



Using the Orchestrating Numeracy and the Executive (THE ONE) Programme to improve maths attainment, a two-armed cluster randomised trial

Evaluation Report

July 2025

Anna Brian, James Merewood, Elena Rosa Speciani, Fin Oades,
Emily MacLeod, Merrilyn Groom

Co-funded by:

London South Early Years Stronger Practice Hub

REACHout – Early Years Stronger Practice Hub, East of England

Derbyshire and Nottinghamshire Early Years Stronger Practice Hub

Yorkshire and Humber Together Early Years Stronger Practice Hub



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

We do this by:

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

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

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



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



info@eefoundation.org.uk



www.educationendowmentfoundation.org.uk



Department
for Education

What
Works
Network 

Contents

About the evaluator	3
Executive summary	4
Introduction	6
Methods	15
Impact evaluation results	37
Implementation and process evaluation results	51
Conclusion	58
References	61
Appendix A: EEF cost rating	64
Appendix B: Security classification of trial findings.....	65
Appendix C: Effect size estimation and additional tables.....	67
Appendix D: Outcome Distributions.....	72
Appendix E: Residual plots from analysis models	75
Further appendices:	88

About the evaluator

The project was independently evaluated by a team from RAND Europe; Elena Rosa Speciani, Anna Brian, James Merewood, Fin Oades, Emily MacLeod, Marilyn Groom

Contact details

Elena Rosa Speciani
RAND Europe
Eastbrook House
Shaftesbury Road
Cambridge, CB2 8DR, United Kingdom
+44 1223 353 329

erspeciani@randeurope.org

Acknowledgements

We would like to thank a number of key people without whom the project would not have been successful: Marilyn Groom, Emily MacLeod, Sarah Angell, and Rachel Hesketh (formerly RAND Europe) who supported the project at various points and the delivery team—Gaia Scerif, Emma Blakey, Caroline Korell, Rosemary O'Connor, Toni Loveridge, Carmel Brough, Sophie Smith, Joanna Archibald, Alicia Mortimer, Siobhan Murray, Alexandra Turner, Hannah Palmer, Molly Staley, Holly Amos, Victoria Simms, Zachary Hawes, Steven Howard, Rebecca Merkley, and Fionnuala O'Reilly—who were collaborative and were very supportive of our research, as were our testing partners at Qa Research, in particular Katie Morris and Rosie Walker-Lyne. The project received the support of the Stronger Practice Hubs, for which we are very grateful. Finally, many thanks to the EEF team—Lauren Spinner, Toby Whittaker, and Ben Simuyandi (formerly of the EEF)—for all your hard work.

Executive summary

The project

The Orchestrating Numeracy and the Executive Programme (the ONE) aims to support early years practitioners to deliver engaging, short, play-based activities that develop children's early executive functioning and numeracy skills. It includes 25 co-developed activities supported by activity cards that provide guidance for integrating executive function elements into each task. Each activity lasts between five to ten minutes and can be easily incorporated into routines, including small group activities and free play. During the 12-week implementation, practitioners must conduct at least three activities each week, with flexibility in group size and format. All children aged three to four who are preparing to transition into reception are encouraged to participate in these activities to enhance their engagement and learning.

Training sessions for practitioners are provided by the University of Oxford and the University of Sheffield and comprise four weekly 30-minute meetings centred on early numeracy and executive function integration. These sessions are set in terms of content and learning goals for educators, but they allow for time for reflection, questions, and are scheduled at times that are flexible through the day to meet the specific needs of each setting. Follow-up support is scheduled in the eighth and twelfth weeks to ensure adherence to the programme and provide further guidance for practitioners.

This efficacy trial was conducted as a cluster randomised controlled trial with randomisation at the setting level across 150 settings and 1,859 children with half randomly assigned to receive the programme starting in either January or February 2024 and the other half assigned to a waitlist control group who will receive the programme in the 2024/2025 academic year. All participating settings were visited by assessors who tested children on their executive functioning and numeracy skills to look at the impact of the ONE on child outcomes. A mixed methods implementation and process evaluation, which included training observations, semi-structured interviews, and surveys, explored how the intervention was delivered in practice as well as understanding usual practice.

As part of the Department for Education's Early Years Recovery Programme, the Education Endowment Foundation (EEF) is working with Stronger Practice Hubs across England to fund Early Years settings' access to evidence-informed programmes and study the programme's influence on practice and children's outcomes. This initiative aims to support education recovery following the pandemic, whilst also developing our understanding of effective professional development in the early years. The EEF has worked with the London South, Derbyshire and Nottinghamshire, REACHOut East of England, and Yorkshire and Humber Together Early Years Stronger Practice Hubs to fund settings' access to the ONE Programme and evaluate the programme through an efficacy trial.

Table 1: Key conclusions

Key conclusions	
1.	Children in the ONE settings made no additional progress in maths, on average, compared to children in control settings. This result has a high security rating.
2.	Children in the ONE settings made no additional progress in executive functioning, on average, compared to children in control settings.
3.	Among children receiving Early Years Pupil Premium (EYPP), those in the ONE settings made two additional months' progress in maths, on average, compared to children in control settings. These results may have a lower security than the overall findings because of the smaller number of children.
4.	There is evidence to suggest that the training and support offered by the ONE team were well received and led to changes in practitioners' understanding of the importance of executive functioning to mathematical attainment.

EEF security rating

These findings have a high security rating. This was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a large number of settings. The trial was a well-designed two-armed randomised controlled trial and was well-powered. Child and setting characteristics at randomisation and endline were well-balanced across the two trial arms in terms of setting type, region, gender, and EYPP eligibility. Relatively few children (9%) who started

the trial were not included the final analysis. Implementation fidelity risks and risks over the control group implementing similar activities make it harder to accurately estimate the size of the impact on the pupils in the trial.

Additional findings

Children in settings that delivered the ONE made no additional progress in maths attainment compared to those 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 for children in settings that delivered the ONE also includes small negative effects of two months less progress and positive effects of up to two months additional progress compared to children in control settings. The evaluation found similar outcomes in executive functioning attainment for children in intervention settings: no additional progress was found compared to control children. This possible lack of impact includes negative effects of two months less progress and positive effects of up to two months additional progress.

Children receiving EYPP in the ONE settings made two additional months' progress, on average, compared to EYPP children in control settings. This possible impact includes negative effects of one month less progress and positive effects of up to five months additional progress. These results may have a lower security than the overall findings because of the smaller number of children.

Measurement and administration errors at baseline, and to a lesser extent endline, may have affected the results, particularly with regard to the secondary outcomes of executive functioning. Administration and measurement error at baseline increased the risk of some non-random sampling at the child level in some settings (which may affect up to 17% of settings), increased attrition in the secondary outcomes at baseline, and exacerbated floor effects at baseline on secondary outcomes. These issues were largely not repeated at endline, with the exception of the persistence of substantial floor effects in HTKS-R (Heads-Toes-Knees-Shoulders, see page 11). An issue with the coding of the primary outcome meant that 16% of the analytical sample violated the normal stopping rules which limits confidence in the primary outcome findings. While additional robustness checks have been carried out, these checks cannot alleviate all concerns, and the secondary outcome analysis should be interpreted with caution in light of these issues.


There is evidence that the ONE intervention was well received by practitioners and lead to real change in practitioners' understanding. Insights gathered from practitioners highlighted their positive reception of the training and materials provided by the ONE team. There is some evidence that the ONE increased practitioners' understanding of the importance of executive function in early maths development. Overall, the ONE demonstrated potential benefits, but its effectiveness may have been impacted by intervention duration or other implementation factors that will require attention in future iterations of the project.

Cost

The average cost of the ONE for one setting for the first year was £3,389.80. When averaged over three years, the average cost per setting is £2,246.96 per year, or £6,740.88 over three years, or £69.14 per child per year.

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 children	P Value	EEF cost rating
Early Years Toolbox Numeracy (EYTN) (overall sample)	0.01 (-0.12; 0.13)	0		1689	0.92	£ £ £ £ £
Early Years Toolbox Numeracy (EYTN) (EYPP subgroup)	0.14 (-0.09; 0.37)	2	N/A	267	0.22	N/A

Introduction

Background

Studies have shown that early mathematics achievement is highly predictive of later mathematics performance (Verdine et al., 2014). We also know that children who fall behind their peers in mathematics early usually continue to develop their maths skills at a slower rate than their more advanced peers and are likely to remain behind them (Purpurpa and Lonigan, 2015).

There is a growing body of evidence that highlights the critical connection between early maths learning and executive functions (for example, Coolen et al, 2021). 'Executive functioning' (EF) refers to a set of cognitive processes that are responsible for planning, organizing, initiating, and regulating goal-directed behaviour. These processes include working memory, cognitive flexibility, inhibitory control, and attentional control (Coolen et al., 2021). Executive skills have been shown to predict domain-specific maths skills for four-year-olds prior to school entry, although these rarely feature in early years practitioner training. Studies with disadvantaged children have also shown that certain elements of executive functioning are highly correlated with early mathematics ability, suggesting that EF may be a key means of narrowing the attainment gap (Blair and Razza, 2007). One reason for this could be that socioeconomically disadvantaged children may have fewer opportunities to practice executive functions (Blair and Raver, 2014). This suggests a vicious cycle of poor exposure and practice for these two inter-related skills.

A programme integrating executive challenge into play-based activities (without the maths focus) has been trialled in Australia (Howard et al, 2020). It resulted in improvements in executive functions for the intervention settings, though improvements in attainment did not reach statistical significance. This current project—Orchestrating Numeracy and the Executive (the ONE)—builds on this work by adapting the Australian programme to the United Kingdom early years context and by incorporating activities with well-evidenced maths-specific content (for example, Moss et al., 2016), given mounting evidence that executive functions are key for early mathematical development. This content has already been co-developed with early years teachers and practitioners in pilot settings and underwent a feasibility RCT in 16 settings in the 2021/2022 academic year (Scerif et al., 2023).

This evaluation is a two-armed, randomised waitlisted controlled trial, with randomisation at the setting level. A waitlisted design allows for delivery across all settings recruited as part of the Stronger Practice Hubs, with those in the treatment condition receiving the intervention in 2023/2024 and those on the waitlist (that is, the control group) receiving the programme in the following academic year. Setting level randomisation is best suited to the whole-class delivery model of the ONE.

The impact evaluation measured maths attainment as a primary outcome and executive functioning as secondary outcomes. Maths attainment and executive attainment were measured at baseline and endline. Our approach to the implementation and process evaluation combined a number of data collection methods allowing for triangulation, comparison between arms of the trial, and captured the experience of all key stakeholders in an efficient and timely manner.

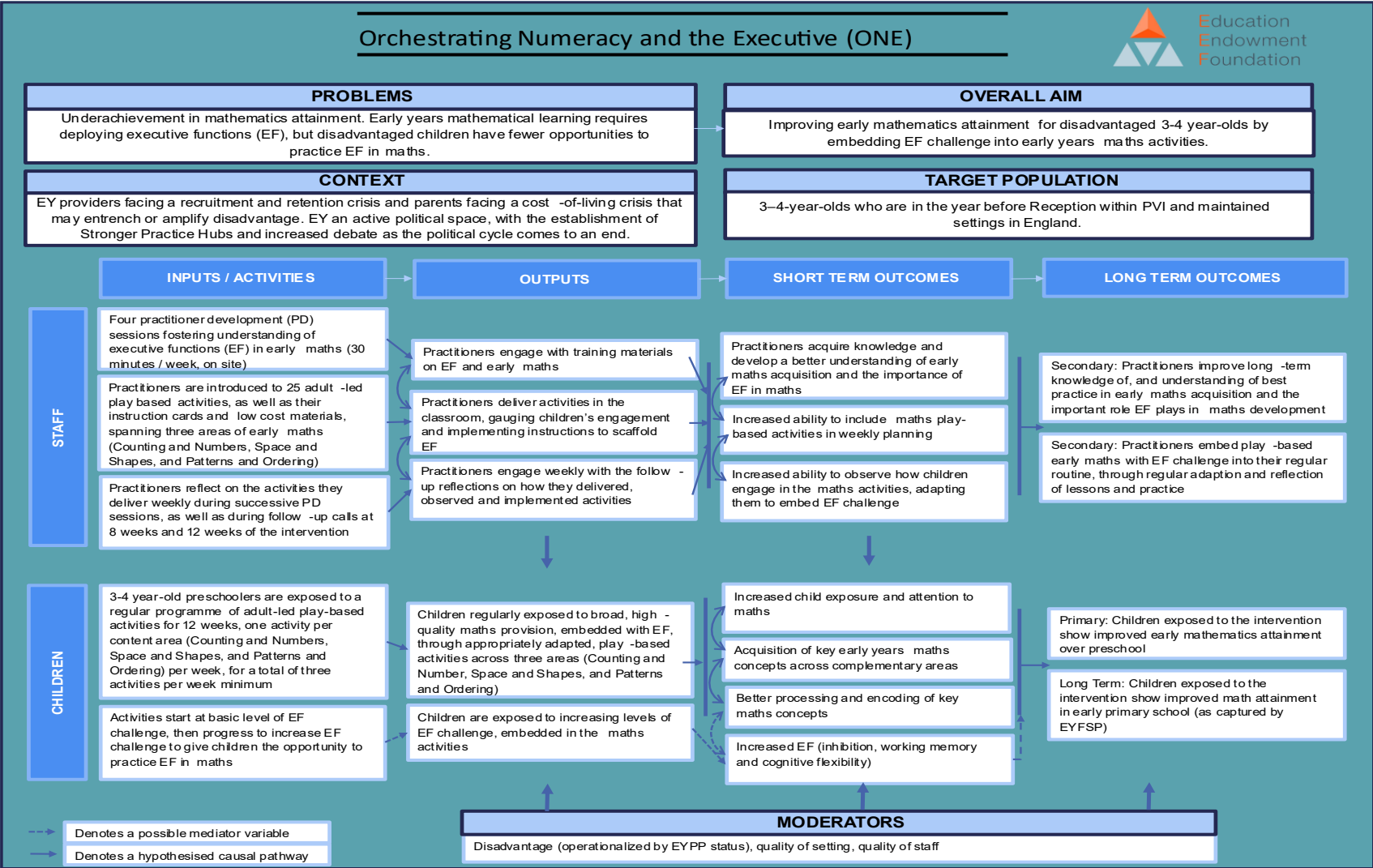
The project is part of a wider programme of work focusing on interventions in early years settings, co-funded with the Department for Education's (DfE) Stronger Practice Hubs (SPH). SPHs were set up to provide advice, share good practice, and offer evidence-based professional development for early years practitioners as part of the DfE's early years education recovery support package. The projects are a major part of the EEF's increased focus on generating evidence for the early years sector.

