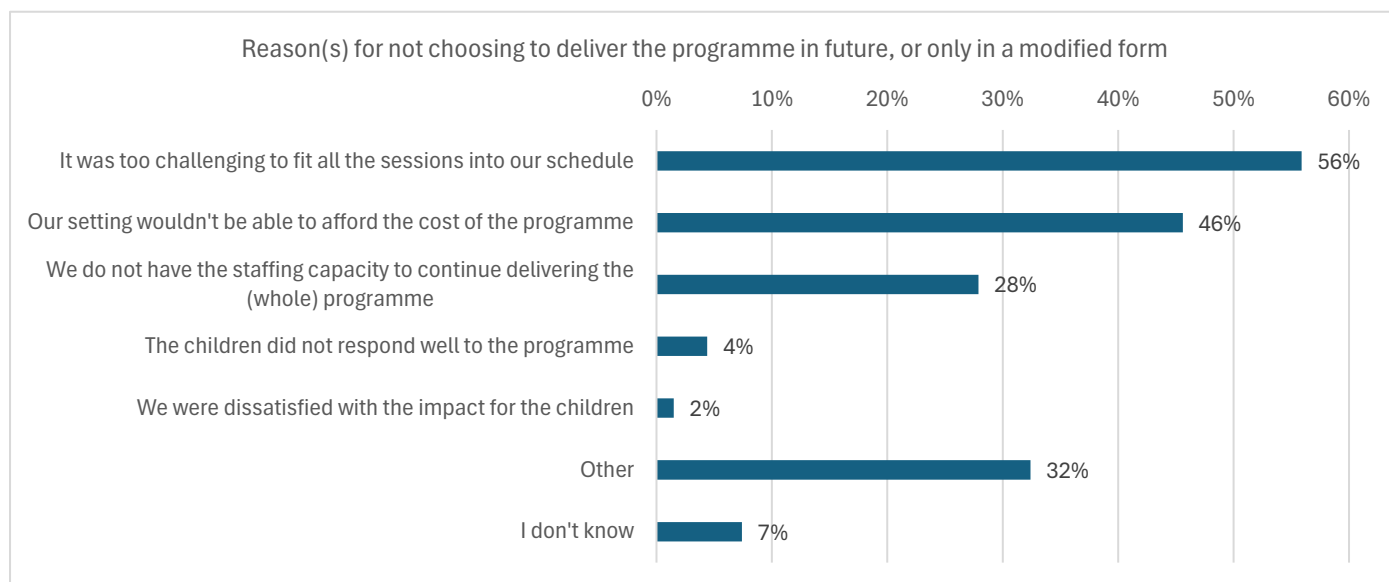
Almost all settings reported that they would use NELI Preschool again; however, most would opt to modify it

Almost all of the surveyed NELI Preschool Leads (93%) reported that they would use NELI Preschool again in future, and this was echoed by some of the case study settings as well, both in relation to the programme and LanguageScreen. However, two-thirds (59%) of surveyed NELI Preschool Leads reported that they would implement it in a modified form for example, by delivering fewer sessions and/or different activities, spreading the sessions out across more weeks, having more children doing the targeted sessions and/or reordering the books to fit better with their curriculum, times of the year, or external events. NELI Preschool Leads in less deprived settings were more likely to wish to continue delivering the programme in a modified form (70%—26 out of 37 vs 58%—29 out of 50), whereas those in more deprived settings were more likely to wish to continue delivering it as is (34%—17 out of 50 vs 22%—8 out of 37).

The most common reason for wishing to modify the programme, or preferring not to continue with the programme at all, was because it was too challenging to fit all sessions into the setting's schedule (see Figure 16). Where NELI Preschool Leads

provided 'Other' reasons, these related to their dissatisfaction with the nature of the programme (for example, it was too time-consuming or not sufficiently play-based), challenges for implementation (for example, child absence, technical issues, and level of child need), and/or the programme not being suitable for the setting's circumstances (for example, repetitive for children in a multi-age classroom or not aligned with their curriculum).

Figure 16: NELI Preschool Leads' reasons for not choosing to deliver the programme again in future, or only in a modified form



Source: Question P from the Endline NELI Preschool Lead survey: 'Please select the main reason(s) you would not choose to deliver the NELI Preschool programme in future (or would only deliver it in a modified form)?' (n = 68).

Note: Only those who indicated that they would not choose to deliver the NELI Preschool programme in their nursery again, or would only do so in a modified form, were asked to respond to this question. Respondents could select multiple responses.

IPE RQ5.3: What are the unintended consequences (positive or negative) of the intervention?

Reflecting on the information presented above, possible unintended consequences of implementing NELI Preschool may include:

- **Additional workload:** Practitioners completed training in their own time, sometimes unpaid—although the trial timeline conditions mean this is not necessarily representative of what would occur if scaled up.
- **Learning displacement:** The large amount of time required for session delivery meant that other aspects of nursery practice were negatively impacted, such as children's attendance at other interventions, other learning opportunities, or interaction time as part of continuous provision.
- **Widening language development gaps for children with SEN:** The inclusion of children with SEN in the sessions was found to vary, and SEN was the most common reason provided by settings for children being ineligible for the targeted intervention. This could result in children with SEN falling further behind their peers.
- **Widening language development gaps for socio-economically disadvantaged children:** Children eligible for EYPP were found to be more likely to have lower hours of attendance, which means they would receive lower levels of exposure to the programme and consequently receive less benefit from it compared to their peers.

While setting staff demonstrated increased confidence in relation to supporting early language development, programme participation also carried some unintended negative impacts for many practitioners in relation to workload. As outlined above, many practitioners completed the training in their own time, sometimes unpaid, and several case study practitioners reported that this was challenging to fit in around their personal lives. Constrained timelines in the trial may have led to a higher occurrence of this than would be seen outside of the trial. However, additional funds were made available to settings

to pay practitioners for time spent on the training as part of the trial, so instances of unpaid training completion may still increase outside of the trial. Administration of LanguageScreen and preparation for the sessions also represented additional workload on top of usual practice.

Similarly, while practitioners reported perceiving positive impacts of the programme for children's confidence and language development, it is possible that participating in the programme resulted in some learning displacement, as the sessions may have resulted in children missing out on other interventions, learning opportunities, and/or interaction time that could have benefitted them more and/or in other areas. Finally, some programme activities may not be suitable for some children with SEN, so it will be important for settings to consider whether their SEN children are likely to be able to participate and, if not, what alternative activities can be provided.

The finding that children eligible for EYPP were over-represented among the children with the lowest levels of setting attendance also indicates that programme implementation risks increasing the disadvantage gap, as children with the lowest levels of attendance will receive the lowest levels of programme exposure. While the expansion of the free childcare entitlement to 30 hours per week from September 2025 will likely increase the number of hours disadvantaged children in general spend in the setting, as the entitlement only applies to families where both parents are employed, the most disadvantaged children will likely continue to have the most limited opportunities to benefit from the programme. While these constraints are obviously outside the remit of what the developers or settings can control, settings may consider how they can ensure the most disadvantaged children may be given additional support to still benefit from the programme, including for example, additional targeted exposure to the programme content during continuous provision at the times they are present at the setting.

These will be important considerations for settings when deciding whether to adopt the programme in any future scale-up. However, there are measures that settings can put in place to mitigate the risk of these negative consequences, including supporting staff to complete the training during working hours, ensuring that children do not miss out on other important learning opportunities as a result of the sessions, and providing additional support to any children who are unable to access the sessions, whether this is due to additional needs or attendance patterns.

Summary of findings for IPE RQ5

Overall, the programme was perceived to have positive impacts for both participating children and practitioners. Practitioners perceived particularly strong impacts for children's vocabulary and confidence speaking, and this held for children eligible for EYPP and EAL. However, perceptions of the programme benefit for children with SEN were more mixed. Positive impacts were also observed for practitioner confidence in their knowledge and abilities in relation to supporting early language development, as well as their motivation to do so. NELI Preschool Leads perceived that this had a positive knock-on effect for practitioner practice. As it is likely that the change in practitioner practice will not feed through to child outcomes until the next round of programme delivery, it is possible that the impact on children may even increase. However, most NELI Preschool Leads indicated that they would make modifications to the programme if they were to deliver it again, which raises the question of whether the outcomes observed in this trial will be sustained.

Usual practice

IPE RQ6: What was 'business as usual'?

In this section, we will address firstly the extent to which usual practice in control and intervention settings differed prior to the trial. Next, we will look to understand the extent to which the programme offer differs from usual practice in these settings, to understand what (if any) it adds to settings' existing offer. As we are interested in understanding the suitability of the programme for both maintained and PVI settings, we will also look for any differences by setting type. Finally, we will look to understand what usual practice looked like in control settings during the trial to understand whether any change in practice compared to baseline may have confounded the impact results.

IPE RQ6.1: What was usual practice in relation to oral language development in all settings prior to the intervention? To what extent (if at all) did this differ between PVI and maintained settings?

- Overall, there were no meaningful differences between control and intervention settings in terms of existing practice before the trial. Small differences were observed in the use of their language screening tool results, but not to the extent of compromising trial results.
- Before the trial, most settings reported not using any language development programmes; however, specialist speech and language support and the use of language screening tools were more common—particularly among PVI settings.
- Before the trial, most settings had not participated in early language development CPD in the last 12 months.
- All settings were broadly comparable in their early language development practices, such as using group discussions or one to one conversations; however, maintained settings were more likely to explicitly teach new vocabulary.
- In most cases, the programme offered a variation on usual practice, although the extent of this depended on the setting.

Formalised language programmes were relatively uncommon across both intervention and control settings, but specialist speech and language support was more widespread

Before the trial, 33% of surveyed settings reported that they used an externally produced language development programme for children aged 3–4 years, with the WellComm Early Years Toolkit being the most common, closely followed by Early Talk Boost. The WellComm Early Years Toolkit provides ideas for activities settings could do with children to improve their speech and language skills, but does not constitute a structured programme (GL Assessment, 2017). Early Talk Boost is a structured programme involving regular small-group sessions, similar to the NELI Preschool ones, but there is no enrichment component, and it is much more light-touch, constituting 27 sessions across nine weeks (Speech and Language UK, 2023). As a result, while both offer additional language support for at least some children, they cannot be considered an equivalent offer to NELI Preschool given the purely targeted and short-term nature of the support they offer. These findings indicate that use of a formalised language programme is in itself relatively unusual practice within the sector.