Intervention

The ONE programme: 'Orchestrating Numeracy and the Executive' was developed by a team from the University of Oxford and the University of Sheffield. The ONE is a professional development-based programme that involves training and support for early years practitioners to run play-based maths activities that support maths development by embedding executive functioning skills into maths learning. Early years practitioners deliver this play-based intervention to children who are due to start school in the following academic year (three- to four-year-olds). This can be seen in the

Logic Model (Figure 1). The ONE consists of face to face training for educators, a pack of 25 activity cards, and resources to be used with the activities. The ONE is a whole-class intervention so all children in the classroom or playgroup, inclusive of those due to start school in the following year, have access to the intervention.

Figure 1: The ONE Logic Model



Who is trained

Settings were asked to nominate at least one practitioner per setting.

The following inclusion criteria were applied: at least one member of staff who was directly involved in the day-to-day education of three- to four-year-olds at each setting was required to attend each session. However, settings were allowed to nominate more than one practitioner if the setting had the capacity and interest for more practitioners to be trained. In this trial, between one and twelve staff members per setting were trained. *Only practitioners that were trained were asked to deliver the intervention.*

Training and ongoing support

The ONE training consisted of four weekly, 30-minute, face to face professional development sessions for the first four weeks of the programme, during which staff were trained in their own settings. These sessions are scheduled, in-person, and at times and in formats that best suit practitioners (for example, one to one or in a group). The sessions support educators' understanding of how early maths and executive functions co-develop and they explain how executive functions can be embedded into a range of early maths learning activities while ensuring that children across a range of different ability levels are adequately challenged. The sessions also introduce practitioners to the activity cards. The aim is to help practitioners develop a better understanding of early maths acquisition and the role of executive functioning in maths, to provide practical activities that incorporate this evidence for delivery with young children, and to increase practitioners' confidence in their own ability to run play-based activities that embed executive functions into maths learning. Each of the four sessions had a different theme.

- The first session introduced practitioners to the ONE, executive function, the structure of the PD sessions, and the delivery team.
- The second session focused on executive functions in more detail and detailed the three main components of working memory, inhibitory control, and flexible thinking.
- The third session focused on different kinds of early maths skills.
- The fourth session concluded with the tie between maths skills and executive functioning and how to introduce more executive functioning into maths based activities.

Training and ongoing professional development support was provided by the delivery team and also included opportunity for practitioners to reflect on their implementation of the activities. In addition to providing opportunities for reflection during the initial four-week professional development programme, one representative per setting has a follow-up session four weeks and eight weeks after initial training with the delivery team to allow the team to provide support, check fidelity, and encourage practitioner reflection. These additional reflection sessions are aimed at encouraging practitioners to consciously observe how children engage with the activities and embed executive challenge within activities to scaffold children's development, adjusting the level of challenge where necessary.

Training during the trial took place between January and March 2024. Delivery of the intervention is concurrent with training, so practitioners began implementing the ONE during the first week (that is, post training Session 1).

Materials

Practitioners are provided with 25 activity cards which describe play-based maths activities across three key areas of early years mathematics (numbers and counting, ordering and patterns, and shapes and spatial awareness), all informed by the evidence-basis provided by early years mathematics experts within the extended delivery team. All activities in their basic format include EF challenge.

All activity cards were developed to follow a consistent format for this intervention. This format included direct guidance on how to prepare for the activity, how to carry it out, how to increase executive challenge and differentiate across children, as well as a summary of the key numerical and EF skills involved in each activity. While the level of executive challenge is designed to be gradually increased, there are elements of EF involved in every activity in its most basic format, even without the additional scaffolded challenge. The content for the majority of the 25 activity cards was developed for this intervention by the delivery team (Scerif et al., 2023; Scerif et al., 2025) based on work from other studies (Howard et al., 2018; Moss et al., 2016; Scalise et al., 2017).

Each of the activity cards gives instructions on how to deliver the activities, describes the materials needed, highlights the key mathematical and executive skills they foster, as well as how to gradually increase executive function demands within all of the activities once children and educators are familiar with the basic EF in maths version of each. Some of these activities will be familiar to educators, with additional maths and executive function elements. For example, 'What time is it Mr Wolf?', where a child ('Mr Wolf') stands at one end of the classroom and other children must walk forwards by the number of steps indicated by the 'wolf'. Other activities are likely to be less familiar, such as 'see it, build it, check it' where children are asked to recreate a pattern from memory. These activities are meant to extend the breadth of maths skills that educators can support and are designed to make use of commonly available resources, supplemented by a low-cost resource pack. Explicit guidance on adaptations focused on increasing executive challenge and differentiation across children were presented on each of the activity cards and exemplified during each of the professional development sessions. The overall aim is for practitioners to scaffold children's maths learning at the optimal level of executive challenge, to boost early maths development.

Prior to the trial being implemented (that is, before baseline testing) a refinement phase further adapted some of the materials from the previous study (Scerif et al., 2023). The aim of this refinement was to better support practitioners serving low-income communities. Activity cards were refined to highlight ways of differentiating activities for children who may start off from a lower knowledge basis in mathematics, or children with special educational needs (SEND), or children with English as an Additional Language (EAL). Conceptual clarifications to individual activity cards were also implemented to help educators understand what key elements of the activities they should retain, and how these activities could be differentiated. For example, 'Number Robot' was refined by providing examples of logical rules that started from matching to simple addition and subtraction, facilitating differentiation, while also retaining key mathematics and executive demands.

How—format and dosage

The ONE is a 12-week intervention. Practitioners are asked to implement a minimum of three activities per week within the setting, including one activity from each identified area of mathematics in each week. The activities last five to ten minutes and can be embedded into preschool routines such as small group activities, outdoor play, and free play.

Practitioners begin running these activities from the first week of the intervention, concurrently with the four-week professional development programme. Practitioners have the flexibility to choose how to implement them (big groups, small groups, or a combination), as long as trained staff (that is, those taking part in professional development) and the children in the year preceding the move into reception (that is, eligible children) are included in these activities. The unadjusted instructions on the cards offer a starting point, which can be adapted to reduce the executive functioning challenge so children can access the activity if they are struggling, or have executive functioning challenge added to stretch children once they are comfortable with the activity. Practitioners are given explicit guidance on how to make adaptations focused on increasing executive challenge and differentiation across children on each of the activity cards, and are further exemplified during each of the professional development sessions.

During the trial, children received the intervention between January and June 2024, depending on when staff at settings started the initial training

Where—location

The delivery team recruited settings from West London, East of England, East Midlands, and Yorkshire and Humber. Both the professional development sessions and activities were carried out within the setting.

Evaluation objectives

This evaluation had one primary research question:

RQ1 What is the difference in maths attainment, measured by the Early Years Toolbox Numeracy, of children in the year prior to entering reception in early years settings receiving the ONE intervention in comparison to those in control settings receiving business-as-usual?

This evaluation had one secondary research question:

RQ2 What is the difference in executive functioning, as measured by Heads-Toes-Knees-Shoulders (HTKS-R) and Corsi Blocks, of children in the year prior to entering reception in early years settings receiving the ONE intervention in comparison to those in control settings receiving business-as-usual?

The implementation and process evaluation (IPE) sought to answer seven questions based on the EEF's IPE guidance (EEF, 2022a). Within each research question there were also further subsidiary research questions.

IPE RQ1 To what extent, how, and why was the ONE delivered as planned (including training of practitioners and the implementation of the intervention by teachers)?

IPE RQ1a. Which components of the intervention were delivered with the highest fidelity and which were implemented with the lowest fidelity (and why)?

IPE RQ1b. To what extent does fidelity moderate outcomes of the ONE?

IPE RQ1c. What, if any, adaptations are made to the ONE during implementation? Why are these made (are they logistical or philosophical, pro-active or reactive, in keeping with intervention logic or deviating from it)? What impact did they have on child responsiveness and outcomes?

IPE RQ2 To what extent does variation in attendance of children in settings and engagement of children in activities affect the perceived impact of the intervention?

IPE RQ2a. Do patterns of child attendance at their setting affect the impact of the intervention?

IPE RQ2b. To what extent do children vary in their engagement with the activities, given the free-play environment of most early years' settings, and is there a perceived difference in the impact of the intervention due to engagement with the activities?

IPE RQ3 What are the expected activities, outputs, outcomes, and impacts of the ONE?

IPE RQ3a. To what extent does the ONE result in changes to teachers' knowledge and understanding of executive functions in math and EY skills?

IPE RQ3b. To what extent does the ONE result in changes to teachers' ability to incorporate maths learning into routine, play-based activities?

IPE RQ3c. To what extent does the ONE result in changes to teachers' ability to adapt activities to appropriate levels of challenge based on observed child engagement?

IPE RQ3d. When can outcomes and impacts reasonably be expected to materialise, and what would make them sustainable in the longer term?

IPE RQ4 How do the activities, outputs, outcomes, and impacts of the ONE differ from business-as-usual?

IPE RQ4a. What characterises business-as-usual in settings? How often do they engage in maths and EF activities with the children?

IPE RQ4b. To what extent do settings engage in other structured pedagogical activities with the children (for example, reading and language-based activities, science activities, etc.)?

IPE RQ4c. How often do they receive professional development? Has recent professional development been targeted at numeracy and executive function?

IPE RQ5 To what extent does the ONE result in positive or negative unintended consequences for children, practitioners, and settings?

IPE RQ5a. Does *engagement* with the ONE alter staff retention? Does it place increased pressure on staff?

IPE RQ5b. Does *engagement* with the ONE crowd out other professional development?

IPE RQ5c. Does *compliance* with the ONE reduce the use of other activities (for example, activities designed to target language development and early literacy)?

IPE RQ6 What are the barriers and facilitators to successful implementation?

IPE RQ6a. To what extent, if at all, does the ONE particularly benefit disadvantaged children, compared to business as usual? What are the barriers and facilitators to the ONE benefitting disadvantaged children?

IPE RQ6b. To what extent, if at all, does the ONE particularly benefit EAL children, compared to business as usual? What are the barriers and facilitators to the ONE benefitting EAL children? Where does this fit in the intervention logic?

IPE RQ6c. What are the barriers and facilitators to the ONE improving teaching practice and teachers' knowledge?

IPE RQ7 To what extent does EF function as a mediator, as suggested by the logic model? What evidence is there that EF drives outcomes (that is, can the intervention logic model for EF as a mediator be validated)?

Finally, there was a research question on cost:

Cost RQ What is the cost of delivering the ONE and how does this compare to business-as-usual?

Ethics and trial registration

The trial was registered on the International Standard Randomised Controlled Trial Number (ISRCTN) registry, which is used to describe randomised controlled trials (RCTs) and efficacy trials at inception. The assigned registration number is 69745606.

The ethics and registration processes are in accordance with the ethics policies adopted by RAND Europe and Oxford University. The evaluation is approved by both the RAND U.S. Human Subjects Protection Committee (HSPC) and the University of Oxford Central University Research Ethics Committee (CUREC).

The protocol was published on the EEF website; the statistical analysis plan (SAP) was peer reviewed and also published on the EEF website.¹

Data protection

Several teams are involved in controlling and processing data. RAND acted as controller during data collection, with Qa Research acting as processor. Following the submission of the report, the delivery team will also have access to the data and act as controller. Further details on this are outlined in the data flow diagram in Appendix D.

RAND obtained personal data from settings as a data controller under the lawful basis of 'legitimate interest' under the General Data Protection Regulation (GDPR). Legitimate interest is an appropriate basis because the data collected as part of this evaluation will be used in ways that people would reasonably expect (that is, for the benefit of improving support for executive functioning and early mathematical development in children) and that have minimal privacy impact. Legitimate interests apply where processing is necessary for the purpose of the legitimate interest pursued by the controller (see GDPR Article 6 (1) (f)) and for statistical and research purposes (see GDPR Article 89). RAND conducted a data privacy impact assessment (DPIA) that was reviewed and approved by RAND's Data Protection Officer (DPO) prior to commencing and data collection was in accordance with ICO guidance on processing of child data. A Legitimate Interest Assessment (LIA) was also completed and signed off by RAND's DPO.

RAND obtained outcome data from its testing subcontractor, which acted as a processor following data-sharing terms in the subcontract. At the end of the study, RAND will submit the data in pseudo-anonymised format to the Office for

¹ <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/the-one-project-2022-23-trial>

National Statistics Secure Research Service (ONS SRS) for archiving in the EEF data archive. This data will only be identifiable to the DfE and may be matched to the National Pupil Database (NPD) and other administrative data in subsequent research. The EEF and the DfE will act as data controllers for the archive, along with contractors appointed to manage the archive. The University of Oxford and its collaborator, the University of Sheffield (the delivery team) also have access to the data collected by RAND as data controllers once the evaluation report has been submitted to the EEF. The delivery team rely on 'public interest' as the legal basis for use of the data.

Prior to data collection, settings were given information sheets and asked to sign memorandums of understanding (MOUs) and parents of children at each setting were sent information sheets, privacy notices, and withdrawal forms drafted by RAND and shared with parents/carers by the setting. Parents/carers were given two weeks between when information sheets were sent out and when child data was collected from settings to allow them to withdraw their children from the evaluation before any data was collected. Parents/carers were also informed that they could withdraw their children at any time from the data collection activities and that if they so chose, their children's data would not be collected or would be deleted if already collected.

RAND Europe collected informed consent forms for all practitioners, managers, and trainers that participated in interviews. The front page for each online survey contained a privacy notice informing respondents that participation in the surveys was entirely voluntary. The consent form in the surveys was built into the data collection tool so that those moving past a certain page (following the privacy notice and information on the research) had given consent for their data to be used in this research.

The evaluation team takes information security extremely seriously and all team members have appropriate technical and organisational measures to protect personal data and special category data. Access to information is kept on secure servers and restricted to the named researchers on this evaluation. Data transferred between teams was done via a secure file transfer platform (Egress). The evaluation team collects and stores all personal and special category data in accordance with the Data Protection Act (2018) and UK GDPR requirements. No personal information collected as part of this study was transferred outside of the European Economic Area (EEA).

Research data will be kept securely by the evaluation and delivery teams for the duration of the study and deleted one year thereafter (RAND Europe) or in 2028 (delivery team) to allow for completion of DPhil theses associated with research. Data in the EEF's archive in the ONS SRS will include data only individually identifiable to the Department for Education—the government department responsible for children's services and education—and is kept indefinitely for the purposes of future research. Anonymous data will be kept indefinitely by the University of Oxford.

These aspects were detailed in documents provided for all participants and parents/carers of children in the study, in the memorandum of understanding, information sheets, withdrawal forms, and privacy notices (see Appendices E, F, G).

No member of the evaluation team has any conflicts of interest and all members of the study team approved the protocol prior to publication.

Project team

Delivery team—University of Oxford and University of Sheffield

Project leaders: Gaia Scerif (University of Oxford), Emma Blakey (University of Sheffield).

Research project manager: Caroline Korell.

Research officers: Rosemary O'Connor, Toni Loveridge, Carmel Brough, Sophie Smith, Joanna Archibald, Alicia Mortimer, Siobhan Murray, Alexandra Turner, Hannah Palmer, Molly Staley, Holly Amos.

Research co-investigators: Victoria Simms (Ulster University), Zachary Hawes, Steven Howard, Rebecca Merkley, Fionnuala O'Reilly.

Evaluation team—RAND Europe

Principal investigators and overall project leaders: Elena Rosa Speciani and Marilyn Groom.

Project managers: Marilyn Groom (current), Emily MacLeod and Rachel Hesketh (formerly RAND Europe).

Fieldwork and analysis team: Anna Brian, Rachel Hesketh, Fin Oades, Miguel Subosa, Teresa Turkheimer, James Merewood, Bhavya Singh, Ivana Cardamore (current), and Sarah Angell, Anirudh Agarwal (formerly RAND Europe).

Test administrator: QA Research.

Methods

Trial design

Table 3: Trial design

Trial design, including number of arms		Two-arm, waitlisted, cluster randomised controlled trial.
Unit of randomisation		Early years settings.
Stratification variable(s) (if applicable)		Region (West London, East of England, East Midlands, and Yorkshire and Humber); setting type (private, voluntary, independent (PVI) or school based setting (SBS)).
Primary outcome	Variable	Maths attainment related to acquisition of maths concepts.
	Measure (instrument, scale, source)	Early Years Toolbox (EYT) numeracy measure, 0–120 (Howard et al., 2022). *
Secondary outcome(s)	Variable(s)	Executive functioning (composite measure); executive functioning (visual-spatial).
	Measure(s) (instrument, scale, source)	HTKS-R (composite measure), 0–118 (Gonzales et al., 2021); Corsi Blocks (visual-spatial measure), 0–15 (as used in Blakey et al., 2020).
Baseline for primary outcome	Variable	Maths attainment.
	Measure (instrument, scale, source)	Early Years Toolbox (EYT) Numeracy measure, 0–120 (Howard et al., 2022).
Baseline for secondary outcome(s)	Variable	Executive functioning (composite measure); executive functioning (visual-spatial).
	Measure (instrument, scale, source)	HTKS-R (composite measure), 0–118 (Gonzales et al., 2021); Corsi Blocks (visual-spatial measure), 0–15 (as used in Blakey et al., 2020).

* While both the original Australian programme integrating executive challenge into play-based activities and the EYT numeracy measure were developed by the same lead author (Howard et al., 2020; Howard et al., 2022), the version of the programme to be evaluated in this trial ('The ONE') has been heavily adapted for U.K. early years settings and is fundamentally different from Howard et al.'s 2020 programme. Members of the ONE delivery team have not been involved in the development of the EYT numeracy measure, therefore, there is no conflict of interest between the programme and the primary outcome measure in this trial.

As detailed in Table 3, the trial was designed as a two-arm, waitlisted, cluster randomised controlled trial that primarily assesses the impact of the ONE on early maths attainment among children aged three to four in early years education. The trial was an efficacy trial given the previous level of evidence (an underpowered, developer-led trial).

Given that the intervention has a whole-class focus, randomisation occurred at the setting level, with each setting being allocated to either a group that receives the ONE intervention (the treatment group), or a group that continues with business as usual (the control group). Randomisation was stratified according to region (West London, East of England, East Midlands, and Yorkshire and Humber) so that each region was proportionately represented across both trial arms while also ensuring that the delivery team had an appropriate number of settings per trainer in each region. Stratification was also based on setting type (PVI or SBS) to ensure a similar proportional representation of each type in each region. Having a balance of both setting types in the treatment and control group ensured that findings from the trial were applicable to all setting types.

Outcomes for this trial reflect the intervention's theory of change. The primary outcome is attainment in mathematics and the secondary outcome is EF. Maths attainment is measured by EYT Numeracy (EYT Numbers 2 app; Howard et al., 2022), which measures all three aspects of early maths targeted by the ONE: spatial awareness and shapes, patterning and order, and counting and numbers. EF is measured both by a composite measure, Heads-Toes-Knees-Shoulders revised (Gonzales et al., 2021) and a domain-specific measure, Corsi Blocks (as used in Blakey et al., 2020). EF, as conceptualised in the intervention and its theory of change, consists of three domains—cognitive flexibility, working memory, and inhibition control—so the composite measure is best suited to capturing the overall effect on EF. However, concerns over possible floor effects in HTKS-R, particularly at baseline, led to the inclusion of the domain-specific Corsi Blocks, which has been validated in this younger age group. For further information, please refer to the Outcome Measures section and the protocol (Speciani et al., 2023).

Participant selection

Settings

The trial was open to both school-based settings (SBS) and private, voluntary, and independent (PVI) early years settings. The recruitment goal for Stronger Practice Hub trials was to include at least 30% of settings from both the SBS and PVI sectors. Settings with fewer than ten children within the relevant age range enrolled in September were initially waitlisted and were included on a case-by-case basis if necessary to reach the recruitment target of 150 settings.

The delivery team recruited settings from four regions: West London, East of England, East Midlands, and Yorkshire and Humber. The recruitment used multiple complementary strategies:

- sending out direct emails to all eligible educational establishments within the specified local authorities (LA), whose contact information is publicly accessible;
- engaging LA early years specialists to assist with recruitment;
- actively reaching out to educational establishments in low-income areas—identified by an Index of Multiple Deprivation (IMD) score of less than five—to increase participation from settings eligible for Early Years Pupil Premium (EYPP); and
- collaborating with Stronger Practice Hubs to extend recruitment efforts through their networks.

Settings could only take part in one SBH programme and could not be involved in another trial that included the same children and the same outcomes of interest (that is, mathematics and EF). Settings could not take part in both the evaluation of the ONE and the DfE early years professional development programme in the same year. However, during the baseline and delivery period, some settings were offered the Maths Champions intervention, having previously been control settings for the trial. The EEF monitored the sign-up list for the Maths Champions intervention, with any settings already enrolled in the ONE asked to delay participation in Maths Champions until June 2024, or April 2024 at the earliest if delaying until June is not feasible. Given endline testing took place in most settings during April and May 2024, delaying participation until June 2024 mitigated the possible contamination effects of Maths Champions participation on the ONE trial. Nevertheless, RAND Europe was provided with a list of all settings participating in Maths Champions by the EEF so that additional sensitivity analysis could be included to allow for possible spillover effects.

Early years settings deliver their activities in many ways. Some are single room settings where children of different ages mix, while others have separate 'classes' for different age groups. Settings with more than one eligible class (that is, where eligible children were located in more than one classroom) were asked to nominate a classroom for testing, but the intervention was delivered across the whole setting (with trained practitioners).

Children

All children were eligible to take part in the intervention, however, for the purposes of the evaluation, settings were encouraged to focus activities on children who were in the year preceding the move into reception (three- to four-year-olds). There were no exclusion criteria on the basis of SEND or EAL status. Children were assessed at baseline regardless of SEND or EAL background, given that neither were accurately nor consistently recorded in early years settings. The lack of exclusion on the basis of SEND and EAL status mimicked the design of the ONE intervention as a

whole-class intervention suitable for all children within the age range. However, if children refused, or were evidently unable, to engage with any of the assessments at baseline, they could not be included in the trial due to a lack of valid baseline results.

Settings provided parents or carers of eligible children with a parent information sheet and withdrawal form prior to baseline data collection. This provided carers with an opportunity to request for their child's data not to be collected or used as part of the trial. They were also informed of their right to withdraw their child's data from the evaluation at any time. While classrooms may have included children who would not be attending school the following year, younger children were not in scope for this evaluation.

To ensure adequate power (see Sample Size calculations), settings with 15 or more eligible children were prioritised for inclusion. To remove bias that could be introduced by assessors selecting children not-at-random, testers conducting assessments in settings with more than 15 eligible children were provided with a randomised class list by the evaluation team and instructed to assess children in the order they appeared on the list.

While many of the baseline assessors were able to test according to the randomised class list, others appeared to not follow the randomised class list. It was planned that only children within the first 15 children included in the randomised list generated by RAND Europe would be tested at baseline. In 17% of settings, over a third of children assessed were outside of the first 15 children included in the randomised list; in total, this represents 310 children. Therefore, we cannot rule out the possibility of non-random selection in some settings. The sample population was also affected by data loss by the independent test administrators at baseline, with failure to upload all primary outcome data for one setting and partial primary outcome data for another two settings; while regrettable in adding to attrition overall, it does not introduce any further risk of non-random selection.

We acknowledge, as a result, that there is the possibility of bias in the baselined sample due to this limited, non-random selection that we cannot control for within our analysis. To assess whether selection into the trial appeared to be truly random at the child-level, when examining balance at randomisation of the sample, we have examined whether the baselined sample are significantly different from the sample available for baseline with regard to available setting-provided covariates, such as gender, age, EYPP status,² and EAL status. This is outline in Table 4. Unfortunately, there are no accurate and readily available indicators of SEND for this age group and this information was not collected from settings. In addition, experience from other RAND Europe-led EEF trials indicates that a low prevalence of SEND status may make analysis under-powered. Therefore, we were not able to include this in our analysis.

Table 4: Pupil characteristics—comparison between sampled and non-sampled children with p-values for non-zero difference

Pupil-level characteristic	Intervention group					Control group				
	Sampled		Non-sampled		P-value	Sampled		Non-sampled		P-value
	n/N (missing)	Count (%)	n/N (missing)	Count (%)		n/N (missing)	Count (%)	n/N (missing)	Count (%)	
EYPP	123/961 (12)	12.8	115/651 (19)	17.7	0.0084	112/962 (19)	11.6	100/738 (9)	13.6	0.2423
EAL	266/954 (19)	27.9	219/658 (12)	33.3	0.0213	285/970 (11)	29.4	286/742 (5)	39.5	0.0001
Gender (male)	475/960 (13)	49.5	362/667 (3)	54.3	0.0570	492/981 (0)	50.2	374/746 (1)	50.1	0.9938
	Sampled		Non-sampled		P-value	Sampled		Non-sampled		P-value
	N (missing)	Mean	N (missing)	Mean		N (missing)	Mean	N (missing)	Mean	

² While we collected EYPP data at baseline, we are aware that doing this can lead to incomplete data given that settings cannot confirm EYPP until the spring term. Therefore, the evaluation team collected EYPP data in the summer term (at endline). The data collected at endline is more likely to be complete since it will have been confirmed during the spring term, and this data was used during the analysis.

Age (months)	973 (0)	43.6	670 (0)	43.8	0.2874	967 (14)	43.5	737 (10)	43.6	0.3416
--------------	------------	------	------------	------	--------	-------------	------	-------------	------	--------

The evidence is suggestive of some non-random sample selection at the child level. This analysis suggests that EAL children were underrepresented in the analytical samples of both intervention and control settings, with the proportion in the analytical sample being significantly lower in both trial arms, with this difference larger for the control arm than the treatment arm. There is also evidence for imbalance between the proportion of children who are EYPP between the sampled and non-sampled pupils in intervention settings, which does not appear to impact control settings. However, the data for this does not provide the complete picture: EYPP baseline data is unreliable as it was collected before many families will have applied for EYPP. While we only collected endline EYPP data for children that were tested at baseline, we cannot compare the sampled and non-sampled children using more reliable EYPP figures. There is some evidence of a slight gender imbalance in intervention schools, with the difference in the proportion of male children in intervention schools larger than it is in control schools. There is no indication that age influenced sampling in either treatment group.

While there are concerns about non-random selection of children, whether due to failure to follow guidance by test administrators, as outlined above, or due to higher absence rates in the target population, there are fewer concerns regarding imbalance across the two trial arms. Statistics reported in Table 10 suggest that the difference in the EYPP eligibility (based on more recent endline data, which is more reliable) and gender between the treatment groups is small, despite potential for selection bias. Furthermore, there appears to be little difference (associated with a p-value of 0.4674) in the proportion of EAL pupils between intervention (27.9%) and control settings (29.4%), suggesting that the analytical sample is balanced on EAL, even in the presence of non-random selection at the baseline sampling stage.

The number of hours a child attended was also not an exclusion criterion at the setting level as the intervention is a whole-class intervention delivered to all children regardless of attendance pattern. However, given the attendance patterns of nursery children vary, not all who attended the setting were in attendance during baseline assessment and, therefore, not included in endline assessment. Baseline assessment was conducted on at least two different days of the week to minimise this attrition.³ For endline, settings were asking to share updated attendance patterns so that endline assessments could be conducted on days when the majority of the baseline children were in attendance. Additional 'mop-up' days were included in the design to minimise attrition due to assessors visiting settings on days when children were not in attendance.

Outcome measures

Baseline and endline assessments consisted of the following:

- mathematics attainment using the Early Years Toolbox's numeracy (EYTN) subset (EYT Numbers 2 app); and
- executive functioning using two measures: Heads-Toes-Knees-Shoulders Revised (HTKS-R) and Corsi Blocks; while HTKS-R is a composite measure (measuring multiple components of executive function), Corsi Blocks is a domain-specific measure of visuo-spatial memory.

Baseline and endline testing was conducted by Qa Research with trained, blind-to-allocation administrators, on a one on one basis, in person. Assessors were provided training by Qa Research at baseline. As a result of data loss, incorrect administration, and measurement error at baseline, prior to endline tests being administered assessors were provided with additional training by Qa Research, with support from the delivery team, to support the administration of the HTKS-R and Corsi Block measure.⁴

³ With the exception of one smaller setting, where all children in the setting were in attendance and could be baselined on that day.

⁴ The initial and additional training used indirect training methods, where assessors were trained in the assessment of the measures through role-play with QA researchers, rather than by observing QA researchers undertake the assessment with children.

Baseline assessment was, for most settings, completed prior to randomisation (see discussion in Randomisation section for further details). Measures were administered in a fixed order for all children: EYTN, Corsi Blocks, and HKTS-R. Given the young age of the children and that shorter attention spans could introduce attrition in measures over the course of the 30-minute assessment, to ensure primary data is available for many children as possible EYTN was prioritised as the first assessment for all children. To break up the longer assessments (EYTN and HKTS-R), the shortest assessment, Corsi Blocks, was placed second to provide variety and keep children engaged. Ultimately the fixed order of assessments prioritised maintaining child engagement and administration ease, but has introduced possible order effects into the measures. Each of the measures are discussed in more detail below.