Specialist speech and language support was, however, more common than formalised language programmes, with three-quarters of surveyed intervention (75%) and control (71%) settings reporting that all children identified as requiring additional support received specialist speech and language support. This was more common in PVI settings (78%) than in maintained settings (69%).

Screening practices were more common, and the results were used slightly differently by intervention and control settings

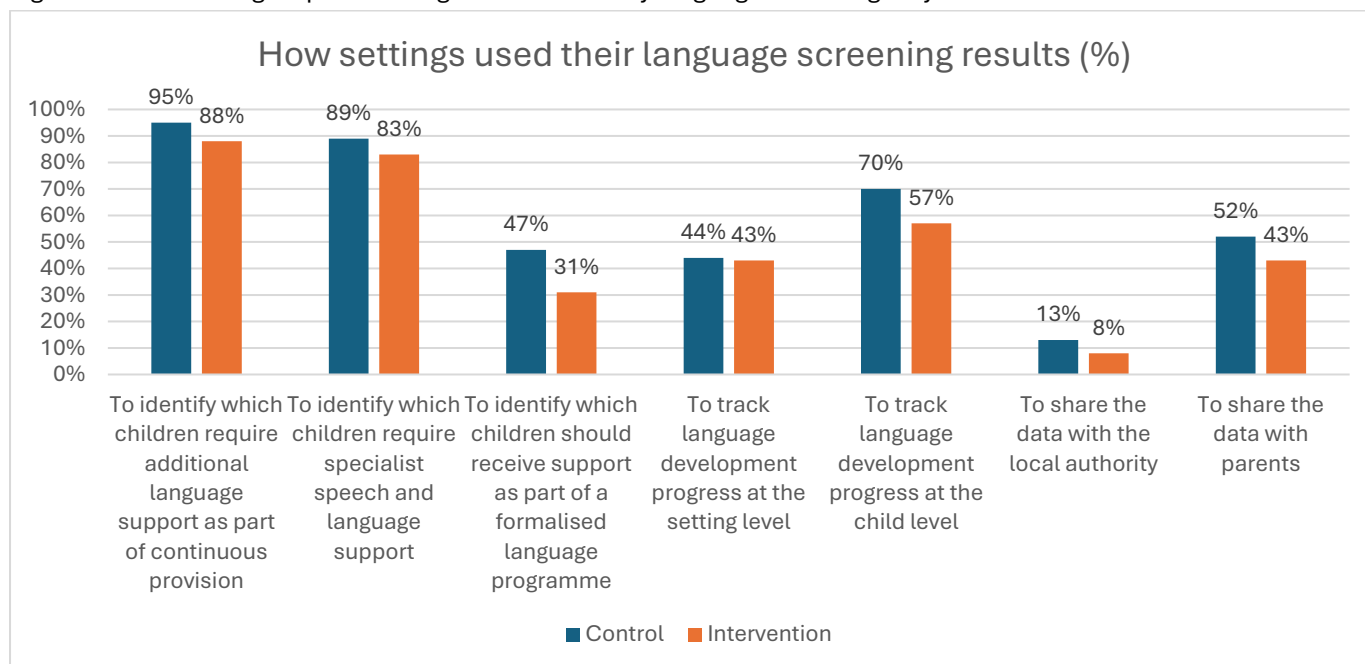
A slightly higher proportion of settings (52%) indicated that they used a language screening tool rather than a formalised language programme, with WellComm again being the most common. The WellComm assessment is quite comparable to LanguageScreen, suggesting that where screening did already take place, the LanguageScreen component of the programme was unlikely to offer anything new. There was no notable difference between control and intervention groups in terms of their use of screening tools.

Of the surveyed settings that reported using screening tools at baseline, nearly all (92%) reported using the results to identify where children required additional language support for continuous provision, and most (86%) to identify children requiring specialist support (Figure 17). Over half (63%) also reported using the results to track individual children's progress.

Control settings were slightly more likely than intervention settings to report using the results of any screening they do to identify which children should receive support as part of a formalised language programme and/or to track language development progress at a child level. These are two of the key ways in which the NELI Preschool programme promotes the use of the language screening results in order to support children's language development. However, the differences are

sufficiently small that this is very unlikely to have biased the impact evaluation results. Moreover, while tracking individual language development is relatively widespread, NELI Preschool is unusual in employing a screening tool to select children for a formalised language programme.

Figure 17: How settings reported using the results of any language screening they did before the trial



Source: Question E from the Baseline BAU Survey: 'How does your nursery use the results of the screening? Please select all that apply.' (Control $n = 64$, Intervention $n = 77$)

Most settings indicated that they had not participated in CPD related to early language development in the previous 12 months

Before the trial, 39% of surveyed settings reported having participated in any internal or external CPD opportunities relating to early language development in the previous 12 months. This was the same for maintained and PVI settings as well as control and intervention groups. Of these settings, 55% reported that the training had lasted half a day and 29% that it had lasted more than one day. This CPD was most commonly delivered in person (41%), with a further 30% being delivered online in real-time, and 17% online but asynchronous. In-person training was much more common in PVI settings than in maintained ones (51% vs 27% of settings). Most of the CPD (80%) was delivered by an external specialist, particularly in PVI settings (93% vs 72% in maintained settings). There was no meaningful difference between control and intervention groups in terms of the duration and nature of CPD opportunities.

These findings indicate that the NELI Preschool training was relatively unusual in duration and format compared to the CPD around language development that settings were otherwise doing. Less than a third (31%) of surveyed intervention settings reported having completed training of equivalent length in the previous 12 months, and less than a quarter (22%) had completed training in an equivalent format. These findings also demonstrate that most settings draw on external specialists for CPD around language development, as in the case of NELI Preschool.

All settings regularly used practices similar to those employed by the programme, although explicit introduction of new vocabulary occurred less often

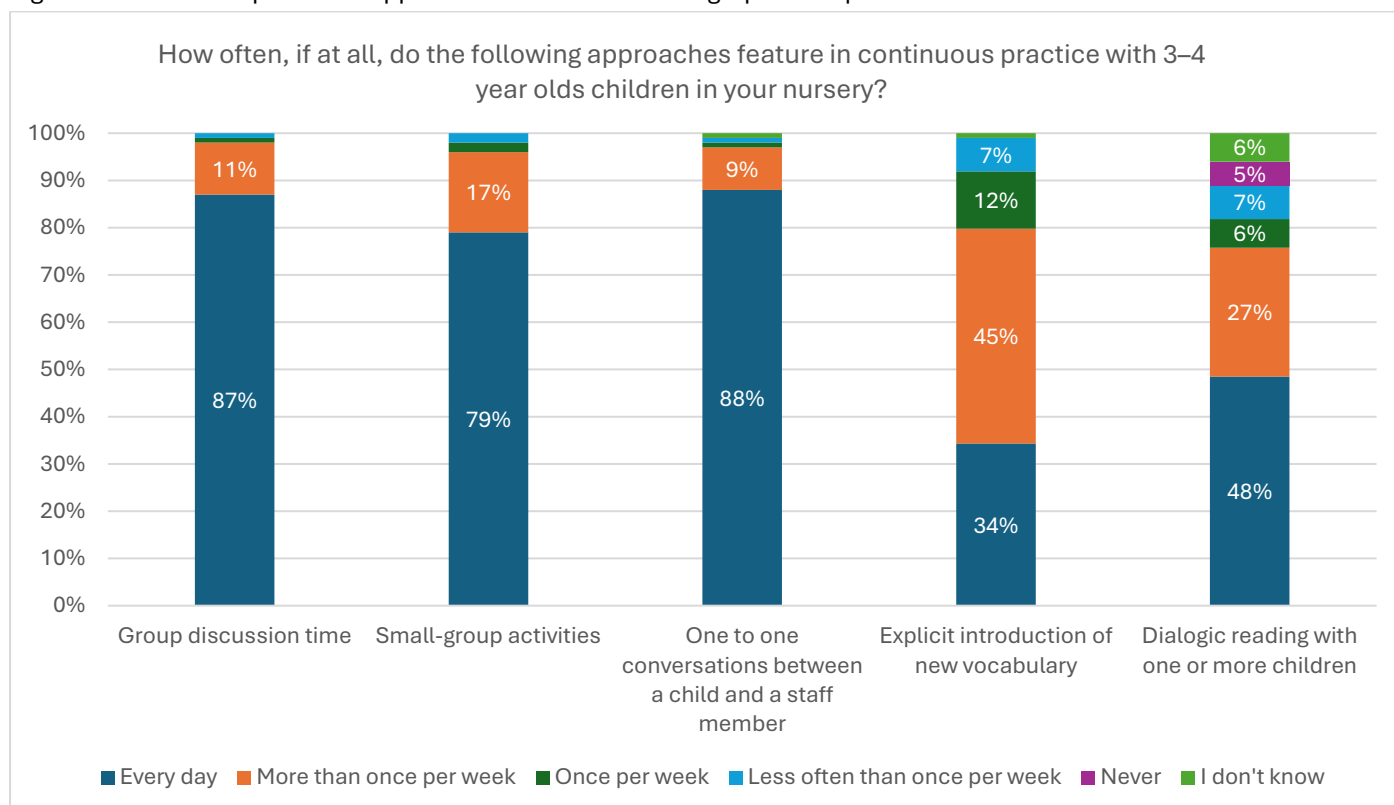
Settings were asked at baseline to indicate how often they already did some of the practices that constitute core components of the NELI Preschool programme on a daily basis. As Figure 18 shows, nearly all settings reported using discussion time, small-group activities, and one to one conversations between children and staff occurring more than once a week to develop children's language skills, and for most (79 to 88%), this was daily. However, only a third (34%) of settings indicated that they explicitly introduced new vocabulary to children daily, although this proportion was higher in maintained

settings (38%) than in PVI settings (28%). No differences were observed between the intervention and control groups in either of these cases.

The emphasis on vocabulary may consequently constitute a key component of the NELI Preschool offer, particularly given the emphasis placed on this as an outcome by NELI Preschool Leads at the end of the trial (see Perceived impact).

While only 48% of settings reported doing dialogic reading every day, the erratic spread of data suggests that there may have been some confusion around the term 'dialogic reading', which may have resulted in an under-representation of the prevalence of this practice. Importantly, there were no differences between the control and intervention groups in terms of their practices for supporting early language development.

Figure 18: How often particular approaches featured in settings' practice prior to the trial



Source: Question F in the Baseline BAU Survey: 'How often, if at all, do the following approaches feature in continuous practice with 3–4-year-old children in your nursery?' (n = 284)

Note: Data labels for proportions under 5% have been removed to support visual clarity.

Most settings (88%) also reported that most or all of their 3–4-year-old children had the opportunity for an active discussion with a staff member on a daily basis, and most settings (73%) reported that all children who were identified as requiring this additional support received it.

In most cases, the programme offered a variation on usual practice, although the extent depended on the setting

These findings suggest that many of the core components of the programme were already usual practice in settings, and half (49%) of the surveyed NELI Preschool Leads described the programme as 'moderately' different from the usual speech and language support they provide. However, a further 39% described it as 'very' or 'extremely' different, suggesting that for some settings the practices offered by the programme provided something new. Furthermore, whilst a few settings indicated that NELI Preschool was similar to other story-based teaching they had done, others described differences in the emphasis the programme places on reading, its structured approach, and the increased focus on specific vocabulary. Whereas NELI Preschool focuses on one book per week, one case study setting said that they would switch books more often than this; however, another said that they would usually focus on one book every two weeks instead. Overall, the NELI Preschool programme appears to have been in line with usual practice within settings while still offering something new.

'I've worked in nurseries now for 17 years, and I've not seen anything quite like this. I think it's amazing. It's not that something new has been invented; it's just the way it's been brought together.'—Practitioner

IPE RQ6.2: What was usual practice in control settings in relation to language enrichment during the intervention?

- Slightly more control settings delivered a language development programme during the trial than they did before the trial, suggesting that some may have looked to compensate for missing out on NELI Preschool.
- However, control setting practice around language screening, specialist speech and language support, and practices to support language development as part of continuous provision remained unchanged.
- There were some practitioners in both groups who had had previous experience with NELI for Reception, resulting in some contamination of the sample. A sensitivity analysis excluding settings with these practitioners produced a higher effect size than the primary analysis, suggesting that prior NELI experience in some settings may have led the main results to underestimate the intervention's impact.
- During the trial, a minority of control settings reported delivering another language development programme, representing a small increase from baseline and suggesting some compensatory rivalry; however, sensitivity analyses excluding these settings produced impact estimates similar to the primary analysis, indicating no detectable bias.

Slightly more control settings delivered a language development programme during the trial compared to baseline

Less than half (41%) of surveyed control settings reported using a language development programme during the intervention window. However, this is a slight increase from baseline (33%), which suggests that some of the control settings may have proceeded to implement other language development programmes in place of NELI Preschool. No equivalent increase was seen for the proportion of intervention settings (32%) who delivered another language development programme in addition to NELI Preschool during the trial window. This means that control settings might have compensated for the lack of taking part in NELI by participating in another language development programme. Therefore, as outlined in the impact evaluation section, a sensitivity analysis was run to address potential threats to internal validity. The results of this analysis indicated that changes to participation in other language development programmes among control group settings did not introduce detectable bias into the impact estimates (see Post-hoc sensitivity analyses in Impact evaluation results for further details). As at baseline, the most common programme was the WellComm Early Years Toolkit.

Surveyed control settings' use of specialist speech and language support remained unchanged, with three-quarters (77%) reporting that all children identified as requiring additional support received it.

Use of screening tools and their results was broadly the same in control settings during the intervention window as before it

During the trial window, 52% of surveyed control settings used a language screening tool with children aged 3 to 4 years, with WellComm being the most common. This is comparable to baseline. Some intervention settings likewise reported ongoing use of WellComm alongside LanguageScreen, although some also reported replacing WellComm with LanguageScreen.

The majority of surveyed control settings (88%) reported using screening tool results to identify where children require additional language support as part of continuous provision or, to a lesser extent (70%), specialist support. Over half (55%) also reported using the results to track the children's progress at an individual level. Overall, these are the same patterns observed in the use of screening tools and their results as at baseline. However, the proportion of control settings using the screening tools for these purposes was consistently lower at endline and closer to what was seen among intervention settings at baseline. While the baseline data raised concerns that control settings were more likely than intervention ones to already have similar screening practices to the programme already in place, these endline findings indicate that no significant bias occurred during the trial in this respect.

Control setting practice around early language development was the same during the trial as before it

As at baseline, nearly all control settings reported using group discussion time, small-group activities, and one to one conversations between children and staff more than once a week to develop children's language skills, and for most this was daily. It was less common for settings to explicitly introduce vocabulary to children every day, as at baseline, and trends relating to dialogic reading were also similar. A comparable number of settings indicated that most or all of the children had been given the opportunity to have an active discussion with a staff member of the nursery on a daily basis during the trial window as compared to baseline. This was likewise the case for the reported proportion of children identified as requiring specialist speech and language support who then received it. This illustrates that practice in control settings during the intervention window was representative of the practice observed in all settings at baseline.

There were some practitioners in both groups who had had previous experience with NELI for Reception

A significant minority (19%) of the surveyed practitioners at endline reported having completed training for NELI for Reception in the past. This was slightly more common in the control group (22%) than in the intervention group (16%). A slightly lower proportion (13%) reported having previously delivered or supported delivery of NELI for Reception sessions. Again, this was slightly more common in the control group (16%) than in the intervention group (10%). Given the similarities between NELI Preschool and NELI for Reception, and the evidence to indicate the impact of NELI for Reception training and delivery on staff knowledge and skills (Dimova *et al.*, 2020), these findings indicate that some amount of contamination in terms of control group practice may have occurred. It is possible that those who had experience with NELI for Reception may already have embedded some of the practices and approaches common to both programmes within their usual practice, bringing it closer to practice in intervention settings during the trial. However, the baseline BAU survey results show that many practitioners across both intervention and control were already employing similar approaches to NELI Preschool prior to the trial. Moreover, the results of the mediation analysis indicate that practitioner outcomes were not a significant vector for child outcomes, which suggests that child exposure to the programme material is more important for impact than broader practitioner practice. As a result, the risk of this contamination affecting the results of this trial is deemed to be minimal.

It should be noted that 17% of the control practitioners surveyed at endline reported that they had been involved in NELI Preschool training as a part of the trial, and 7% reported that they had been involved in delivery. However, follow-up contact with some of these respondents indicates that this was in fact a representation of the LanguageScreen preparation and administration they had done as part of the trial, rather than an indicator of contamination. Nonetheless, as a response was not received from all respondents concerned, there still remains a small risk of contamination having occurred.

Two control practitioners reported having completed NELI Preschool training outside the context of the trial, and two each of intervention and control practitioners reported having delivered or supported delivery of NELI Preschool sessions outside of the trial. The size of these numbers and the relative parity between intervention and control means that the risk of this contamination meaningfully altering the results of this trial is minimal.

Overall, given the risk of contamination to the control group from several areas (increased delivery of language development programmes, prior experience with NELI, and possible exposure to NELI Preschool during the trial), there is a small risk of bias having been introduced that would have resulted in the trial's findings representing an underestimate of the intervention effect. However, this was an effectiveness trial, and the prevalence of NELI for Reception in recent years as a result of government funding should now be considered business as usual.

Summary of findings for RQ6

BAU in intervention and control settings was very similar prior to the trial, and the control condition held during the trial with only a small number of control settings changing their practice around supporting language development compared to baseline. The programme offer was also confirmed to be distinct from BAU, with both formalised language programmes and CPD in relation to language development found to be relatively uncommon. At the same time, many of the practices involved in the programme were reported to be a common part of continuous provision BAU, suggesting that the programme was aligned with existing practice, which is likely to facilitate adoption.

Cost

This section provides an overview of the amount of time required for programme delivery in the context of the trial, as well as an estimate of the time that would be required were the programme to be delivered fully as intended. The financial costs of delivery will then be outlined, with clarification around which of these costs settings were exempt from as part of the trial. Details about the sources, methods, and assumptions behind these calculations can be found in the Costs methods section. Table 40 **Error! Reference source not found.** outlines the average staff time spent on the programme per staff member per setting as part of the trial.

Table 40: Time spent on NELI Preschool delivery

		Year 1		Year 2		Year 3	
		Mean number of staff members (min, max)	Mean number of hours per setting (min, max)	Mean number of staff members (min, max)	Mean number of hours per setting (min, max)	Mean number of staff members (min, max)	Mean number of hours per setting (min, max)
Training	Online training course	4 (0, 11)	31.5 (14, 48)	1	8.5 (3, 12)	1	8.5 (3, 12)
	Webinars	2 (1, 3)	2.75 (1, 3.5)	0	0	0	0
	Delivery Support Hub	2 (0, 5)	2.75 (0, 6)	2 (0, 5)	2.75 (0, 6)	2 (0, 5)	2.75 (0, 6)
Preparation	LanguageScreen	1 (1, 2)	2.5 (1, 6)	1 (1, 2)	2.5 (1, 6)	1 (1, 2)	2.5 (1, 6)
	Set-up	3 (1, 4)	5.25 (2, 12.5)	3 (1, 4)	5.25 (2, 12.5)	3 (1, 4)	5.25 (2, 12.5)
	Session preparation	3 (2, 4)	67.25 (60, 80)	3 (2, 4)	67.25 (60, 80)	3 (2, 4)	67.25 (60, 80)
Delivery	Whole-group session delivery	3 (1, 5)	27.5 (11.75, 51.25)	3 (1, 5)	27.5 (11.75, 51.25)	3 (1, 5)	27.5 (11.75, 51.25)
	Small-group session delivery	2 (2, 4)	14.25 (0, 31.75)	2 (2, 4)	14.25 (0, 31.75)	2 (2, 4)	14.25 (0, 31.75)
	Individual session delivery	2 (1, 4)	8 (0, 62.75)	2 (1, 4)	8 (0, 62.75)	2 (1, 4)	8 (0, 62.75)

Note: All amounts of time in this table have been rounded to the nearest quarter of an hour (i.e., 0.25). Personnel data has been rounded to the nearest one person.