Primary outcome measure

Attainment in mathematics was evaluated using the Early Years Toolbox (EYT) Numbers 2 app (Howard et al., 2022). This numeracy subtest of the EYT (EYTN) consists of 120 items covering number sense, counting, numerical operations, spatial concepts, and patterning (Howard et al., 2023). EYTN is thus well suited to this evaluation as a primary outcome as it covers all three mathematics domains targeted by the ONE intervention: spatial awareness and shapes, patterning and order, and counting and numbers. Despite the absence of U.K.-specific norms for the EYTN, its use is justified by its validation on Australian children, with EYTN exhibiting good validity, reliability (test-retest reliability, $r = 0.89$), and sensitivity to developmental changes (Howard et al., 2022).

The EYTN has the additional advantage of being straightforward to administer. Instructions and stopping rules are integrated into the iPad app, thus no decision-making is required of the administrator. The test automatically adapts to each child's age, as entered by administrator, and ends after five consecutive incorrect responses,⁵ typically lasting seven minutes (Howard et al. 2022). The administrator's role requires resolving technical issues, recording responses by the child accurately, overcoming shyness, and ensuring the child is attending to the task. Data collected is held locally within the app on the device until it can be directly uploaded to a GDPR-compliant cloud. Despite this ease of administration, there were some issues at baseline and endline collection. Due to failure to upload locally-held data before wiping devices, the independent test administrators lost all EYTN data for one setting and partial data for another. It was noted during baseline that 20 individuals had faster than expected administration times (under one minute); however, it is not clear whether these fast administration times were entirely due to test administration rather than true zeros so, as outlined in the statistical analysis plan (Speciani et al., 2024), this small number of fast tests were ultimately retained in analysis. By comparison, at endline, only four individuals had administration times under one minute.

Unfortunately, a bug appeared in the EYT app between baseline and endline, which unfortunately affected a minority of EYTN scores. This bug was caught by the evaluation team during the endline period and a patch implemented by the developers to fix the issue. Nevertheless, it did affect a minority of EYTN scores (276 endline scores in total, which is 16.32% of the endline scores in the analytical sample).

Further details on EYT Numbers 2 (the EYT numeracy subtest) are outlined in the protocol (Speciani et al., 2023).

Secondary outcome measures

Executive functioning was measured as a secondary outcome using two tests: the Heads-Toes-Knees-Shoulders Revised (HTKS-R) and Corsi Blocks. HTKS-R is a composite measure of executive functioning suitable for four-year-olds, assessing all three executive functioning components (working memory, inhibitory control, cognitive flexibility). Corsi Blocks is a domain-specific measure, testing visuo-spatial working memory, and is well validated for three- and four-year-olds and predictive of early maths outcomes (Blakey et al., 2020).

HTKS-R

HTKS-R integrates multiple executive functioning domains into a single game-like measure (McClelland et al., 2021). The game introduces behavioural rules, where children are asked to do the opposite (for example, 'when I say touch your head, you touch your toes'). It includes four parts with increasing complexity through the changing or introduction

⁵ Due to the stopping rules (test ending after five consecutive incorrect responses), not all children will receive questions covering all these mathematical concepts as they will not progress sufficiently to allow complete coverage of concepts.

of new rules and a scoring system that awards points for correct responses and self-corrections (Gonzales et al., 2021; McClelland et al., 2021). All children complete the first two parts—a spoken part, without any gross motor demands, and the first action-based sequence. For Parts II, III, and IV, the child is required to reach a score of at least four points to continue to the following part. The measure, as validated in Gonzales et al. (2021) and McClelland et al. (2021), aggregates responses to both the practice rounds and test rounds, but the continuation rules only consider the scores on test rounds for each part. For this evaluation, we similarly aggregated scores according to the process used in the validation studies (Gonzales et al., 2021; McClelland et al., 2021). Using this approach, incorrect responses are scored zero, self-corrected responses are scored as one, and correct responses as two for each item in practice and test rounds, with aggregated scores ranging from zero to 118 for HTKS-R.⁶ HTKS-R is short, taking just five to seven minutes to complete, and straightforward to administer (Gonzales et al., 2021; McClelland et al., 2021).

After consultation with the delivery team, a puppet was introduced as a prop for illustrating the rules before baseline administration with the aim to mitigate possible floor effects for the youngest age groups. The delivery team piloted this approach and reported it helped with engagement and understanding for the youngest children. We include a full distribution of HTKS-R at baseline and endline (see Appendix D) and compare the distribution at endline in this trial with the distributions documented in the original papers (Gonzales et al., 2021; McClelland et al., 2021), which did not use puppets, to ascertain whether the inclusion of puppets may have altered the statistical properties or validity of the test.

HTKS-R is well-suited to this evaluation as it reflects the broad conceptualisation of executive functioning in the intervention and theory of change, rather than focusing on a single domain. Moreover, it is strongly correlated with other measures of executive functioning (Gonzales et al., 2021), predictive of young children's academic achievement (McClelland et al., 2021), and displays construct and predictive validity (McClelland et al., 2021). It is preferred over HTKS due to fewer floor effects in younger, socioeconomically diverse children (Gonzales et al., 2021; McClelland et al., 2021). However, while it is well-validated in four-year-olds, HTKS-R has not been validated in three-year-olds, so concerns over possible floor effects for the youngest children at baseline remain.

An examination of baseline data suggests there was high attrition from HTKS-R, compared to EYTN, and substantial floor effects in HTKS-R. Although these issues were substantially reduced at endline, the floor effects persisted, albeit to a lesser degree, affecting 3.03% of the analytical sample. These issues may have been due to a number of factors: it is a longer assessment, it was the last assessment faced by the children, and it was more complex for assessors to administer. Given there was lower attrition and less likelihood of floor effects for Corsi Blocks, and evidence from the literature that working memory is moderately correlated with both HTKS-R and early maths achievement (McClelland et al., 2021), we use Corsi Blocks as the baseline measure of executive functioning in the headline model for secondary outcome analysis (see Analysis). We additionally report two measure-specific models alongside this headline model to examine the degree to which estimated effect sizes depend on how executive functioning is measured. The pre-test/post-test correlations of these two models are likely to be higher than the mixed-measure model (HTKS-R at endline and Corsi Blocks as baseline), but the sample size is likely to be lower given higher attrition in HTKS-R at baseline. The pre-test/post-test correlations for all models are provided in appendix Table 10.

Corsi Blocks

Given the possibility of floor effects among youngest children assessed at baseline, Corsi Blocks, an alternative measure which focuses on visuo-spatial memory (Corsi, 1972; Arce and McMullen, 2021) was also used as a measurement of EF. This task involves a child replicating a sequence of block taps demonstrated by an assessor, starting with just two blocks in each sequence. The test builds complexity by increasing the sequence length by one block each time until the child cannot recall two out of three sequences. While domain-specific, Corsi Blocks is validated for young children, correlates with executive functioning and maths ability, and remains predictive of maths performance in the nursery years (Blakey et al., 2020).

The measures of executive function are not combined in this evaluation. Generating a single latent factor model of executive functioning from the two measures was considered, however, previous examination of the factor structure for HTKS-R and executive functioning in the early years suggests that the best model fit is a one-factor solution (Gonzalez et al., 2021) and that HTKS-R is the only measure available which is a consistent independent predictor of early maths

⁶ Very low scores of between zero and four are indicative of failing even the practice rounds of HTKS-R.

achievement (McClelland et al., 2021). Evidence suggests that as a single measure, the original HTKS can perform similarly or more strongly than individual measures of executive functioning (McClelland et al., 2014; Lipsey et al., 2017), and provides an efficient composite measure in terms of the predictive relationship between executive functioning and early academic achievement (Lipsey et al., 2017). There is little evidence to suggest that augmenting HTKS-R with a single domain-specific measure of working memory would create a more efficient and predictive estimator for the purposes of this evaluation.

The inclusion of Corsi Blocks was to guard against possible floor effects in the youngest or most disadvantaged children, particularly at baseline, given that previous evidence suggests HTKS—and, to a lesser extent, HTKS-R—may suffer from these issues (McClelland et al., 2021; Gonzalez et al., 2021). However, histograms of Corsi Blocks, included in Appendix B, showed evidence of floor effects at baseline. These floor effects were reduced at endline, affecting only 3.33% of the overall analytical sample, but not eliminated. While there were fewer issues with test administration of Corsi Blocks at baseline than with HTKS-R, there is still some cause for concern that these floor effects are due to administration errors by the independent test administrators (such as incorrectly coding incomplete responses), rather than due to the age of the children. Given these issues, caution should be used in interpreting the secondary outcome analysis in this evaluation.

Baseline

Given there is no statutory requirement to collect academic administrative data for this age group, we included baseline tests. Tests at baseline allow us to improve pre-test/post-test correlation estimates, which in turn will improve power with a smaller number of settings and also allow us to explore differential attrition, if necessary. We used the same tests at baseline and endline in order to maximise pre-test/post-test correlation and improve statistical power.

There was high attrition from HTKS-R and, to a lesser extent, Corsi Blocks, compared to EYTN (see Analysis section for more discussion) and test administrators faced difficulties administering the executive functioning tests at baseline (for example, following stopping rules or correctly coding responses). Additional training was implemented alongside quality assurance checks to improve administration at endline, but we do acknowledge that measurement error at baseline could impact on findings. At baseline, 40.04% of Corsi Blocks scores were within one standard deviation of the floor, and 20.02% of HTKS scores were within one standard deviation of the floor. However, in the HTKS test, there were substantial issues with the recording of the data, which introduced ambiguity as to whether the results were valid: for example, 352 (18.7%) tests were marked as invalid by assessors despite the stopping rules apparently being followed, and 63 (3.34%) tests where the stopping rules were not followed correctly were marked as valid (see Analysis section for additional discussion).

Sample size

The minimum detectable effect size (MDES) for this study was calculated using a two-level random assignment design, to reflect the design of the trial, with randomisation occurring at the setting level and analysis occurring at the individual level. In calculating the MDES, a number of assumptions were used: randomisation at the setting level with 50:50 allocation, alpha of 0.05, and power at 0.8. All MDES calculations were made using PowerUp! (Dong & Maynard, 2013).

As can be seen in Table 8, at protocol stage, we assumed an average of 12.5 children per setting and pre-test post/test correlations of 0.8 at the child level and 0.2 at the setting level. In line with other published early years trials at the time, we assumed an intra-cluster correlation (ICC) of 0.18. The MDES at protocol stage was 0.204. At randomisation we updated the number of children per setting to reflect the number of children baselined using EYTN (1,859 children overall, 12.4 children per setting) with all other assumptions remaining the same.⁷ The MDES at randomisation remained 0.204.

⁷ The published statistical analysis plan (SAP) had calculated 13.3 children per setting. This was adjusted after data cleaning had removed incorrect data (for example, double entries, incomplete baseline).

At analysis the MDES was calculated using data from the trial: pre-test/post-test correlations of 0.68 at the child level, 0.5 at the setting level, an ICC of 0.188, and an average of 11 children per setting.⁸ This places the MDES at analysis at 0.196. This is slightly more sensitive than what had been planned. This is due predominately to the improved pre-test /post-test correlation at the setting level.

As is standard in EEF trials, we also looked at the impact on a subgroup of children from disadvantaged backgrounds. In early years interventions, disadvantage can be operationalised by the number of children in receipt of Early Years Pupil Premium (EYPP). At protocol stage, we estimated that the average number of three- and four-year-olds registered for EYPP in England and eligible in our sample to be 2.5 per setting.⁹ Assuming the intervention settings are representative of those across England, we thus estimated that 360 children in the sample at protocol stage would be in receipt of EYPP. Keeping the same assumptions for this subsample as for the whole sample, this provided an MDES for EYPP children of 0.250. At randomisation, the average number of children reported as eligible for EYPP was substantially below expectations, at just 1.5 per setting, leading to an MDES of 0.280. At analysis the MDES for pupils registered as EYPP was calculated using data from the trial:¹⁰ an ICC of 0.188 and an average of 11 children per setting. For the pre-test/post-test correlations a negative R-squared of 0.54 at the setting level was found, precluding the possibility of using the correlation in the MDES calculation. This suggests that in our small sample of EYPP pupils, the full model of endline scores in the sub-sample, which captures the effect of setting characteristics, explains the data poorly compared to a null model that does not. In essence, due to the restricted nature of our sample, on average the variance in the outcomes between similar settings is greater than the variance between all settings. As such, we proceeded with the MDES calculations with two different estimates for this figure. The most conservative estimate would be to use zero for the setting level. This yields a MDES of 0.242. However, being less pessimistic we also ran the MDES using the setting-level estimates for all pupils. This gives an MDES of 0.235.

This places the MDES at analysis between 0.235 and 0.242. This is slightly more sensitive than that planned. This is due predominately to the improved pre-test/post-test correlation at the child level.

Randomisation

Randomisation of the settings to one of the treatment arms took place on 1 December 2023. In total, 150 settings were randomised to either intervention or control group, with randomisation occurring with a 50:50 allocation resulting in 75 randomised to intervention and 75 to control. The settings were the unit of randomisation, but children are the unit of analysis, reflecting the clustered-RCT design of this evaluation. The nature of the intervention, which involves professional development for the educators and group run activities in free-flow playroom environment, means that one cannot avoid contamination between groups within a setting, thereby making individual-level randomisation unfeasible and a clustered design more suitable.

Randomisation was stratified by region (West London, East of England, East Midlands, and Yorkshire and Humber) and setting type (private, voluntary, independent, PVI, or school based setting, SBS). A stratification by region helps ensure balance between control and treatment group since key covariates (such as EYPP eligibility) are likely to vary across the region and the recruitment is organised regionally. Stratification by setting type is crucial given there may be some differences between setting types regarding staff qualifications, availability of additional and specialist services, and differences in proportion of EYPP-eligible children (Paull and Popov, 2019; Bonetti, 2020).

The randomisation was conducted by a member of the evaluation team who was provided with meaningless setting identifiers so that they were blind to the setting identities. A tailored package in Stata, `randtreat`, was used to implement the settings randomisation with regional and setting type stratification. A second senior researcher at RAND Europe then checked the randomisation code and the outcome to verify independence. The code used to randomise

⁸ We note that here, and with the raw correlations in Appendix Table 10, the correlation between baseline and endline EYTN is substantially lower than the test-retest correlation of $r = 0.89$ reported in Howard et al. (2022). This is likely due to the intervening time period between baseline and endline (approximately six months), as compared to the test-retest duration of one week in Howard et al. (2022).

⁹ The Department for Education reports that 116,500 three- and four-year-olds were in receipt of EYPP in 2022 across 47,121 providers. Source: <https://explore-education-statistics.service.gov.uk/find-statistics/education-provision-children-under-5>

¹⁰ These estimates are based on the 98 settings in which there are valid test statistics for EYPP pupils.

settings can be found in Appendix N. A master copy of the final allocation was retained in a locked folder on RAND Europe's servers to prevent editing and the final allocation communicated to the delivery team checked against it to ensure no edits occurred in the processing or transfer of data.

Settings allocated to treatment were expected to deliver the intervention in the 2023/2024 academic year. Those allocated to control were expected to carry on with business as usual that year and receive the training in 2024/2025 given the study's waitlisted design. As these children are expected to transition to primary school by the start of 2024/2025, they will not be exposed to the intervention, ensuring the waitlist design does not interfere with the potential for longitudinal analysis in future (outside the scope of this current evaluation).

Originally, randomisation was to occur after all baseline testing had been completed (see Speciani et al., 2023). However, during testing it became clear that more time was needed to conduct testing across all settings owing to setting-level factors (such as illness and Ofsted visits) and test administrator availability. To increase time for baseline testing, while allowing the delivery team to contact settings to arrange professional development sessions for January, a decision was made by the evaluation team, the EEF, and the delivery team to use concealed randomisation. The following conditions needed to be met:

- at least 90% of all settings had finished baselining by the randomisation date; and
- all settings included in the randomisation had to have provided child data (names and dates of birth at a minimum).

Randomisation was initially concealed from all settings that had not completed baseline assessment (N = 14). However, for two, treatment allocation was revealed to the setting (but not to baseline assessors) by the delivery team prior to the conclusion of baseline assessment. One of these was a treatment setting that had started but not completed baseline assessment prior to its allocation being revealed and the other a control setting that had not undergone any baseline assessment prior to allocation being revealed. After consultation with the EEF and the delivery team, it was agreed that we would conduct additional sensitivity analysis of the primary outcome, excluding all children at these settings baselined after allocation was revealed. If the sensitivity analysis indicated that inclusion of these children significantly alters the estimated treatment effects, primary outcome estimates would have been reported without these children included.

Table 5 outlines the actual allocations for the overall sample of participating setting by stratification variables. In total, 75 settings were each allocated to the control and intervention groups. The numbers of settings in the trial varied by the stratification regions, from 22 settings in the Yorkshire and Humber (where initial recruitment was limited to fewer and heavily recruited upon LAs) and 44 in West London. As expected, the randomisation produced as equal an allocation to the intervention and control group as possible among the overall sample of participating settings across the regions. The number of settings in the trial varied by setting type as well, with 83 (55%) settings classed as PVI and 67 (45%) classed as SBS, meeting the EEF's target of at least 30% of each type.

Table 5: The ONE randomisation results

	Control	Intervention	Total settings
Region			
West London	22	22	44
East of England	20	21	41
East Midlands	22	21	43
Yorkshire and Humber	11	11	22
Setting type			
PVI	42	41	83
Maintained	33	34	67

Statistical analysis

Primary analysis

As detailed in the protocol (Speciani et al., 2023), this efficacy trial has one primary research question:

RQ1 What is the difference in maths attainment, measured by the Early Years Toolbox Numeracy, of children in the year prior to entering reception in early years settings receiving the ONE intervention in comparison to those in control settings receiving business-as-usual?

To address the primary research question, we used a multilevel model of mathematical attainment, the primary outcome for this efficacy trial, on an intention-to-treat (ITT) basis. As outlined above, mathematical attainment was assessed by the Early Years Toolbox Numeracy subtest (Howard et al., 2022). Under an ITT approach, analysis included all randomised settings and baselined children, grouped according to random assignment, regardless of programme compliance or treatment dosage. It is an inherently conservative approach, estimating the average effect of offering the intervention, and is key to ensuring an unbiased analysis of intervention effects in line with EEF guidance (see EEF, 2022b).

More specifically, using multilevel modelling (MLM), we estimated a two-level random-intercept model. The two-level model, with the first level the unit of analysis (the child) and the second the unit of randomisation (the setting), reflects the trial design and nested nature of the data, as recommended by the EEF (EEF, 2022b). It appropriately clusters the error term at the unit of randomisation to ensure appropriate and unbiased confidence intervals are estimated. The two-level, random-intercept model estimates a single average treatment effect on mathematical attainment across settings while allowing for setting-specific variation in mean mathematical attainment. By adopting a clustered two-level model, we allow for this potential setting-level heterogeneity.

The impact was estimated using the model outlined below in Equation 1. Equation 1 is known as a random-intercept model because the setting-specific intercepts for each setting j ($\beta_{0j} = \beta_0 + u_j$) vary randomly with the setting-level residual ($\beta_{0j} \sim i.i.d N(\beta_0, \sigma_u^2)$). The model additionally controls for pre-test (baseline) attainment, as measured by pre-test EYT Numeracy scores, and estimate fixed effects for stratification variables (region, setting type) at the setting level.

$$(1) \quad Y_{ij} = \beta_0 + ONE_j\tau + \beta_1 Z_j + \beta_2 X_{ij} + u_j + e_{ij}$$

where:

Y_{ij} is the EYT numeracy score for child i in setting j , at endline;

β_0 = cluster-level coefficient for the slope of a predictor on number skills;

ONE_j = binary indicator of the setting assignment to intervention [1] or control [0];

Z_j = setting-level characteristics, that is, the stratifying variables of geographical location and setting-type (as used for randomisation);

X_{ij} = child-level characteristics for child i in setting j , or more, specifically the baseline EYT numeracy score;

u_j = setting-level residuals; and

e_{ij} = child-level residuals.

The coefficient τ is the outcome of interest, as an estimate of the conditional effect of treatment on endline EYT numeracy scores. We then calculated a standardised effect size using Hedges' g for τ (more in Calculation of Effect Sizes).

The use of the raw scores for EYT Numeracy follows EEF guidance (EEF, 2022b) as age-standardised scores are not recommended by the developer. The age-adjusted starting rule does not appear to affect the validity of the raw scores

(Howard et al., 2022).¹¹ Despite using raw scores instead of age-standardised scores, the estimated average treatment effect should remain unbiased even without the inclusion of age, as long as there is balance in age across treatment arms at baseline. To check this assumption we conducted a sensitivity analysis including age in Equation 1 (see Additional Analysis).

All analyses were done in R.

Secondary analysis

As outlined in the protocol (Speciani et al., 2023), this study answers the following secondary research question:

RQ2 What is the difference in executive functioning, as measured by Heads-Toes-Knees-Shoulders (HTKS-R) and Corsi Blocks, of children in the year prior to entering reception in early years settings receiving the ONE intervention in comparison to those in control settings receiving business-as-usual?

As outlined in the Outcome Measures section above, this trial has two secondary outcome measures: one is a composite measure of the child's executive functioning (HTKS-R score) while the other is the measure of visual spatial ability of the child (Corsi Block score). While the HTKS-R, as a composite measure, is better suited to the theory of change, a possibly larger incidence of floor effects given the age of the children, particularly at baseline, prompted the collection of an additional domain-specific measure, Corsi Blocks. There are no plans to combine these two measures into a single latent EF measure using structural equation modelling. (For justification of this approach, see Secondary Outcome Analysis section above.)

For the secondary analysis, we used the same multilevel modelling approach as in the primary analysis, that is—more specifically—we estimated a two-level random-intercept model (see Primary Outcome Analysis section for justification). As with the primary outcome analysis, the model accounts for baseline achievement, determined by pre-test scores, and estimates fixed effects for variables used in stratification (region and setting type) at the level of each setting. In all models, raw scores were used for both baseline and endline EF tests.

$$(2) \quad Y_{ij} = \beta_0 + \text{ONE}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij}$$

where:

Y_{ij} = EF score for child i in setting j , at endline (either endline HTKS-R or endline Corsi Blocks);

β_0 = cluster-level coefficient for the slope of a predictor on number skills;

ONE_j = binary indicator of the setting assignment to intervention [1] or control [0];

Z_j = setting-level characteristics, that is, the stratifying variables of geographical location and setting-type (as used for randomisation);

X_{ij} = child-level characteristics for child i in setting j or, more specifically, the baseline EF score (either baseline Corsi Blocks or baseline HTKS-R);

u_j = setting-level residuals; and

e_{ij} = child-level residuals.

As with the primary outcome model, the coefficient τ is the outcome of interest. We then calculated the effect size using Hedges' g for τ (more in Calculation of Effect Sizes).

¹¹ Howard et al. (2022) found a correlation of $r = 0.97$ in a sample of 126 children aged three to five between the raw scores under the full scale (where children answered all questions regardless of age or ability) and the raw scores using the age-adjusted starting rules and performance-based stopping rules.

We ran three different secondary outcome models, with the first being presented as the headline secondary outcome model, as outlined in Secondary Outcome Measures:

1. a mixed-measures model, which uses raw HTKS-R scores at endline for Y_{ij} and raw Corsi Blocks scores at baseline for X_{ij} ;
2. a HTKS-R model, which uses raw HTKS-R scores both at baseline (X_{ij}) and endline (Y_{ij}); and
3. a Corsi Blocks model, which uses raw Corsi Blocks scores both at baseline (X_{ij}) and endline (Y_{ij}).

As a composite measure, HTKS-R is best suited to the evaluation of the ONE intervention's theory of change. The first two models use this composite measure at endline, with the generated effect sizes interpreted as the effect of the intervention on overall EF. This ensures the headline effect size is aligned with the conceptualisation of EF in the theory of change and for this reason is our preferred endline measure. Appendix Table 10 shows the second model has a higher correlation between endline and baseline test scores, given it uses the same measure at both timepoints, the presence of higher rates of attrition in HTKS-R at baseline raised concerns about the available sample size and power of the HTKS-R model (see discussion in Secondary Outcome Measures). For this reason, we specified a mixed-measures model as the headline model, mimicking the primary outcome analysis model outlined in Equation 1, using HTKS-R at endline and Corsi Blocks at baseline given these measures are correlated (as shown in Appendix Table 10 and Speciani et al., 2024).¹² This ensures our headline secondary outcome model has the highest possible sample size, to improve the power of the analysis.

As outlined in Secondary Outcome Measures, we were additionally concerned about potential floor effects in HTKS-R, particularly at baseline given the age of the children. The Corsi Blocks model allows for a comparative analysis (albeit domain-specific) to understand how large the floor effects might be and how they might influence the estimated effect size. However, given the presence of floor effects in Corsi Blocks as well, all three secondary outcome models may suffer from similar issues. We report these effects and have interpreted our findings with an understanding of their potential impact on the efficacy assessment of the intervention.

We noted above (Secondary Outcomes) that missingness and attrition is higher in the secondary outcomes than for the primary outcome at baseline. Therefore, we caveat all findings appropriately and ensure this secondary outcome analysis is accompanied by a series of sensitivity analyses, as outlined in the statistical analysis plan (Speciani et al., 2024) and below.

Given we have multiple secondary outcomes, we employ a Romano-Wolf correction to statistically correct for over-rejection of null hypotheses under multiple hypothesis testing. Since the HTKS-R measure is a composite measure of EF, and the Corsi Block measure is a domain specific measure of EF, they are likely to be at least moderately correlated. In such circumstances, the Bonferroni correction might be too conservative and lead to an over-correction. For this reason, we will apply the Romano-Wolf correction in secondary outcomes analyses. This takes into account the dependent nature of the test statistics and provides a strong control against the family-wise error rate (Clarke et al., 2019) as recommended in the EEF's evaluation guidance (EEF, 2022).

All analyses were done in R. Multiple hypothesis testing corrections were made using `crctStepdown` in R (Watson, 2024).

Analysis in the presence of non-compliance

Compliance

As the ITT approach is inherently conservative, capturing the average effect of offering the ONE intervention, we also estimated treatment effects for complying settings. The ITT approach, while reducing bias associated with non-random attrition, may dilute the estimated effect of an intervention due to non-compliance. Therefore, we used complier average causal effect (CACE) analysis to measure the effect of the programme among settings that were fully compliant with the intervention. This measures the average effect of fully compliant participation in the ONE on numeracy outcomes.

¹² Correlation at baseline is 0.35.

Participation in the ONE intervention requires settings to both participate fully in a series of professional development sessions over the course of the 12-week intervention and implement the intervention activities three times a week over the intervention period. After discussions with the delivery team, the EEF, and the evaluation team it was decided that full participation in the professional development arm of the training was seen as the necessary pre-condition to successful implementation. Therefore, compliance was a binary measure with full compliance with the intervention defined as at least one staff-member from each setting participating in each of the professional development sessions.

We undertook a complier average causal effect (CACE) analysis using a two-stage least squares (2SLS) estimation with random group allocation serving as the instrumental variable (IV) for the compliance indicator following the EEF guidance (EEF, 2022). The CACE analysis rests on two main assumptions:

- treatment and control groups have the same probability of non-compliance, which holds true given the randomisation procedure adopted in this evaluation; and
- being offered the intervention has no direct effect on outcomes unless the intervention is actually received and complied with (Raudenbush and Bloom, 2015). The validity of this assumption is argued theoretically in this evaluation.

As discussed above, compliance was a binary measure (where 1 = compliant; 0 = non-compliant) defined at the setting level based on attendance logs for professional development sessions. In line with EEF guidance, it was estimated for the primary outcome, EYT Numeracy, only.

The first stage of this 2SLS approach estimated the extent to which the assignment to the intervention affects setting to take up the treatment (the first stage regresses treatment assignment on compliance). This estimates a compliance rate and was estimated using the following equation:

$$(3) \quad Y_j = \beta_0 + \tau ONE_j + \beta_2 Z_j + u_j$$

where:

Y_j = compliance score of setting 'j';

β_0 = intercept;

ONE_j = binary indicator assigned to setting 'j' indicating if it is treatment [1] or control [0];

Z_j = setting-level characteristics of setting 'j' (region and setting type); and

u_j = setting level residual.

The second stage of the IV estimation predicts the outcome as a function of all covariates included in Equation 1 (see Analysis: Primary Outcomes) but substitutes the treatment indicator (ONE_j) in Equation 1) with the compliance rate estimated in the first regression (Angrist and Krueger, 1991; Angrist, 2006). Due to ease of estimation, we used an OLS IV approach to analysis, clustering the errors at the school level. This does not mimic exactly the multilevel hierarchical analysis employed throughout the rest of analysis, but still controls for intracluster correlations at the setting level to ensure appropriate standard errors and confidence intervals are used.

Dosage

While compliance indicated the impact of settings that are compliant with the intervention requirements (the training) it does not indicate how frequency of the intervention (that is, dosage) impacts outcomes. To this end we looked at dosage at the setting level as measured by the frequency of intervention activities provided to the children over a 12-week period. Since the intervention requires settings to deliver three activities per week during the intervention period, the dosage was set at 36 (three sessions over twelve weeks), creating a continuous dosage measure with a potential range from zero to 36.