Sources: Case study setting cost pro formas (n = 7), Endline NELI Lead Survey (n = 107), training completion data (n = 151)

On average, each setting spent 112 hours on training and preparation for the sessions, in addition to 49.75 hours of session delivery. Preparation for the sessions was the most time-consuming single element, with all settings that provided a cost pro forma reporting between three and four hours spent on preparation each week. The sample of practitioners involved in the programme covered a range of different roles and qualification levels, with no clear trend in relation to the nature of their involvement (see Compliance and fidelity section for more details). These calculations do not include time for integrating elements of the programme into continuous provision, as this was an optional and highly variable extension of the programme. Given the high rates of turnover per setting and the aim for as many practitioners as possible in the setting to complete the training, we have assumed that one additional staff member will need to complete the training course each year. As an average of 1 in 7 PVI staff members and 1 in 14 maintained setting staff leave their setting each year (DfE, 2025a), accounting for an additional one staff member trained each year is still a conservative estimate.

Time required for set-up may reduce in Years 2 and 3 of delivery, but settings will still have to select the children and schedule the sessions anew each year. However, they may save time on preparing the materials. Likewise, preparation time might reduce slightly in Years 2 and 3 of delivery if the same practitioners are involved in their delivery, but this will not necessarily be the case, and some level of preparation for each session would still be required.

What is shown in Table 40 represents what happened in the trial, which in some cases is different to the time required for intended practice—as shown in Table 41. The time requirement for intended practice assumes only one practitioner being involved in each instance of each activity. The time per setting in the trial is calculated by multiplying the mean time spent and the mean number of practitioners involved, based on Table 40. The reported time spent on session preparation and delivery in the trial is higher than the amount for intended practice. This may reflect overestimates of the proportion of each session type each practitioner delivered and/or the fact that in practice more than one practitioner is often involved in session preparation and delivery.

Table 41: Time requirements for intended programme delivery

Activity	Intended practice	Time requirement per setting (hours)—intended practice	Time per setting (hours)—in the trial
Webinars	At least one practitioner attends each webinar	At least 3.5	2
LanguageScreen	Administered to all children at start and end of delivery*	5	2.5
Session preparation time	20 weeks of delivery, average of 3.5 hours of session preparation per week	60.5	77.25
Whole-group session delivery	100 sessions of between 15–20 mins	29.25	38.25
Small-group session delivery	120 sessions of between 10–15 mins	25	15.5
Individual session delivery	120 sessions of between 5–10 mins	15	16

Note: All amounts of time in this table have been rounded to the nearest quarter of an hour (i.e., 0.25). Time spent on the online training course and the Delivery Support Hub are not included in this table because there were no expected requirements around the amount of time spent on these activities as part of intended practice.

**LanguageScreen was administered at endline by NFER's Test Administrators for the purpose of the trial, but delivery outside of the trial would require the practitioners to re-administer LanguageScreen once completing programme delivery.*

As outlined in Table 42, implementing the programme would cost each setting approximately £2,682.80 over 3 years. Assuming 22 children per setting, this means the programme costs £40.65 per child per year over this period. The large majority of this is driven by the cost of purchasing the programme materials and digital resources. The cost of the annual digital resources subscription in Table 42 is for a setting with one class/group/room. For a larger setting, the price of this would be £660. However, it is likely that the relative increase in the number of children in that setting would result in a similar cost per child per year over three years.

Not all of these costs were incurred by all settings. For example, only 60% of settings needed to be supplied with a tablet as part of the trial. Different settings also took different approaches to some of these costs. Four of the seven settings that completed cost pro formas indicated that staff completed the training in paid time on top of their usual working hours, with an average (mean) of 16.5 hours per setting. As settings were paid £120 per practitioner for up to five practitioners to complete the training as part of the trial, it is likely that staff will not so commonly be paid to complete the training in their own time were the programme to be implemented outside of the trial. However, as settings may still need to pay for cover even for staff completing the training during their usual working hours, we have calculated the cost of staff time for training based on the average (mean) time spent on the training per setting, regardless of whether this occurred within or outside of their usual working hours.

We have assumed that wear and tear for items purchased will be sufficiently minimal to not require any replacement purchases over the three-year period.

Table 42: Cost of a setting delivering NELI Preschool

Item	Type of cost	Cost (average per setting in this trial)	Total cost over 3 years	Total cost per child per year over 3 years
Staff time for completing the training	Running cost per setting	£404.16 (£107.36 for subsequent years)	£618.88	£9.38
Tablet	Start-up cost per setting	£125	£125	£1.89
Programme materials	Start-up cost per setting	£500	£500	£7.58
Annual digital resources subscription	Running cost per setting	£460	£1,380	£20.91
Other—extra copies of the texts, storage for the materials, additional props	Start-up cost per setting	£23.79	£23.79	£0.36
Other—printing, photocopying, internet	Running cost per setting	£11.71	£35.13	£0.53
Total		£1,524.66	£2,682.80	£40.65

The one prerequisite for delivery was having a screen in the setting that could be used for whole-group session delivery (Table 43). In the trial, these screens were purchased for 36% of the settings. Outside of the trial, this would constitute an additional start-up cost of £199, resulting in an increase in the per child cost over three years to £43.67.

TV trolleys and wireless TV keyboards were also provided to 30% and 17% of settings, respectively, as part of the trial. However, these were optional and specific to particular set-ups within the settings. If a setting were to purchase a TV trolley and keyboard outside the trial, this would constitute an additional start-up cost of £95.98, increasing the per child cost over three years to £42.10. If a setting required a screen, TV trolley, and keyboard, this would increase the start-up cost by £235.84, resulting in a cost per child over three years of £44.22.

Table 43: Prerequisites for the implementation of NELI Preschool

Year 1			
	Quantity required (mean)	Mean additional quantity required (min, max)	% of settings which purchased items
Screen	1	1	36%

Source: OxEd administrative data

The cost of the programme for settings involved in the trial was very different to usual practice. The programme materials and digital resources were provided free, as were any tablets, screens, TV trolleys, or wireless TV keyboards required. In addition, settings were provided with the funds to pay up to five practitioners £120 each for the time taken to complete the training. The average cost of a practitioner completing the training, based on mean time taken multiplied by median hourly pay (Flemons and Worth, 2025), is £107.36, and 90% of settings had five practitioners or fewer complete the training. This means that in most cases, there was no additional cost to settings for providing staff with additional paid time to complete the training. This means that in the context of the trial, the programme cost an average of £35.50, or £1.61 per child for that year.

Table 44: Cumulative costs of NELI Preschool (assuming delivery over three years)

	Year 1	Year 2 (cumulative)	Year 3 (cumulative)
NELI Preschool	£1,524.66	£2,103.73	£2,682.80

While the cost per child of NELI Preschool is very low, cost was still raised as a significant barrier to continuing delivery (see Context and moderators). Moreover, the required time commitment may appear expensive compared to other Communication and Language Approaches in the EEF Toolkit, which are also rated as very low cost, with few direct financial requirements beyond optional resources. Toolkit estimates indicate that additional materials such as discussion books typically cost £10–£20 per child, although staff professional development can raise these costs.

Conclusion

Table 45: Key conclusions

Key conclusions

1. Children in NELI Preschool settings made an average of two additional months' progress in their oral language skills compared to children in control settings. This result has a high security rating.
2. Children eligible for Early Years Pupil Premium (EYPP) also made an average of two additional months' progress in their oral language skills compared to children in control settings. These results may have lower security than the overall findings because of the smaller number of children.
3. Children receiving additional targeted support made around two additional months' progress. This was similar to the progress of children receiving only whole-class enrichment, suggesting that targeted sessions help children with the weakest language skills benefit fully from whole-class provision.
4. Most practitioners reported that the NELI Preschool training was useful and prepared them well to deliver the programme effectively and perceived the programme to have a strong positive impact on children's language development—including vocabulary development, confidence in speaking, and engagement.
5. Practitioners in NELI Preschool settings demonstrated improvements in their confidence, practice, and motivation to support early language development compared to practitioners in control settings.

Impact evaluation and IPE integration

Evidence to support the ToC

Analysis from this effectiveness trial provides broad support for the core assumptions of the NELI Preschool ToC. IPE findings demonstrated that the NELI Preschool programme was delivered across a much larger and more diverse group of settings than in the efficacy trial, including a significant proportion of PVI settings. The programme was largely delivered as intended, with high rates of training course completion, widespread use of the structured teaching materials and recommended strategies in the sessions, and high compliance with the guidance in selecting children for the targeted component. These processes align well with the hypothesised causal pathway in which high-quality interaction and structural vocabulary enrichment lead to improvements in children's oral language skills. Findings from the impact evaluation reinforced this, with the programme producing a positive effect on the primary outcome of latent oral language skills and larger effects seen in fully compliant settings (see full details in the Interpretation section below). This suggests that the intended mechanisms of change were functioning as expected.

Additional targeted support for children with weaker language skills was delivered with some variation. While individual sessions were delivered broadly in line with expectations, with high levels of child attendance, it appears that fewer small-group sessions were delivered than planned, and these sessions may have involved slightly larger groups of children than recommended. Practitioners perceived the targeted element to offer participating children the opportunity to make progress at a faster rate than those participating in just the enrichment, including by helping to better engage with the whole-group sessions. The sessions were seen to be particularly beneficial in this respect for children who were reluctant to speak in the larger group, as it gave them the chance to practise speaking outside that context and provided them with the opportunity to familiarise themselves with the content ahead of the whole-group session, which supported their confidence to then contribute to it. These findings support the hypothesised function of the targeted intervention as presented in the ToC. Interestingly, impact estimates for the targeted subgroup were slightly higher than for the enrichment-only group, which also supports the ToC and conclusions of the IPE. This is in contrast to the findings from the efficacy trial, where the

targeted intervention appeared to have less impact than the enrichment component. Note that despite the slightly higher impact estimate for the targeted children, our estimate of the confidence interval for the differential effect between targeted and enrichment-only children included zero, suggesting that targeted sessions help children with poor language skills engage with and benefit from whole-group activities, but do not necessarily confer additional benefit in their own right. Overall, the balance between whole-class and targeted activities is probably about right to allow all children to progress at a similar rate regardless of prior ability.

The pattern of results across the secondary outcomes aligns with the programme's ToC, which places strong emphasis on vocabulary building, narrative development, and scaffolded language production. Impact evidence was particularly strong for expressive language outcomes: expressive vocabulary subtest of LanguageScreen, RAPT information, and RAPT grammar. This mirrors practitioner perceptions of the largest impact being in relation to vocabulary, as well as the consistently observed support for vocabulary development in line with the programme guidance, and the results from the BAU surveys that suggest daily explicit teaching of new vocabulary is one of the primary ways in which the programme differs from usual practice.