However, as outlined in protocol, while dosage measures at the setting level might be an appropriate approximation for setting-level dosage, the varying attendance patterns of children in early years means this may not accurately represent

child-level dosage. This is because children may not attend days when the ONE was being delivered. In the protocol, we additionally suggested collecting electronic attendance data, where possible, but this added an extra burden on settings and so we did not proceed with this dosage measure. Additionally, asking settings to provide attendance of children at each activity was deemed too burdensome, therefore, we proposed measuring child-level dosage by the registered attendance hours of children at baseline as an estimate for child-level dosage.

The same analytical approach was used as compliance analysis above, with the outcome variable in the first stage the appropriate dosage measure rather than the compliance indicator. The model was estimated for primary outcome measure only (EYTN score). Using this approach, for setting-based dosage we estimated the average effect of attending a setting that offered one additional intervention activity on child-level numeracy outcomes. For child-level dosage, we estimated the average effect of an additional hour spent in an intervention setting on child-level numeracy outcomes.

This dosage analysis is not able to specifically capture the effect of additional hours of intervention on child outcomes meaning the estimated dosage effect will represent the average impact of attending a setting that is delivering the intervention, rather than the specific impact of attending specific intervention activities. However, given the difficulties in collecting accurate session attendance data, this is our closest approximation. This, along with the fact that the programme's logic model is concerned with children regularly exposed to broad, play-based activities across three areas (numbers and counting, ordering and patterns, shapes and spatial awareness), this limitation is well within bounds.

Compliance and dosage analysis was undertaken in R.

Missing data analysis

Missingness may occur due to attrition at setting or child level, child non-response to primary or secondary outcome testing (for example, refusing to participate in one test), or test administration errors. Unfortunately, non-random missingness can introduce bias into the ITT approach outlined in the analysis above. To understand better the impact of missingness on the analysis, we report the extent of—and sources of—missing data and whether there is a pattern in the missingness. For all primary, secondary, and subgroup analyses, we report the extent of missing data through cross-tabulations. For the primary outcome measure, we have analysed the pattern of missingness and perform a multiple imputation analysis, as recommended in the EEF evaluation guidance (EEF, 2022).

To assess whether there are systematic differences or a clear pattern in missingness, we modelled missingness at follow-up (defined as children with missing primary outcome data at endline) as a function of covariates available at baseline,¹³ with the exception of HTKS-R and Corsi Blocks at baseline due to the high level of missingness already prevalent. The analysis model for this approach mirrors the multilevel level model specified in Equation 1, with children clustered at the setting level. Given the binary nature of the missingness variable (where 1 = missing; 0 = complete), we used a multilevel mixed-effects logistic regression model (using the R statistical software). Given issues with baseline assessment, albeit concentrated largely in the secondary outcomes, we also modelled missingness at baseline using this same approach to understand patterns of missingness at baseline on the sample as randomised.

We followed the protocol for missing data suggested by the EEF guidance (EEF, 2022b). When missingness from the primary outcome model is less than 5% of the sample as randomised, we will conduct a complete-case analysis. This assumes that data is missing completely at random (MCAR), which we will be able to test only partially with the logistic model outlined above. If the missing data exceed 5% of the sample as randomised, our approach will depend on the pattern of missingness observed. If the missing data pattern appears to be unrelated to the effect of the treatment (for instance, solely due to child absences or test administration disruptions), we will presume that the data is Missing Completely At Random (MCAR) and proceed with an analysis based only on complete cases. We will repeat the logistic model of missingness at baseline as well to ensure all sources of missingness are analysed.

If we cannot assume data is MCAR, our approach will depend on the pattern of missingness revealed by the multilevel logit model outlined above. If there is evidence that missingness is correlated with observable covariates, then data is likely at least missing at random (MAR) and a complete-case analysis will be biased. Given the missingness at baseline

¹³ At the child-level, available baseline covariates are treatment group, gender, EYPP status, EAL status, EYTN baseline, HTKS-R baseline, and Corsi Blocks baseline. However, given degree of missingness on HTKS-R and Corsi Blocks at baseline, we could exclude these baseline measures from this logit model.

discussed above, in this evaluation we expect to employ a multiple imputation (MI) approach to address MAR. Both full-information maximum likelihood (FIML) and MI have been shown to be broadly equivalent (Lee and Shi, 2021). We will follow the guidelines for MI recommended in Jakobsen et al. (2017). Given children with valid primary or secondary outcome data were tested at endline, we likely face missingness in both endline and baseline EYTN, so our MI must allow for both types of missingness in primary outcome. Given this, we used a Multiple Imputation using Chained Equations (MICE) method for imputation to allow us to impute both missing baseline and endline.

MI only alleviates bias if the pattern of missingness is MAR; if the reason for missingness is due to unobserved variables, namely missing not at random (MNAR), then MI (or FIML) will not improve estimates. Note, however, that while the logistic model investigating pattern of missingness is informative, MAR and MNAR are not distinguishable based on observed data. If it seems likely data could be MNAR, sensitivity analysis will be conducted and reported alongside headline estimates.

Multiple imputation analysis was undertaken using the R statistical software. The advantage of using the `mice` package in R is that multiple imputation can be carried out on the multilevel models outlined in the primary outcome section (Grund et al., 2018).

Subgroup analyses

This trial is not powered to detect moderate effect sizes of treatment on children from disadvantaged backgrounds, however, a subgroup analysis was undertaken given the EEF's focus on this group. There are two indicators of disadvantage in the early years:

- children's eligibility for the Early Years Pupil Premium eligibility (EYPP), an additional top-up for eligible three- to four-year-olds; and
- Free Early Education Entitlement (FEEE) at the age of two.

During baseline child data collection, some settings reported not knowing FEEE eligibility as the children were not attending the setting at the age of two, and this information is not held administratively. Furthermore, the target population is three- and four-year-olds, making EYPP the more relevant measure of disadvantage—and therefore the one that we used.

We had initially collected EYPP data at baseline, however, it became clear that EYPP is not officially recorded until January. With this in mind, we asked schools in spring 2024 to provide us with data on which children were classed as EYPP.

The EYPP analysis was done in two stages, as recommended in the EEF evaluation guidance (EEF, 2022b). First, we completed the primary outcome analysis on the EYPP subsample, with effect sizes and statistical uncertainty calculated and reported as per the procedure outlined in Primary Outcome, using Equation 1. Second, we estimated a second model, using the full sample, with the additional inclusion of EYPP eligibility to estimate the effect of EYPP eligibility on the intervention effects. More formally, we added EYPP eligibility and its interaction with treatment assignment to the two-level random intercept model outlined in the Primary Outcome analysis section, given by Equation 4 below:

$$(4)Y_{ij} = \beta_0 + ONE_j\tau + EYPP_{ij}\beta_1 + (EYPP_{ij} * ONE_j)\beta_2 + Z_j\beta_3 + X_{ij}\beta_4 + u_j + e_{ij}$$

This is the same model specification as in Equation 1, with the addition of an $EYPP_{ij}$ indicator (taking on the value of 1 if a child is eligible for EYPP) and an interaction term combining EYPP eligibility and treatment allocation, $EYPP_{ij} * ONE_j$. We report both the interaction term coefficient, β_2 and its associated p-value and CI. If the β_2 coefficient is positive, it will indicate that the EYPP eligibility increases the treated children's endline score, as compared to their non-EYPP treated peers, indicating the intervention has a 'gap-closing' effect. However, if the β_2 coefficient is negative, it will indicate that the EYPP decreases treated children's endline scores, indicating the intervention has a 'gap-widening' effect.

For the first analysis undertaken on the subsample of EYPP children, the effect size was calculated and outlined in the same way outlined under Primary Outcomes Analysis. As a sensitivity check, we additionally calculated the effect size

of treatment for EYPP children using the interaction model outlined in Equation 2, run on the full sample. This was done according to the following formula:

$$(5) \quad ES = \frac{ONE_j \tau + (ONE_j * EYPP_{ij}) \beta_2}{sd}$$

The coefficients in the numerator come directly from Equation 4 and the standard deviation used in the denominator is the unconditional standard deviation of the EYPP subsample (both treatment and control).

Additional analyses and robustness checks

Given several differences between the proposed analyses and EEF guidance and issues encountered during baseline testing, we propose a series of additional analyses. All analyses outlined below were conducted using R, as outlined in the Primary Outcome and Secondary Outcome Analysis sections.

Accounting for age

In the primary outcome analysis, the coefficient τ is the estimated average treatment effect, with respect to the primary outcome measure (EYT Numeracy). The use of the raw scores for EYT follows EEF guidance (EEF, 2022), as age-standardised scores are not recommended by the developer. While using raw scores instead of age-standardised scores keeps the estimated average treatment effect unbiased, as long as there is balance in age across treatment arms at baseline, we know that both early numeracy and executive function are age-dependent (Howard et al., 2020; Howard et al., 2022; McClelland et al., 2021; Gonzales et al., 2021). We thus conducted, as a sensitivity analysis, an analysis including child's age as a variable.

This analysis repeats that described in the Primary Outcome analysis section (Equation 1 repeated here for ease), but now includes age in the child-level characteristics matrix X_{ij} :

$$(1) \quad Y_{ij} = \beta_0 + ONE_j \tau + Z_j \beta_1 + X_{ij} \beta_2 + u_j + e_{ij}$$

where:

Y_{ij} = EYT numeracy subscale score for child i in setting j ;

β_0 = cluster-level coefficient for the slope of a predictor on number skills;

ONE_j = binary indicator of the setting assignment to intervention [1] or control [0];

Z_j = setting-level characteristics, that is, the stratifying variables of geographical location and setting-type (as used for randomisation);

X_{ij} = characteristics of child i in setting j , that is, the pre-intervention EYT numeracy subscale score and age;

u_j = setting-level residuals; and

e_{ij} = individual-level residuals.

The coefficient τ represents the treatment's conditional effect on the primary outcome (EYT Score), with the treatment effect size calculated using Hedge's g , as outlined in the Effect Size Calculation section.

Studies suggest that EF is also age-dependent (McClelland et al., 2021; Gonzalez et al., 2021) so we repeated the above analysis on the secondary outcomes as well. As with the primary outcome analysis, we used an identical model, but with the addition of age to the child-level characteristics matrix, X_{ij} . We conducted this only for the mixed-measures model as we expected this to have the largest sample size and power.

Accounting for possible bias introduced by the randomisation process

As outlined under Randomisation, the switch to concealed randomisation meant that treatment status was revealed to two settings prior to the completion of all baseline assessments. To ensure reported results are robust to possible bias introduced by this revelation, we repeated the primary analysis excluding all children at these settings baselined after

allocation was revealed. If the sensitivity analysis had indicated that inclusion of these children significantly alters the estimated treatment effects, the headline primary and secondary outcome estimates would have been reported without these children included.

Accounting for possible measurement error

We are aware that test administrators faced some difficulties in administering the EF tests at baseline. This introduces measurement error as well as subjectivity as to what constitutes a ‘correct’ case versus an ‘incorrect’ case. At endline, the stopping rule was followed more consistently, and any difficulties were better reflected in the data, suggesting much more limited potential for measurement error. As part of the SAP development, we discussed with the delivery team and the EEF how data should be managed and agreed what we would include in the main analyses outlined above. However, these decisions could introduce potential bias and so we suggested the following sensitivity analyses to ensure that our approach to classifying measurement error did not unfairly bias results (Speciani et al., 2024). While this sensitivity analysis attempts to provide a more robust analysis, given the identified issues, we still recommend that all findings, particularly with regard to the secondary outcomes, are interpreted with caution.

For HTKS-R, we found that in some cases assessor reports directly contradicted the data. Assessors had the option of flagging cases where children did not complete the assessment. However, in over 300 cases (of 375), valid HTKS-R scores can be calculated and there appears to be no violation of stopping rules, despite the flag indicating the assessment was not completed. Given this inconsistency, for the main analysis all data that was flagged as incomplete, but nevertheless has valid, non-zero scores, was included in the analysis. As this data is a potential source of measurement bias, we ran a sensitivity analysis to exclude data that had been flagged this way. This analysis repeated the secondary outcome models involving HTKS-R (mixed-measures model and the HTKS-R model), excluding all children who had been flagged as not completing by assessors despite having valid, non-zero raw scores.

For Corsi Blocks, there were a higher than expected number of incomplete responses, with indications in the data that incomplete responses logged by assessors may have been incorrect responses (for example, where incomplete responses were then followed by correct responses, suggesting children were still willing to engage and participate). It was impossible to determine ex-post in a robust and non-biased way which responses were truly incomplete and which were more accurately incorrect and excluding responses could introduce non-random missingness into the data if these incomplete tests had been recorded as incorrect, as missingness will likely be higher at the lower end of the distribution. As a result, we conducted the mixed-measure model, outlined above under Secondary Outcome Analysis, scoring all incomplete responses at baseline as incorrect instead, and increasing the power of the analysis at the risk of introducing measurement error.

Finally, during the tail end of endline testing it became apparent that there was a routing error with the EYTN app where stopping rules were incorrectly applied in a small number of cases (N = 276; 16.34% of the endline primary analytical data sample). This error is apparent at the item level, making it easy to identify children who were affected by this. Given a large amount of data had been collected up to this point and the relatively small number of children effected, we retained the full dataset in the primary analysis. However, as an additional sensitivity check, we removed all cases where this error occurred—as it could introduce measurement error—and ran the analysis outlined in Equation 1. If this sensitivity analysis had suggested bias had been introduced into the primary outcome model, we would have reported headline results excluding these cases.

As discussed above, there are substantial floor effects in both HKTS-R and Corsi Blocks at baseline, and particularly with the former at endline. Floor effects were less prevalent in the EYTN dataset, making up approximately 2% at baseline (when the children are at the lower end of the validated distribution for EYTN) and less than half a percent at endline (which is in line with the presence of floor effects in the original validation sample, as presented in Howard et al., 2022). We cannot entirely overcome the impact of floor effects using the above sensitivity checks, so all secondary outcome findings should be interpreted with caution.

Analysis with endline data only

We are regrettably aware that the independent test administrators failed to upload baseline child data on the primary outcome for five settings, including all primary outcome child data for one setting and partial data for the remaining four. As a result, despite having 1,955 children baselined across one of the three assessments, we only have valid primary data available for 1,859. We still undertook endline testing for all children who were baselined in at least one of the three

assessment tasks. Given the unexpected reduction of sample size, we ran the primary outcome model outlined above in Equation 1 excluding the baseline EYTN variable.

Analysis excluding settings participating in Maths Champions

As noted in the Participant Selection section, a small number of settings were offered a similar early years intervention, Maths Champions, during the course of the evaluation of the ONE. These settings were asked to delay participation to at least April 2024, and ideally until June 2024. However, we acknowledge that there could be some risk to a similar intervention being delivered in assessments during the final months of endline assessment.

Given this risk, we repeated the primary outcome analysis excluding those settings: we ran the primary outcome model outlined above in Equation 1, excluding all settings that the EEF indicated participated in Maths Champions. If this sensitivity analysis had suggested bias had been introduced into the primary outcome model from the inclusion of Maths Champions settings, we would have reported headline results excluding these settings.

Mediation analysis

Our theory of change identifies EF as a key intermediary factor in the development of maths skills. The existing body of research, including findings by Blakey et al. (2020), robustly endorses EF's intermediary role in maths achievement, lending strong theoretical backing to our mediation model. To further understand the theory of change, we had proposed to conduct a mediation analysis to explore how executive functioning potentially influences the impact of our intervention on numeracy learning outcomes (see Speciani et al., 2024).

Meaningful mediation analysis can only be conducted by establishing first that there may exist a possible causal and statistically significant relationship between the intervention, mediation variable, and outcome variable. However, there was insufficient evidence that the intervention effected both early maths (the primary outcome) or EF (the secondary outcome), and as such we did not proceed with mediation analysis.

Estimation of effect sizes

As outlined in the Analysis section, unless otherwise stated, we calculated effect sizes (hereafter ES) for cluster-randomised trials using Hedges g as outlined in the EEF evaluator guidance (Education Endowment Foundation, 2022):

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{sd}$$

where:

$(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ = mean difference between the intervention and control group adjusted for baseline test score and other stratification variables; and

sd = estimate of the pooled unconditional standard deviation.

The pooled unconditional standard deviation is the weighted average of standard deviations of treatment and control (Coe, 2002). The pooled unconditional standard deviation across the two trial arms was used in the denominator, as we assume the standard deviations of both the treatment and control groups were drawn from the same underlying population distribution.

Effect sizes were computed for each of the estimated models and reported alongside their 95% confidence interval (CI). The effect sizes were then converted into months of progress (for attainment measures) to facilitate interpretability. All ES were estimated using R.¹⁴ If there was evidence of non-normality in residuals of any models, we had proposed to report non-parametric bootstrapped confidence intervals instead, as per the *eefanalytics* package.

¹⁴ See here for more information: <https://ideas.repec.org/c/boc/bocode/s458904.html>

Estimation of ICC

The ICC is a crucial metric for trials involving clusters. It quantifies the fraction of variance in a specific outcome attributable to differences between clusters (such as settings), rather than variance occurring within these clusters. To calculate the ICC on the primary outcome measure at analysis we followed two approaches: (i) we used the model corresponding to Equation 1 and (ii) employed a model akin to that in Equation 1 but without any covariates. This second model accounts for the clustering of children in schools and is referred to as the ‘empty model’.

ICCs were estimated using the R equivalent.

Longitudinal analysis

While longitudinal analysis for this study was not in scope, data-collection during evaluation does enable long-term follow-up using the National Pupil Database (NPD), despite the lack of Unique Pupil Numbers (UPNs). RAND collected the following identifiable information to allow subsequent matching of children with the NPD: first name, last name, date of birth, and setting postcode. These variables will be archived in the EEF's data archive. In addition, the delivery team proposed to approach parents via settings in 2024 to gather permission to collect further information on children while in reception. As part of this, they plan to collect information on the children's school, which will be archived by the delivery team with the EEF to facilitate long-term follow-up of these children. In addition, if UKRI permits, data on children's numeracy and executive function skills in reception will also be archived by the delivery team with the EEF. All data protection documentation, from privacy notices and project information sheets, to DSAs, clearly state that data collected will be linked to school-level information and follow-up analysis conducted.

Implementation and process evaluation

Implementation is the process by which an intervention is put into practice and can be considered a multi-dimensional construct consisting of compliance, fidelity, participant responsiveness, programme differentiation, monitoring of control or comparison conditions, and adaptation (EEF, 2022b).

Research methods

We developed a mixed-methods implementation and process evaluation (IPE) data collection plan, which included training observations, semi-structured interviews, and surveys, as outlined in Table 6. The tools and approaches are based on the theory of change, which was co-constructed with RAND, the delivery team, and the EEF. This ensures that the approach is theory-based and intervention-led. Finally, we balanced the need to triangulate data (that is, to increase reliability of findings by asking a number of sources) with the need to reduce burden on school staff.

Table 6. IPE methods overview

Research methods	Data collection methods	Participants/data sources	Data analysis methods	IPE research questions addressed	Implementation/ logic model relevance
Training observations	In-person training session observations by RAND	Observations of each of the four training sessions, at four different treatment settings, one in each of four different regions involved in the trial.	Thematic analysis	1, 2	Fidelity and adaptation, dosage.
Interviews	Semi-structured interviews	Trainers (4) delivering training sessions in all four regions involved in the trial.	Thematic analysis	1, 3	Fidelity and adaptation, dosage, programme differentiation.

	Semi-structured interviews	Managers (12) and practitioners (10) in 12 treatment settings selected through random sampling.	Thematic analysis	1, 2, 3, 4, 5, 6, 8	Fidelity and adaptation, dosage, programme differentiation, unintended consequences, context/moderators, cost.
Surveys	Online questionnaires	Managers and practitioners in control settings (75).	Thematic analysis; descriptive statistics	4, 6, Cost RQ	Programme differentiation, context/moderators, cost.
	Online questionnaires	Managers and practitioners in treatment settings (75).	Thematic analysis; descriptive statistics	1, 2, 3, 4, 5, 6, 7, Cost RQ	Fidelity and adaptation, dosage, programme differentiation, unintended consequences, context/moderators, mediators, cost.

Training observations

Four in-person observations of the ONE training were carried out by RAND to understand the way in which training imparts the core components of the intervention and how these are cascaded from Oxford trainers to setting practitioners and, where relevant, managers. These observations were conducted in each of the four regions of England participating in the programme and included one of each of the four training sessions.

This spread of observations helped to shed light on the extent to which training imparts core elements of programme to those responsible for delivery. This developed understanding of the extent to which the in-person training worked (or did not work) in practice. RAND researchers made notes on key features of the training, including content, structure, and timing. Notes were made during training sessions. These notes were analysed using a core components framework that allowed researchers to map key details of training against the core components.

Interviews

RAND conducted two rounds of interviews as part of the IPE. First, semi-structured interviews were conducted with all four trainers of the ONE after all in-person training had been delivered. These interviews were designed to gather trainers' reflections on whether and how the ONE training worked in different settings, practitioners' responses to the training and activities, and their own experiences of delivery of the ONE training. The interview schedule for trainers can be found in Appendix F.

Second, semi-structured interviews were conducted with practitioners and managers at twelve settings that had received the ONE intervention. These interviews were designed to gain managers' and practitioners' perspectives of the impact of the ONE in their setting, the extent to which the ONE activities needed adaptation in their setting, and their thoughts on the training and support they received during the intervention. The interview schedule for practitioners can be found in Appendix G and the interview schedule for managers in Appendix H.

All interviews were conducted after the ONE had been implemented in settings and were held online via Microsoft Teams at a time agreed by interviewees in order to minimise any potential burden.

Surveys

Our IPE activities included two rounds of surveys. First, a baseline survey exploring settings' usual practice with regard to maths and executive functioning was sent to all setting managers shortly after randomisation using an online survey tool. The baseline survey can be found in Appendix I.

An endline survey including topics such as potential unintended consequences of implementing the ONE activities, delivery, fidelity, and adaptations was sent to intervention setting managers and practitioners. Managers were also asked about the cost implications of running the programme. Control setting managers and practitioners were also surveyed at endline to describe business as usual in the control groups. Survey questions explored the extent to which control schools have implemented targeted maths or executive function interventions, including their cost. The endline surveys for practitioners can be found in Appendix J (control group) and Appendix K (treatment group) and the endline surveys for managers can be found in Appendix L (control group) and Appendix M (treatment group).

Analysis

Qualitative data from training observations and interview data were analysed using a thematic analysis framework informed by the IPE research questions (see Evaluation Objectives). These were coded independently by two researchers and discussed to promote reliability of findings. Deductive coding was used primarily, cross referencing against the theory of change, though researchers were encouraged to use inductive coding in instances where responses promoted alternative interpretation.

Quantitative data from surveys was cleaned and then analysed using descriptive statistics. Findings from the relevant sections of the survey were included alongside qualitative data in the thematic analysis framework to provide a fuller picture of each research question through triangulation. Findings from the thematic analysis were then compared against the theory of change, with any similarities and divergences noted.

Finally, IPE findings were discussed alongside the impact evaluation at an evaluation team synthesis workshop in November 2024. This workshop focused on integrating findings from the impact and IP evaluations to understand what may have led to the observed results.

Costs

Data on implementation costs was collected through endline surveys with practitioners in both control and treatment settings and through a brief questionnaire with the delivery team. The cost evaluation considered both direct and indirect costs incurred for settings that delivered the ONE and those incurred in control settings implementing similar programmes. These direct and indirect costs include but are not limited to: (a) time away from teaching due to participation in training and other programme activities, (b) staff cover for teaching staff participating in out-of-setting programme-related activities, (c) prices of instructional materials, and (d) additional staff workload required to run the programme.

Analysis was conducted in line with the EEF's 2023 Cost Evaluation Guidance (EEF 2023).

Timeline

Table 7: Timeline

Dates	Activity	Staff responsible/leading
Oct–Dec 2022	Set-up meetings	RAND, University of Oxford, University of Sheffield, EEF
Dec 2022–Apr 2023	Development and finalisation of data protection documents (DPIA, LIA, DSA) Ethical approval	RAND, University of Oxford
Jan–Apr 2023	Development and finalisation of recruitment documents (MOU, information sheets, privacy notices, withdrawal forms)	RAND, University of Oxford
Apr–Jul 2023	Recruiting settings	University of Oxford, University of Sheffield

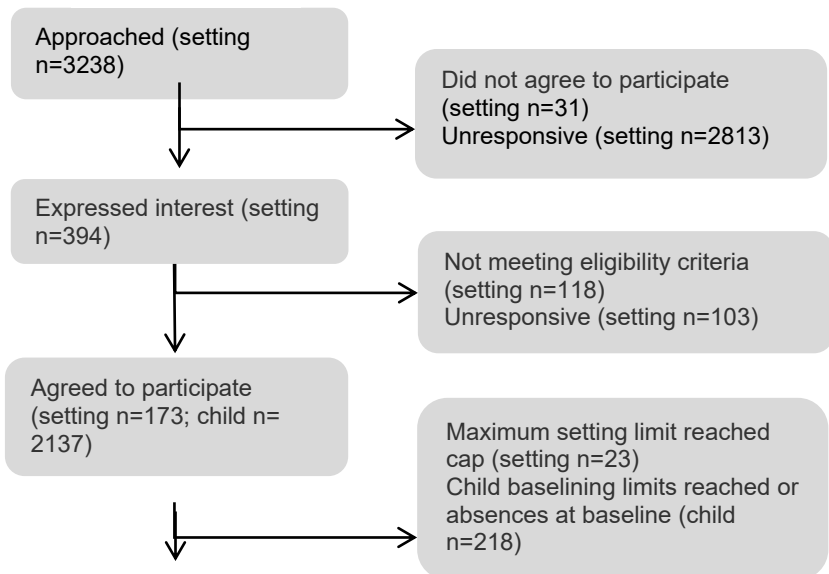
Mar–Oct 2023	Development and publication of the trial protocol	RAND, EEF
Oct–Dec 2023	Baseline testing	RAND, Qa Research
Dec 2023	Randomisation of settings	RAND
Nov 2023	IPE data collection: baseline survey	RAND
Jan–May 2024	Delivery of the ONE programme in settings	University of Oxford, University of Sheffield
Jan–May 2024	Training observations	RAND
Apr–Jul 2024	Endline testing	Qa Research, RAND
May–Jul 2024	Interviews and endline survey	RAND
Jul–Sep 2024	IPE and impact evaluation analysis	RAND
Aug–Jan 2024	Report writing	RAND
Sep 2024	The ONE delivery begins in waitlisted control settings	University of Oxford, University of Sheffield

Impact evaluation results

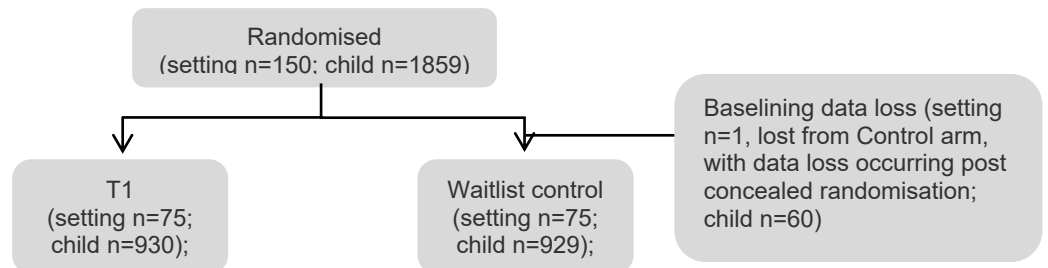
Participant flow including losses and exclusions

Figure 2: Participant flow diagram (two arms)

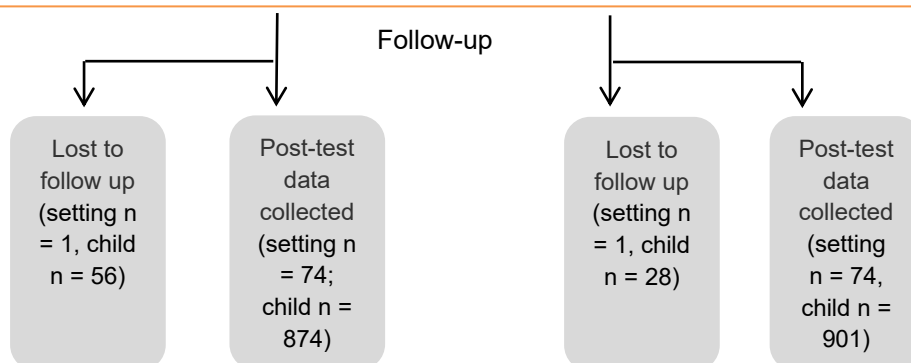
Recruitment



Allocation



Follow-up



Analysis

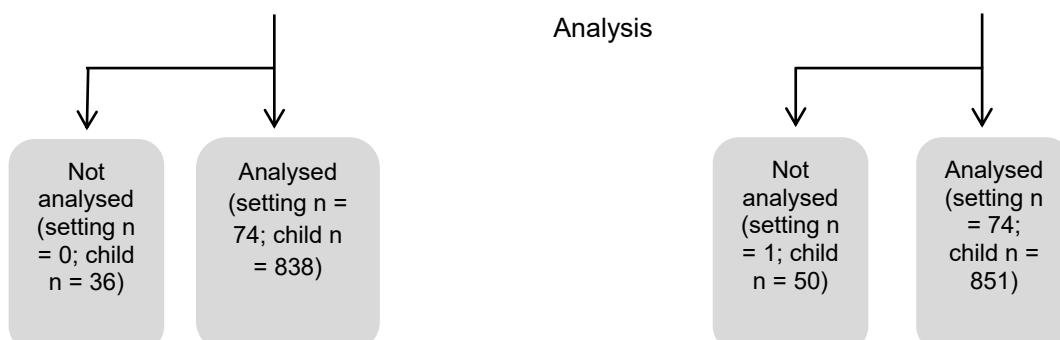


Table 8: Minimum detectable effect size at different stages

		Protocol		Randomisation		Analysis	
		Overall	EYPP	Overall	EYPP	Overall	EYPP
MDES		0.204	0.250	0.204	0.280	0.196	0.235–0.242
Pre-test/post-test correlations	Level 1 (child)	0.8	0.8	0.8	0.8	0.68	0–0.68
	Level 2 (class)	NA	NA	NA	NA	NA	NA
	Level 3 (setting)	0.2	0.2	0.2	0.2	0.5	NA
Intracluster correlations (ICCs)	Level 2 (class)	NA	NA	NA	NA	NA	NA
	Level 3 (setting)	0.18	0.18	0.18	0.18	0.188	0.06
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		12.5	2.4	12.4	1.5	11	2.7
Number of settings	Intervention	75	75	75	75	74	74
	Control	75	75	75	75	75	75
	Total:	150	150	150	150	149	149 *
Number of children	Intervention	1,125	180	974	111	838	130
	Control	1,125	180	981	114	851	137
	Total:	2,250	360	1,955	225	1,689	267

* Estimates are based on the 98 settings in which there were valid test statistics for EYPP children.