Some areas of the ToC were only partially supported and could be considered further for refinement. Engagement with the community of practice via the webinars and Delivery Support Hub was very low, indicating that these elements were not seen to be valuable. If OxEEd feels that these are crucial elements of the ToC, they may wish to consider how to make them more useful to practitioners. Alternatively, they could be removed from the ToC to function as an optional element alongside parental engagement and integration of NELI Preschool practice into continuous provision. A similar case can be made in relation to the Narrative Progress Tracker.

In relation to the digital slides, if OxEEd feels that their use is a core component, settings may need to be provided with additional information around the added value of this approach as compared to using just the handbook for delivery, as settings have shown an inclination to do. Alternatively, the handbook could be expanded so that it can be used instead of the slides without children missing out on any of the programme material.

There are also some areas of the ToC that could benefit from additional support to achieve greater fidelity to intended practice. Both practitioners and children were found to be less confident with the programme elements around narrative structure. Practitioners could also benefit from additional support around facilitating discussion over one-word answers in the whole-group session, as well as engaging reluctant speakers, to support more consistent practice at scale.

Child attendance patterns, particularly among EYPP children, were anticipated in the ToC to be a key moderator, and this has been confirmed by this trial. Those children with lower hours of attendance, among which those eligible for EYPP are over-represented, attended fewer enrichment sessions and were often deemed ineligible for the targeted intervention on the basis of the attendance-related criteria. While the effect found for EYPP children was equivalent to that of their peers, the high rate of targeted support in this group may conceal a group of among the most disadvantaged children with lower levels of progress than their peers. In order to avoid a widening of the disadvantage gap, it is possible that the programme may need to identify adaptations that allow it to provide additional support to children in need but with lower rates of attendance. However, with the expansion of the free childcare entitlement from 15 to 30 hours per week as of September 2025, it is also possible that this concern may become less relevant.

Conversely, children with EAL received more consistent exposure and were also more likely to meet targeting criteria, which may help explain the stronger subgroup effects for them. It is also possible that the programme is particularly effective for EAL children, who are likely to be facing slightly different language development challenges compared to their peers.

Overall, the ToC is broadly supported, but the evaluation indicates specific areas where assumptions require refinement—particularly around delivery conditions, dosage, and the variability of children's attendance.

Interpretation

The findings from this effectiveness trial are consistent with the previous evidence that NELI Preschool has a positive impact on children's early oral language skills. The trial set out to assess whether NELI Preschool, when implemented under real-world conditions and at scale across both maintained and PVI settings, could have a positive impact on children's receptive and expressive language skills. Taken together, the results suggest that the programme has a modest yet

meaningful impact on children's language development, particularly on expressive language skills. These impact results are best understood in conjunction with the implementation findings, which showed generally high fidelity to the core components of the programme. However, the IPE findings also indicate that more than half of settings (59%) would choose to modify elements of the programme, most commonly to address how challenging they found it to fit all the sessions into their schedule, so 'real-world' impact levels may be lower than those seen in the trial. The cost of the programme, even without the additional costs of purchasing a screen or paying staff to complete the training, was seen to be prohibitive for nearly half (46%) of settings, so financial constraints are likely to constitute a significant barrier to scale-up.

The primary analysis showed that NELI Preschool has a positive effect on children's language skills with the effect size of 0.10. This effect is slightly lower than that seen in the earlier efficacy trial, which is expected under effectiveness-trial conditions where implementation support is lighter, settings vary widely in structure, and delivery occurs at scale. The intervention showed its largest effects on expressive language measures: the effect size for LanguageScreen Expressive Vocabulary was 0.11, the effect size for RAPT Information was 0.14, and the effect size for RAPT Grammar was 0.15. These results suggest that NELI Preschool is particularly effective at improving children's ability to produce language. This is supported by the IPE evidence, which found that practitioners consistently implemented high-quality vocabulary scaffolding, repeated modelling, and structured opportunities for children to speak—areas of practice that were observed with higher fidelity than narrative-related activities, perhaps reflecting why expressive, rather than receptive, measures demonstrated the largest gains.

The subgroup analyses also provided useful insight. The effect size for those who were selected for the targeted component was 0.12, which is slightly higher than the effect size for the enrichment-only subgroup which was 0.10. This suggests that the additional support received by this group helped them to keep up with the rest of the class, or even slightly narrow the gap. This view is supported by the interaction analysis, which found no statistically significant differential impact between targeted and enrichment-only children. This means targeted children experienced similar overall impact to enrichment-only children despite starting at much lower baseline levels. Note that this finding contrasts with the somewhat surprising finding from the efficacy trial which found a lower effect for the targeted children exposed to the targeted and whole-class components than for those exposed just to the whole-class component. The effect found in the targeted group is marginally lower than the effect size observed for those who should have been targeted (ES: 0.14). This parity of magnitude indicates that practitioner overrides of the ideally targeted group were, on balance, appropriate because there is no evidence that these decisions disadvantaged children who should have been selected but missed out on the targeted component. The IPE findings help explain this, as practitioners reported applying their professional judgements when they selected children for the targeted component. This was often due to a change in understanding of children's needs or their attendance patterns to accommodate targeted sessions.

There was no evidence of an interaction between setting type (PVI) and the intervention, meaning that NELI Preschool worked similarly well in PVI and maintained settings. However, the IPE findings suggest that different setting types may still face different constraints and opportunities, resulting in different experiences for their children. PVI settings found providing suitable space for the targeted sessions and scheduling sessions around children's attendance patterns more challenging and had a stronger tendency to use the handbook instead of the digital slides. Maintained settings, however, often delivered shorter sessions and invited fewer children to respond to a question, but were also more likely to deliver targeted support to children outside of the targeted group, potentially resulting from a perception that there were additional children who would have benefitted beyond those selected by the guidance. Differences were also found in usual practice—PVIs tended to do more screening, have more specialist in-person language development training, and provide more specialist speech and language support, while maintained settings provided more consistent vocabulary support.

There was no evidence of an interaction between EYPP status and the intervention, indicating that children eligible for EYPP benefitted from NELI Preschool to a similar extent as their peers. This suggests that the programme's impacts were consistent across socio-economic groups, though the IPE findings highlighted attendance at the setting as a major challenge for children eligible for EYPP. They attended substantially fewer whole-group sessions and were often ineligible for targeted sessions (due to low attendance), limiting their exposure to the programme, although this doesn't seem to have resulted in a smaller effect overall for EYPP children. One possible explanation for this apparent contradiction is that EYPP children with sufficient attendance to meet the targeted group eligibility criteria were more likely to receive the targeted intervention which may have offset the smaller impact for the EYPP children with lower attendance levels.

There was a similar story for EAL with no evidence to suggest there is a differential impact for EAL children. However, IPE findings showed that children with EAL tended to have higher attendance and were disproportionately represented in targeted groups, meaning they experienced more consistent access to the elements of the programme most closely linked to the observed impact.

CACE analysis indicated a positive impact for children in compliant settings (ES = 0.13), with the entire confidence interval above zero. This effect is higher than the primary ITT estimate (ES = 0.10), implying that where NELI was implemented as intended, the impact on children's oral language skills was greater. The associated *p*-value (<0.0001) indicates that this positive effect is unlikely to have occurred by chance.

In line with the trial's ToC, we found strong evidence that NELI Preschool increased practitioner confidence, indicating that the programme met its objective of building workforce capability alongside improving children's oral language outcomes. However, the impact evidence suggests that these confidence gains were not the primary mechanism through which children's language improved: mediation results indicate only a small indirect contribution (around 9% of the overall effect), with most of the programme's impact occurring through direct pathways.

In the context of potential future scale-up, it is important to note that the large majority of practitioners reported completing the training in their own time, in some cases unpaid, and that this training was reported to be of a longer duration than other similar training they would undertake. Given that intervention settings were provided with additional funding to pay practitioners for their training time, it is likely that outside of the trial, practitioners who are required to complete the training in their own time will do so unpaid. However, having the option of spreading training completion over a longer period could reduce the number of staff doing the training outside of their usual working hours. This obviously has implications for practitioner workload and wellbeing but also raises the possibility that the training may not be fully completed or completed with the same rigour as in the trial, and that it may also be completed by fewer practitioners per setting, resulting in lower fidelity and quality of delivery overall. These concerns also apply to the amount of time invested in preparing for the sessions, which was seen to be flexible but also important for high-quality delivery. This situation is only likely to be exacerbated as staffing pressures become more acute (Flemons and Worth, 2026). Given the limited perceived usefulness of the Delivery Support Hub and webinar elements, it is possible that making these explicitly optional could go some way to alleviating additional burden.

Other considerations for programme design that emerged from the trial include suitability of all activities, the books and the format for children in this age group, and whether the inclusion of more physical activity could be beneficial for engagement. It may also be worth considering whether reading the book needs to be explicitly integrated into more of the weekly sessions given the limitations of memory in this age group. While many settings took the initiative around reading the book through in later sessions and outside the sessions, making this recommendation more explicit would help make the time commitments around the programme more transparent.

The intensity of programme dosage was flagged as a likely area for adaptation in future scale-up. Although implementation showed generally high fidelity to the core components of the programme, just under three-fifths of settings (59%) reported that they would choose to modify elements of delivery, most commonly in response to the practical challenge of fitting all sessions into existing schedules, suggesting that 'real-world' impact levels may be lower than those observed under trial conditions. Many settings indicated that they would spread the sessions out more to fit them more easily into the setting schedule. However, several also reported that they would expand the targeted component to reach more children. Given that the dosage analysis results suggest that what was delivered in the trial may already exceed saturation, it is possible that expanding the targeted component would reach more children without diluting the effect for the core group. Spreading the sessions out could also support children with lower attendance levels to access more of the content, which could in turn reduce the risk of a widening disadvantage gap.

Finally, it is possible that the requirement for a screen may pose a barrier to scale-up, as the majority of the PVI settings did not already own a screen, and many reported that they would not have been able to purchase one themselves.

The results of this effectiveness trial broadly align with the direction of the existing evidence base showing that structured oral language programmes can improve early vocabulary, listening, and narrative skills; however, the size of the effects

observed here is lower than has been reported in earlier evaluations. Previous evaluations of NELI Preschool (West *et al.*, 2023) and its Reception-age counterpart (Fricke *et al.*, 2017; Dimova *et al.*, 2020) consistently demonstrate positive impacts on children's vocabulary, listening, narrative, and early literacy skills, often equivalent to two to four months of additional progress. These UK-based studies represent the most comparable evidence base, given their similar delivery model, target population, and implementation conditions. In contrast, some of the studies included in the EEF Teaching and Learning Toolkit's estimates for 'Communication and Language Approaches' draw on international evidence and evaluate higher intensity or less comparable interventions. They also contain a high proportion of interventions that were not independently evaluated, which typically results in higher effect sizes; this partly explains why the Toolkit reports an average impact of around +7 months' progress for early years children, which should probably not be interpreted as a direct benchmark for NELI Preschool.

Findings from this effectiveness trial directly complement the results of the earlier efficacy trial of NELI Preschool (West *et al.*, 2023) where the Cohen's *d* was 0.26 for the enrichment (whole-class) component of the intervention. Based on the EEF's effect-size-to-months-progress conversion table, this was equivalent to approximately three months of additional progress (as opposed to two months for this effectiveness trial). However, the conversions are broad indicators rather than precise estimates, and their usefulness depends heavily on the context. In the Early Years Toolkit, 'months' progress' is presented as an accessible summary metric, but learning in the early years does not follow a linear or uniform trajectory. Early development spans wide variations in pace and domain-specific progress, meaning that effect-size conversions should be used cautiously and always interpreted alongside rich contextual information about early years learning. The IPE findings add to the knowledge about what real-world delivery looks like, showing for example, that consistent vocabulary scaffolding is possible to implement, but there remain challenges for the narrative-focused activities. These insights strengthen the evidence base for early language interventions and demonstrate that high-quality language enrichment can be delivered at scale in diverse settings.

The programme appears to have been aligned with existing practice within early years settings while still offering materials and structure that add something new. However, as the use of formalised language development programmes in early years settings is still relatively rare, it is possible that it is simply the use of a structured, consistent approach of this kind that has driven much of the impact, rather than anything specific to the programme itself.

Overall, the findings from this effectiveness trial show a modest but positive impact on children's oral language skills and particularly on expressive language, with strong positive effects in compliant settings when delivered in everyday conditions. The effect sizes are lower in magnitude than those reported in the earlier efficacy trial (West *et al.*, 2023), which concluded that NELI Preschool produced 'substantial improvements' in children's oral language skills in a more tightly controlled context of an efficacy trial. The lower effects observed at scale are in line with expectations for the EEF evidence pipeline: the efficacy trial was conducted in fewer, predominantly maintained, settings with closer delivery team involvement. The current evaluation involved many more settings operating under typical real-world constraints with everyday conditions, including more variable staffing, attendance patterns, and diversity—particularly across PVI and maintained settings. However, emphasis on the importance of following intended practice for the purpose of the trial means that settings may have delivered the programme with greater fidelity within the context of the trial than they would have done in 'real-world' conditions, as indicated by the 59% of NELI Leads who reported that they would modify the programme if they were to deliver it again in their nursery. These contextual differences may help explain the attenuated effect sizes at scale, although it should also be noted that the effect size confidence intervals are overlapping between the efficacy and effectiveness trials meaning we should be wary of overinterpreting these differences in effect size.

Limitations and lessons learned

This effectiveness trial faced several constraints, including 15% missing primary outcome data due to child-level attrition. Variation in implementation fidelity across settings (for example, inconsistent use of digital slides, variable session lengths, lower small-group delivery) may have reduced the observed effect size. There are indications that control settings slightly increased their focus on language development during the trial period, with practitioners demonstrating improvement in confidence around language development knowledge and skills, and the proportion of control settings delivering formalised language development programmes increasing. However, these changes are not significant enough to be a cause for concern. This interpretation is supported by a post-hoc sensitivity analysis that restricted the control group to settings

without concurrent language-development programmes, which produced effect estimates very similar to those from the primary analysis, suggesting that concurrent activity in control settings did not introduce bias into the headline findings.

A significant minority of both control and intervention practitioners reported previous involvement with NELI for Reception, which employs very similar approaches to NELI Preschool. This raised the possibility of contamination through prior practitioner experience, representing a potential threat to internal validity. A second post-hoc sensitivity analysis that restricted the sample to settings with no prior practitioner experience of NELI or NELI Preschool yielded a larger estimated effect on the primary outcome than the main analysis—though with overlapping confidence intervals. This suggests that any bias associated with prior exposure is more likely to have attenuated the observed impact rather than inflated it, indicating that the primary estimate may be conservative for settings with no prior experience of NELI.

Additionally, factors such as attendance-related exclusions from the targeted sessions, noisy environments in some PVI settings, and staffing shortages represent some potential threats to validity by influencing both who received the intended dosage and the quality of the delivery.

In future, addressing attendance-related barriers, especially for EYPP children who were disproportionately likely to miss sessions, should be prioritised in programme delivery.

A further limitation relates to the geographic scope of the trial. Participation was restricted to settings located within designated SPH areas, meaning that we cannot be fully confident that the findings would generalise to settings outside these regions. However, the nine SPHs involved span a very wide range of local authorities across England, covering diverse urban and rural contexts and a broad spectrum of socio-economic characteristics. In practice, this breadth of coverage suggests that the findings are probably applicable to Early Years settings nationally.

Overall, the evaluation demonstrated consistent positive effects across a very large and diverse sample of 303 settings, spanning both PVI and maintained provision, and representing the full range of IDACI scores and percentage of FSM and EAL children. The results were consistent within all tested subgroups including PVI, EAL, and EYPP. All of which suggests that the findings should be capable of generalising to a wide range of settings with characteristics that differ along these dimensions.

Future research and publications

A short follow-up evaluation is planned. A light-touch IPE survey of all NELI Preschool Leads will explore the longer-term influence of the programme on practitioner confidence, motivation, and practice; how settings have continued or adapted elements of the intervention; and any challenges in embedding language-teaching techniques within continuous provision. These data will be analysed descriptively and integrated with a longitudinal impact analysis, in which children's main trial data will be matched to the NPD to obtain EYFSP outcomes, enabling us to re-run the primary analysis using a composite language-related EYFSP measure. An addendum report combining these longitudinal IPE and impact findings is expected to be published in June 2027. In addition, the data underlying this report will be shared with the developers (after publication), and they are intending to use it to publish in an appropriate academic journal.

References

- Bulus, M., Dong, N. and Spybrook, J. (2021) 'PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package'. Available at: <https://cran.r-project.org/src/contrib/Archive/PowerUpR/> (Accessed: 2 July 2026).
- Burgoyne, K., Gardner, R., Whiteley, H., Snowling, M. and Hulme, C. (2018) *Parents And Children Together (PACT): evaluating a parent-delivered oral language enrichment programme for pre-school children*. Available at: https://www.nuffieldfoundation.org/wp-content/uploads/2019/11/Hulme204179320-20Main20public20report_PACT.pdf (Accessed: 5 February 2026).
- Chen, J.-Q. and McCray, J. (2013) *A survey study of early childhood teachers' beliefs and confidence about teaching early math*. Available at: <https://earlymath.erikson.edu/wp-content/uploads/2013/12/2013-1-Chen-McCray-Survey-Study-of-Early-Childhood-Teachers-Beliefs-and-Confidence-about-Teaching-Early-Math.pdf> (Accessed: 5 February 2026).
- Chow, J.C. and Ekholm, E. (2019) 'Language domains differentially predict mathematics performance in young children', *Early Childhood Research Quarterly*, 46(Quarterly), pp. 179–186. Available at: <https://doi.org/10.1016/j.ecresq.2018.02.011>.
- Department for Education (2026) *Early years foundation stage profile results - Academic year 2024/25*, GOV.UK. Available at: <https://explore-education-statistics.service.gov.uk/find-statistics/early-years-foundation-stage-profile-results/2024-25> (Accessed: 8 April 2026).
- DfE (2025a) *Childcare and early years provider survey: reporting year 2025*, GOV.UK. Available at: <https://explore-education-statistics.service.gov.uk/find-statistics/childcare-and-early-years-provider-survey/2025> (Accessed: 4 February 2026).
- DfE (2025b) *Early years foundation stage statutory framework: for group and schools-based providers. Setting the standards for learning, development and care for children from birth to five*. Available at: https://assets.publishing.service.gov.uk/media/68c024cb8c6d992f23edd79c/Early_years_foundation_stage_statutory_framework_-_for_group_and_school-based_providers.pdf (Accessed: 22 January 2026).
- DfE (2025c) *Funded early education and childcare. Reporting year 2025*, GOV.UK. Available at: <https://explore-education-statistics.service.gov.uk/find-statistics/funded-early-education-and-childcare/2025> (Accessed: 28 January 2026).
- DfE (2025d) *Giving every child the best start in life*. CP 1362. Available at: https://assets.publishing.service.gov.uk/media/689b39d55555fb89cf3f5e7d/Giving_every_child_the_best_start_in_life_web_version.pdf (Accessed: 22 January 2026).
- DfE (2025e) *Help for early years providers: communication and language*. Available at: <https://help-for-early-years-providers.education.gov.uk/areas-of-learning/communication-and-language> (Accessed: 22 January 2026).
- DfE (2025f) *Pulse surveys of childcare and early years providers, 2024-2025*. Available at: https://assets.publishing.service.gov.uk/media/68ff9bdebcbb10f6bf9bef938/Pulse_surveys_of_childcare_and_early_years_providers__2024_to_2025.pdf (Accessed: 5 February 2026).
- DfE (2025g) *Pupil absence in schools in England: autumn and spring term 2024/25*, GOV.UK. Available at: <https://explore-education-statistics.service.gov.uk/find-statistics/pupil-absence-in-schools-in-england/2024-25-autumn-and-spring-term> (Accessed: 15 January 2026).
- Dimova, S., Ilie, S., Brown, E.R., Broeks, M., Culora, A. and Sutherland, A. (2020) *Nuffield Early Language Intervention: evaluation report*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Nuffield_Early_Language_Intervention.pdf?v=1770289018 (Accessed: 5 February 2026).
- EEF (2019) *Lessons learnt from EEF early years trials: recommendations for evaluators*. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-syntheses/EY_lessons_learnt.pdf?v=1630582495 (Accessed: 5 February 2026).

EEF (2022) *Statistical analysis guidance for EEF evaluations*. Available at: <https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1709857308> (Accessed: 5 February 2026).

EEF (2023a) *Communication and language approaches: high impact for low cost based on moderate evidence*, EEF. Available at: <https://educationendowmentfoundation.org.uk/early-years/toolkit/communication-and-language-approaches> (Accessed: 5 February 2026).

EEF (2023b) 'Cost evaluation guidance for EEF evaluations'. EEF. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/Cost-Evaluation-Guidance-Feb_2023.pdf?v=1765786924 (Accessed: 15 January 2026).

Fernet, C., Senécal, C., Guay, F., Marsh, H. and Dowson, M. (2008) 'The Work Tasks Motivation Scale for Teachers (WTMST)', *Sage Journals*, 16(2), pp. 256–279.

Field, F. (2010) *The Foundation Years: preventing poor children becoming poor adults*. Available at: http://www.complexneeds.org.uk/modules/Module-1.1-Understanding-the-child-development-and-difficulties/All/downloads/m01p040d/the_foundation_years_preventing_poor_children_becoming_poor_adults.pdf (Accessed: 5 February 2026).

Flemons, L. and Worth, J. (2025) *The early years workforce in England 2025*. Available at: https://www.nfer.ac.uk/media/4uwi4cs5/the_early_years_workforce_in_england_2025.pdf (Accessed: 15 January 2026).

Flemons, L. and Worth, J. (2026) *The Early Years Workforce in England Annual Report 2026*. NFER. Available at: <https://www.nfer.ac.uk/publications/the-early-years-workforce-in-england-annual-report-2026/> (Accessed: 30 June 2026).

Fletcher-Wood, H. and Zuccollo, J. (2020) *The effects of high-quality professional development on teachers and students*. Available at: https://epi.org.uk/wp-content/uploads/2020/02/EPI-Wellcome_CPD-Review__2020.pdf (Accessed: 5 February 2026).

Fricke, S., Bowyer-Crane, C., Haley, A.J., Hulme, C. and Snowling, M.J. (2012) 'Efficacy of language intervention in the early years', *Journal of Child Psychology and Psychiatry*, 54(3), pp. 280–290. Available at: <https://doi.org/10.1111/jcpp.12010>.

Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., Lervåg, A., Snowling, M.J. and Hulme, C. (2017) 'The efficacy of early language intervention in mainstream school settings: a randomized controlled trial', *Journal of Child Psychology and Psychiatry*, 58(10), pp. 1141–1151. Available at: <https://doi.org/10.1111/jcpp.12737>.

GL Assessment (2017) *What is in the Toolkit?*, *GL Education Support*. Available at: <https://support.gl-education.com/knowledge-base/assessments/wellcomm-support/general-information/what-is-in-the-toolkit> (Accessed: 15 January 2026).

Gonzales, C.R., Bowles, R., Geldhof, G.J., Cameron, C.E., Tracy, A. and McClelland, M.M. (2021) 'The Head-Toes-Knees-Shoulders Revised (HTKS-R): development and psychometric properties of a revision to reduce floor effects', *Early Childhood Research Quarterly*, 56, pp. 320–332. Available at: <https://doi.org/10.1016/j.ecresq.2021.03.008>.

Goodman, A. and Gregg, P. (2010) *Poorer children's educational attainment: how important are attitudes and behaviour?*, *Joseph Rowntree Foundation*. Available at: https://www.educationandemployers.org/wp-content/uploads/2014/06/poorer_children_s_educational_attainment_-_jrf_full_report1.pdf (Accessed: 5 February 2026).

Guo, G. and Harris, K.M. (2000) 'The mechanisms mediating the effects of poverty on children's intellectual development', *Demography*, 37(4), pp. 431–447. Available at: <https://doi.org/10.1353/dem.2000.0005>.

Hart, B. and Risley, T.R. (1995) *Meaningful differences in the everyday experience of young american children*. Baltimore: Brookes Publishing. Available at: <https://products.brookespublishing.com/Meaningful-Differences-in-the-Everyday-Experience-of-Young-American-Children-P14.aspx> (Accessed: 5 February 2026).

Haux, T., Butt, S., Rezaian, M., Garwood, E., Woodbridge, H., Bhatti, S., Rogan, R.W., Davies, H., Bain, E. and Hunnikin, L. (2022) *Early years recruitment, retention and business planning during the Coronavirus pandemic research report: a summary of findings*. Available at:

- https://assets.publishing.service.gov.uk/media/62695faae90e0746cd3f0f9c/Early_years_recruitment__retention_and_business_planning_during_the_Coronavirus_pandemic.pdf (Accessed: 5 February 2026).
- Hjetland, H.N., Brinchmann, E.I., Scherer, R., Hulme, C. and Melby-Lervåg, M. (2020) 'Preschool pathways to reading comprehension: A systematic meta-analytic review', *Educational Research Review*, 30(3), p. 100323. Available at: <https://doi.org/10.1016/j.edurev.2020.100323>.
- Hoff, E. (2003) 'The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech', *Child Development*, 74(5), pp. 1368–1378. Available at: <https://doi.org/10.1111/1467-8624.00612>.
- Hornburg, C.B., Schmitt, S.A. and Purpura, D.J. (2018) 'Relations between preschoolers' mathematical language understanding and specific numeracy skills', *Journal of Experimental Child Psychology*, 176, pp. 84–100. Available at: <https://doi.org/10.1016/j.jecp.2018.07.005>.
- Howard, S.J., Siraj, I., Melhuish, E.C., Kingston, D., Neilsen-Hewett, C., de Rosnay, M., Duursma, E. and Luu, B. (2020) 'Measuring interactional quality in pre-school settings: introduction and validation of the Sustained Shared Thinking and Emotional Wellbeing (SSTEW) scale', *Early Child Development and Care*, 190(7), pp. 1017–1030. Available at: <https://doi.org/10.1080/03004430.2018.1511549>.
- Hughes, C., Daly, I., Foley, S., White, N. and Devine, R.T. (2015) 'Measuring the foundations of school readiness: Introducing a new questionnaire for teachers. The Brief Early Skills and Support Index (BESSI)', *British Journal of Educational Psychology*, 85(3), pp. 332–356. Available at: <https://doi.org/10.1111/bjep.12076>.
- Hulme, C., McGrane, J., Duta, M., West, G., Cripps, D., Dasgupta, A., Hearne, S., Gardner, R. and Snowling, M. (2024) *LanguageScreen: The development, validation, and standardization of an automated language assessment app*. American Speech-Language-Hearing Association. Available at: https://pubs.asha.org/doi/10.1044/2024_LSHSS-24-00004 (Accessed: 14 June 2024).
- Hulme, C., Nash, H.M., Gooch, D., Lervåg, A. and Snowling, M.J. (2015) 'The foundations of literacy development in children at familial risk of dyslexia', *Psychological Science*, 26(12), pp. 1877–1886. Available at: <https://doi.org/10.1177/0956797615603702>.
- Hutchinson, J., Reader, M. and Akhal, A. (2020) *Education in England: annual report 2020*. Available at: https://epi.org.uk/wp-content/uploads/2020/09/EPI_2020_Annual_Report_.pdf (Accessed: 5 February 2026).
- Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M. and Lyons, T. (1991) 'Early vocabulary growth: relation to language input and gender', *Developmental Psychology*, 27(2), pp. 236–248. Available at: <https://doi.org/10.1037/0012-1649.27.2.236>.
- 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 socioeconomically disadvantaged backgrounds*. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/guidance/Law_et_al_Early_Language_Development_final.pdf?v=168434613 (Accessed: 5 February 2026).
- Liberman, I.Y., Shankweiler, D. and Liberman, A.M. (1989) 'The alphabetic principle and learning to read', *Phonology and reading disability: Solving the reading puzzle*. Ann Arbor, MI, US: The University of Michigan Press (International Academy for Research in Learning Disabilities monograph series, No. 6.), pp. 1–33. Available at: <https://files.eric.ed.gov/fulltext/ED427291.pdf> (Accessed: 5 February 2026).
- Marulis, L.M. and Neuman, S.B. (2010) 'The effects of vocabulary intervention on young children's word learning: A meta-analysis', *Review of Educational Research*, 80(3), pp. 300–335. Available at: <https://doi.org/10.3102/0034654310377087>.
- McMullin, C. (2021) 'Transcription and Qualitative methods: implications for third sector research', *VOLUNTAS*, 34(1), pp. 140–153. Available at: <https://doi.org/10.1007/s11266-021-00400-3>.
- Murray, A. and Egan, S.M. (2014) 'Does reading to infants benefit their cognitive development at 9-months-old? An investigation using a large birth cohort survey', *Child Language and Therapy*, 30(3), pp. 303–315. Available at: <https://doi.org/10.1177/0265659013513813>.
- Naeem, M., Ozuem, W., Howell, K. and Ranfagni, S. (2023) 'A step-by-step process of thematic analysis to develop a conceptual model in qualitative research', *International Journal of Qualitative Methods*, 22, p. 16094069231205789. Available at: <https://doi.org/10.1177/16094069231205789>.

- Nolan, A. and Molla, T. (2017) 'Teacher confidence and professional capital', *Teaching and Teacher Education*, 62, pp. 10–18. Available at: <https://doi.org/10.1016/j.tate.2016.11.004>.
- Public Health England (2020a) 'Best start in speech, language and communication: guidance to support local commissioners and service leads'. Available at: https://assets.publishing.service.gov.uk/media/5f9be9bbe90e0704157fb12f/BSSLC_Guidance.pdf (Accessed: 22 January 2026).
- Public Health England (2020b) *Best start in speech, language and communication (SLC)*, GOV.UK. Available at: <https://www.gov.uk/government/publications/best-start-in-speech-language-and-communication> (Accessed: 22 January 2026).
- Public Health England (2020c) 'Early language identification measure and intervention: guidance handbook'. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/939872/ELIM_Handbook_December-2020.pdf (Accessed: 22 January 2026).
- Renfrew, C. (2019) *Action Picture Test*. Oxford: Routledge.
- Roulstone, S., Law, J., Rush, R., Clegg, J. and Peters, J. (2011) *Investigating the role of language in children's early educational outcomes*. DFE-RR134. Available at: <https://assets.publishing.service.gov.uk/media/5a7b67a5e5274a319e77f135/DFE-RR134.pdf> (Accessed: 5 February 2026).
- Roy, P., Flemons, L., Schwendel, G., Welbourne, S., Speciani, E.R. and Groom, M. (2024) *A randomised controlled trial of the effectiveness of the Nuffield Early Language intervention (NELI): preschool programme: evaluation protocol*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/neli_preschool_-_evaluation_protocol_2024-09-05-143529_qzpo.pdf?v=1750348042 (Accessed: 21 April 2026).
- Roy, P., Flemons, L., Schwendel, G., Welbourne, S., Speciani, E.R. and Groom, M. (2025a) *A randomised controlled trial of the effectiveness of the Nuffield Early Language Intervention (NELI) preschool programme statistical analysis plan*. EEF. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/neli_preschool_-_statistical_analysis_plan_-_v.1.0.0.pdf?v=1768757383 (Accessed: 4 February 2026).
- Roy, P., Flemons, L., Schwendel, G., Welbourne, S., Speciani, E.R. and Groom, M. (2025b) *A randomised controlled trial of the effectiveness of the Nuffield Early Language Intervention (NELI) Preschool programme*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/neli_preschool_-_evaluation_protocol_-_v.2.0.0.pdf?v=1770209090 (Accessed: 4 February 2026).
- Sampson, R.J., Sharkey, P. and Raudenbush, S.W. (2008) 'Durable effects of concentrated disadvantage on verbal ability among African-American children', *Proceedings of the National Academy of Sciences*, 105(3), pp. 845–852. Available at: <https://doi.org/10.1073/pnas.0710189104>.
- 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*. London: SAGE Publications Ltd., pp. 23–39. Available at: https://www.researchgate.net/publication/306200560_Connecting_early_language_and_literacy_to_later_reading_disabilities_Evidence_theory_and_practice (Accessed: 5 February 2026).
- Semel, E., Wiig, E. and Secord, W. (2006) *Child Evaluation of language fundamentals preschool UK*. 2nd edn. Oxford: Pearson Assessment.
- Sibieta, L., Kotecha, M. and Skipp, A. (2016) *Nuffield early language intervention*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/EEF_Project_Report_Nuffield_Early_Language_Intervention.pdf?v=1710077612 (Accessed: 5 February 2026).
- Sirin, S.R. (2005) 'Socioeconomic status and academic achievement: a meta-analytic review of research', *Review of Educational Research*, 75(3), pp. 417–453. Available at: <https://doi.org/10.3102/00346543075003417>.
- Smith, A., Staunton, R., Sahasranaman, A. and Worth, J. (2023) *Impact evaluation of Nuffield Early Learning Intervention (NELI) wave two*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/NELI-wave-2-scale-up-impact-evaluation-report-FINAL_2023-09-13-094553_bzhhh.pdf?v=1717592765 (Accessed: 5 February 2026).

Snowling, M., Hulme, C., Butterworth, B., Morton, J., Snowling, M.J. and Warrington, E.K. (1994) 'The development of phonological skills', *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 346(1315), pp. 21–27. Available at: <https://doi.org/10.1098/rstb.1994.0124>.

Speech and Language UK (2023) *Early Talk Boost*. Available at: <https://speechandlanguage.org.uk/educators-and-professionals/programmes-for-nurseries-and-schools/talk-boost/early-talk-boost/> (Accessed: 15 January 2026).

Weisleder, A. and Fernald, A. (2013) 'Talking to children matters: early language experience strengthens processing and builds vocabulary', *Psychological Science*, 24(11), pp. 2143–2152. Available at: <https://doi.org/10.1177/0956797613488145>.

Welbourne, S.R., Woollams, A.M., Crisp, J. and Lambon Ralph, M.A. (2011) 'The role of plasticity-related functional reorganization in the explanation of central dyslexias', *Cognitive Neuropsychology*, 28(2), pp. 65–108. Available at: <https://doi.org/10.1080/02643294.2011.621937>.

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

West, G., Snowling, M.J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H. and Hulme, C. (2021) 'Early language screening and intervention can be delivered successfully at scale: evidence from a cluster randomized controlled trial', *Journal of Child Psychology and Psychiatry*, 62(12), pp. 1425–1434. Available at: <https://doi.org/10.1111/jcpp.13415>.

Woollams, A.M., Halai, A. and Lambon Ralph, M.A. (2018) 'Mapping the intersection of language and reading: the neural bases of the primary systems hypothesis', *Brain Structure & Function*, 223(8), pp. 3769–3786. Available at: <https://doi.org/10.1007/s00429-018-1716-z>.

Appendix A: EEF cost rating

Cost rating	Description
£ £ £ £ £	<i>Very low:</i> less than £80 per pupil per year.
£ £ £ £ £	<i>Low:</i> up to about £200 per pupil per year.
£ £ £ £ £	<i>Moderate:</i> up to about £700 per pupil per year.
£ £ £ £ £	<i>High:</i> up to £1,200 per pupil per year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

OUTCOME: Oral language skills

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 (for example, RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%	4		Adjustment for threats to internal validity [0]		4
3	Design for comparison that considers selection on all relevant observable confounders (for example, Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%					
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%					
0	No comparator	≥ 0.6	$>50\%$					

Threats to validity	Threat to internal validity?	Comments
Threat 1: Confounding	Low	Randomisation carried out according to SAP, not blinded but provided with minimum data required. No evidence of imbalance between groups at baseline.
Threat 2: Concurrent Interventions	Low	Evidence of control group taking up concurrent interventions during the trial (increase from 33% at baseline to 41% at endline), though most common intervention at baseline (Wellcomm) quite different from NELI Preschool and no evidence that screening or early language development practices changed. Sensitivity analysis findings suggest that this did not introduce detectable bias into the primary outcome results, as the effect sizes estimates are very similar (0.11 vs 0.10).

Threat 3: Experimental effects	Moderate	Evidence of contamination through more control group practitioners having prior experience with NELI/NELI Preschool, though difference is relatively small (22% control vs 16% intervention) and findings suggest practitioner outcomes do not strongly mediate child outcomes. Sensitivity analysis findings cannot rule out bias having been introduced, as the effect size estimate is larger than that of the primary analysis (0.17 vs 0.10). This may be leading to an underestimation of the treatment effect, but the confidence intervals of the two estimates overlap, prior experience with NELI was considered BAU for this effectiveness trial and exposure to NELI Preschool seems to be limited to LanguageScreen training.
Threat 4: Implementation fidelity	Low	Implementation fidelity and compliance are well defined and aligned with the most important elements of the ToC, and both are high (for example, 81% overall compliance).
Threat 5: Missing Data	Low	15% of children missing primary outcome data, with proportion being slightly higher for intervention than control pupils (19% vs 11%). However, results of sensitivity analysis accounting for observable predictors of missingness (EAL status, two SPHs) are similar to complete case analysis results.
Threat 6: Measurement of Outcomes	Low	Valid, reliable and appropriate/acceptable outcome measures independently administered and marked by evaluators. Measures combined into a latent oral language variable with no evidence of ceiling/floor effects.
Threat 7: Selective reporting	Low	Analysis is comprehensively reported and deviations from the study plan are well explained. Interpretation of findings is appropriately cautious.

- **Initial padlock score:** 4 Padlocks – Randomised design powered to detect a MDES of 0.16. However, suffered 15% attrition, leading to a loss of a padlock (i.e., padlock of four is the highest that can be achieved with attrition of 11-20%).
- **Reason for adjustment for threats to validity:** 0 Padlocks – While there was evidence of contamination that may have led to an underestimation of the programme's effect, uncertainty around the sensitivity analysis estimate and nature of prior experience with NELI/exposure to NELI Preschool make this a moderate threat. Given all other threats are low risk, this does not affect the overall security rating of the trial.
- **Final padlock score:** initial score adjusted for threats to validity = 4 Padlocks.
- **Reason(s) for final padlock rating:** 15% attrition leads to an initial loss of a padlock, with no further padlocks being lost for threats to internal validity.

Appendix C: Changes since the efficacy trial

Feature		Efficacy to effectiveness stage
Intervention	Intervention content	In the current evaluation, the delivery team provided comprehensive guidance on selecting children for the targeted component. In addition to this, settings were required to complete the Narrative Progress Tracker template which was not part of the efficacy trial.
	Delivery model	There were no changes to the delivery mechanism, i.e., the developers led intervention delivery in this trial.
	Intervention duration	There were no changes in the duration of delivery between the efficacy and effectiveness trial.
Evaluation	Eligibility criteria	The efficacy trial predominantly involved maintained nursery settings whereas 42% of randomised settings in this trial were PVI settings. Settings in the previous trial were from seven geographical areas and the settings in the current evaluation were from the LAs in the SPH areas. The current evaluation also excluded settings that took part in the efficacy trial and those who took in other SPH-funded programmes or other EEF-funded evaluations. The target age of the children was the same as the previous evaluation although the delivery team provided additional Targeted Group Selection Guidance which was not part of the previous evaluation.
	Level of randomisation	This trial was a setting-randomised trial, similar to the efficacy trial.
	Outcomes and baseline	Outcome measures were different across both trials. The primary outcome measure for the efficacy trial was a latent variable from eight subtests across LanguageScreen, CELF and RAPT, whereas the current evaluation used six subtests from LanguageScreen and RAPT to derive the primary outcome. The baseline for the previous evaluation was a latent variable from eight subtests across LanguageScreen, CELF and RAPT whereas the current evaluation used LanguageScreen as a baseline measure. In the previous evaluation, LanguageScreen was administered as a screening test to identify children with the weakest language skills prior to randomisation, while CELF and RAPT were administered after randomisation. The current evaluation also utilised LanguageScreen as a screening test to identify children with the weakest language skills and as a baseline measure and was also administered prior to randomisation.
	Control condition	In both evaluations, the control group continued with their usual practice. The efficacy trial was a wait list design whereas control group settings in the current evaluation received £1,000 which was offered as a financial payment or to be used to purchase the NELI Preschool programme at the end of the trial.

Appendix D: Effect size estimation

Appendix Table 44: Effect size estimation

			Variance components obtained from a model with no predictors			
Outcome	Unadjusted differences in means	Adjusted differences in means	N intervention (missing); N control (missing)	Between-schools variance	Within-class variance	Pooled variance
Primary Outcome	0.24	0.20	1,692 (282); 1,606 (289)	0.85	2.95	1.97
EYPP subgroup	0.30	0.21	449 (104); 452 (86)	0.38	3.06	1.86

Further appendices:

Appendix E to Appendix R are provided in a separate document '*Further appendices*' that can be found on to project page.

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

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

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

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



Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

<https://educationendowmentfoundation.org.uk>

 @EducEndowFoundn

 [Facebook.com/EducEndowFoundn](https://www.facebook.com/EducEndowFoundn)