A full picture of attrition is presented in the Participant Flow Diagram (Figure 2). Child loss from baselining data loss is an estimate only, based on test administrators reports. Due to the presence of wi-fi-enabled devices, some data loss was mitigated before tablets were wiped. As presented in Table 8, the MDES at analysis was found to be 0.196 at the overall sample level. This marks a mild improvement to the MDES calculated at randomisation stage (0.204). This appears to be driven largely by a higher setting-level pre-post test correlation at analysis stage (0.5) compared to at randomisation (0.2).

Attrition

As can be seen in Table 9, attrition at the child level was relatively low with only 9.9% missing from intervention and 8.4% missing from control for a total of 9.1% overall; 4% of children were lost to follow-up as they had moved settings and 5.9% were unable to be tested at endline (due to illness or testing dates not coinciding with child attendance patterns). One treatment setting with eight children (less than 1% of the treatment sample) also declined to take part in endline testing.

Table 9: Child-level attrition from the trial (primary outcome)

		Intervention	Control	Total
Number of children	Randomised	930	929	1,859
	Analysed	838	851	1,689
Child attrition (from randomisation to analysis)	Number	92	78	170
	Percentage	9.9%	8.4%	9.1%

Child and setting characteristics

Table 10 presents child and setting characteristics at randomisation and at endline. There were slightly more PVI settings overall compared to SBS, but these were equally distributed across treatment and control. A similar balance across treatment and control was achieved for region. These proportions were maintained at analysis stage—likely due to the relatively low levels of attrition at the setting level (see Attrition section above).

At the child level there was balance on gender and eligibility for EYPP, which was also maintained at analysis stage.

Table 10: Baseline characteristics of groups as randomised

Setting-level (categorical)	At randomisation				As analysed			
	Intervention group		Control group		Intervention group		Control group	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Setting type								
SBS	34/75 (0)	45.33%	33/75 (1)	44%	33/74 (1)	44.59%	33/74 (0)	44.59%
PVI	41/75 (0)	54.67%	41/75 (1)	56%	41/74 (0)	55.41%	41/74 (0)	55.41%
Region								
East Midlands	21/75 (0)	28%	22/75 (0)	29.33%	21/74 (0)	28.38%	21/74 (0)	28.28%
East of England	21/75 (0)	28%	20/75 (0)	26.67%	21/74 (0)	28.28%	20/74 (0)	27.03%
London	22/75 (0)	29.33%	22/75 (0)	29.33%	22/74 (0)	29.73%	22/74 (0)	29.73%
Yorkshire and Humber	11/75 (0)	14.67%	11/75 (0)	14.67%	10/74 (1)	13.51%	11/74 (0)	14.86%

Evaluation Results									
Setting-level (continuous)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	
Proportion of female children	74/75 (1)	0.51 (0.16)	75/75 (0)	0.50 (0.13)	74/74 (0)	0.52 (0.16)	74/74 (0)	0.50 (0.14)	
Child-level (categorical)	n/N (missing)	Count (%)	n/N (missing)	Count (%)	n/N (missing)	%	n/N (missing)	%	
EYPP									
Not EYPP-eligible	821/974 (1)	84.38%	821/983 (3)	83.78%	708/838 (0)	84.49%	714/851 (0)	83.90%	
EYPP-eligible	152/974 (1)	15.62%	159/983 (3)	16.22%	130/838 (0)	15.51%	137/851 (0)	16.10%	
Gender									
Female	450/974 (93)	51.08%	459/983 (73)	50.44%	427/838 (2)	51.08%	429/851 (4)	50.65%	
Male	431/974 (93)	48.92%	451/983 (73)	49.56%	409/838 (2)	48.92%	418/851 (4)	49.35%	
Child-level (continuous)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect Size
EYTN	930/973 (43)	23.58 (14.66)	929/981 (52)	23.66 (15.07)	838/973 (135)	23.53 (14.80)	849/981 (130)	23.87 (14.84)	-0.09
HTKS	952/973 (21)	38.95 (34.06)	933/981 (48)	35.56 (33.17)	842/973 (131)	39.64 (34.1)	843/981 (138)	36.39 (32.95)	0.56
Corsi Block	921/973 (52)	4.45 (3.05)	887/981 (94)	4.28 (2.99)	825/973 (148)	4.42 (3.04)	818/981 (163)	4.26 (2.95)	0.09

Outcomes and analysis

Primary analysis

As presented in the logic model, the ONE programme primarily seeks to improve early mathematics attainment among children exposed to the intervention through structured play sessions focusing on a range of key early mathematics concepts. As outlined in the Methods sections, this evaluation operationalised early mathematics using the Early Years Toolbox (EYT) Early Numeracy Assessment (ENA). At endline, ENA scores had a mean of 32.69 and a standard deviation of 15.64. The distribution of ENA endline scores is presented in Figure 5 (Appendix D) and points towards a broadly normal distribution.

The impact of the ONE intervention on early mathematics was explored using a linear random intercepts regression model, which accounts for the multilevel structure of the data where children are nested within settings. The EYT ENA measure is used as the endline measurement of the primary outcome. Controls are added for region, setting type, and baseline assessment scores. A complete case analysis was used, generating an estimate of the treatment effect on an ITT basis for observations which contain no missing data for outcomes or controls. The residuals from this primary analysis model were not normally distributed (see Appendix E) and, therefore, the confidence intervals and p-value associated with the treatment effect were re-estimated using a bootstrapping technique to correct for this violation of the underlying normal residual assumption.

The results of the primary analysis (Table 11) suggest that receipt of the intervention did not result in a measurable impact on early mathematics scores among the intervention group. The effect size (0.007) is very close to zero and when combined with wide associated bootstrapped confidence intervals (-0.12; 0.13), which span zero, prevents us from rejecting the null hypothesis. While there is some degree of missingness in both the intervention and control

groups—36 and 50 respectively—our results hold during sensitivity and robustness testing, which strengthens the findings of the primary analysis.

We considered the possibility of floor effects in the primary analysis, as the baseline distribution displayed in Figure 4 appear to display a positive skew. However, only 2% of baseline observations scored at the EYTN demonstrate floor effects making it unlikely that floor effects are problematic in this case. This is higher than the overall presence of floor effects in Howard et al. (2022); however, the baseline sample is at the lower end of the validated age range for EYTN, with an average age of 3.64 years, so higher floor effects are to be expected. At endline, the overall floor effects are marginally lower than those found in Howard et al. (2022), which is likely due to the aging of the cohort over the six-month interval between baseline and endline. We perform an endline-only analysis which does not factor in children's EYTN scores at baseline to provide a robustness check to the potential floor effects, or any inadvertent introduction of measurement bias through baseline EYTN. The results in Table 20 suggest that there is not a measurable difference between the endline scores of the intervention and control group.

Table 11: Primary analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
EYTN Numeracy Test	838 (36)	32.58 (31.54; 33.62)	851 (50)	32.8 (31.72; 33.87)	1689 (838; 851)	0.007 (-0.12; 0.13)	0.92

Secondary analysis

A key element of the ONE's proposed mechanism of action for improving children's mathematics attainment is through improving EF, particularly in relation to inhibition, working memory, and cognitive flexibility. Two measures were selected to measure different aspects of EF: HTKS-R, measuring EF more broadly, and Corsi Block scores specifically measuring visual-spatial ability. These two measures were initially intended to be utilised as separate secondary outcomes. However, as previously outlined, floor effects encountered in the HTKS-R baseline scores posed a threat to robustly estimating a treatment effect, as floor effects may suggest that the measure did not capture the variation in EF at the bottom of the scale. To understand the impact of these floor effects on the secondary outcome impact estimates, we conducted the HTKS-R analysis using Corsi Block scores at baseline and HTKS-R scores at endline. This was agreed between all parties as the latent constructs operationalised by these two measures are sufficiently proximal to each other.

Additional models were subsequently run on HTKS-R and Corsi Block scores specifically. The results of the secondary analyses are presented in Table 12, where confidence intervals and p-values are estimated using bootstrapping techniques as a result of non-normal residuals encountered when initially running the analysis.

We do not find evidence that receipt of the ONE programme had an impact on EF, as analysed by the mixed-model that utilises HTKS-R scores as the endline outcome measure and Corsi Blocks scores as the baseline outcome measure. The mixed measures model produced a negligible effect size (-0.03), with this estimate surrounded by wide bootstrapped confidence intervals and a high p-value, preventing us from rejecting the null hypothesis in this instance.

In addition, we do not find evidence of an impact of the ONE intervention on EF in the analysis that uses HKTS-R or Coris Blocks as both baseline and endline. Measure-specific models similarly found null effects, with again negligible effect sizes found for both the HTKS-R (-0.052) and Corsi Block (0.044) estimates, accompanied by wide bootstrapped confidence intervals. There is indication of potential floor effects in the secondary outcomes at baseline as indicated by histograms (see Figure 8 and Figure 10, Appendix D). In the HTKS-R measure, 8.1% of baseline scores are at the floor and, in the Corsi Blocks measure, 10.6%. These figures imply that the interpretation of these results could be severely attenuated by floor effects at baseline, which could bias our results if the secondary measures do not capture the

progression of lower-ability children. We were aware of the possibility of floor and ceiling effects when planning the analysis stage and therefore conducted a mixed-measure model (with Corsi Block baseline and HTKS-R endline) as a robustness analysis. However, as both measures have potential floor effects at baseline, the results of this model do not resolve this issue. Other robustness checks designed to deal with floor and ceiling effects, such as the Tobit model, are only appropriate where they exist in the dependent variable (in this case the endline results) and therefore are unfortunately not useful in this analysis.

The null results observed across the secondary analysis are similar to the results obtained from the primary analysis. A key mechanism, as explained in the theory of change and logic model (Figure 1), is that improvements to early years numeracy skills are mediated through initial improvements to EF; in particular, the ONE programme targets inhibition, working memory, and cognitive flexibility. Our secondary analysis findings indicate that there were no improvements to EF in the children who received the ONE intervention (compared with the control group). The lack of an effect on EF in the secondary analysis may explain why there were no improvements identified in the primary outcome, numeracy. However, given the complexity of missing data, administrative errors experienced during baseline data collection, and the presence of substantial floor effects, these results should be interpreted with caution. We explore the impact of missing data in later sections.

Table 12: Secondary analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (Boot. 95% CI)	Boot. p- value
HTKS Executive Function Test (mixed measures model)	818 (41)	51.85 (49.72; 53.97)	811 (70)	53.22 (51.05; 55.38)	1629 (818; 811)	-0.03 (-0.17; 0.1)	0.65
HTKS Executive Function Test (single measure model)	842 (17)	51.66 (49.57; 53.75)	843 (38)	52.55 (50.41; 54.69)	1685 (842; 843)	-0.052 (-0.19; -0.08)	0.43
Corsi Blocks Executive Function Test	825 (42)	5.6 (5.42; 5.79)	818 (73)	5.46 (5.28; 5.64)	1643 (825; 818)	0.044 (-0.1; -0.18)	0.55

Analysis in the presence of non-compliance

The primary analysis presented above assessed the impact of the ONE intervention on an ITT basis, analysing the sample as randomised. However, analysis on an ITT basis can sometimes include outcomes for children who were assigned to the treatment group but did not eventually receive the intervention (or visa versa for the control group). Therefore, further analysis was conducted to explore the effect of receipt of the intervention (compliance) on the primary outcome. Participation in the ONE intervention requires settings to both participate fully in a series of professional development sessions over the course of the 12-week intervention and implement the intervention activities three times a week over the intervention period. Given the competing pressures facing settings, mandating that settings satisfy the full intervention requirements would likely be too strict a definition of compliance. As discussed in the Methods section, three measures of compliance were explored in this analysis: binary setting-level compliance, continuous setting-level dosage, and continuous child-level dosage. For more detail on these three measures, as well as justifications for their use, please consult Table 13.

Table 13: Compliance and dosage criterion

Compliance and dosage criterion	Data source	Compliance or dosage indicator
Setting-level compliance: attendance at professional development sessions	Attendance recorded by delivery partner	At least one staff member from the setting has attended all sessions
Setting-level dosage: intervention activities offered to the children	Intervention activity delivery recorded by delivery partner	Number of times intervention activities were offered in the setting (ranges from 0 to 36)
Child-level dosage: attendance patterns at setting	Attendance patterns reported by settings to evaluation team	Number of hours a week the child usually attends a setting

Typically, in evaluation of programmes in education settings we see different levels of compliance, however, here, all but one treatment setting met the minimum requirement for staff attendance at professional development sessions, ensuring that at least one staff member from the setting attended all sessions. However, only 17 of the 75 treatment settings were able to offer all 36 activities included in the intervention; 31 of the 75 were able to offer 30 or more, and the average was 28.67 across the intervention period. We next conduct CACE analysis utilising each of the dosage measures contained in Table 13.

The results of the CACE analysis (Table 14) do not provide evidence that intervention receipt had an impact on early numeracy. The CACE model using the binary measure of setting-level compliance found a marginally higher level of early numeracy among children attending compliant settings (Hedges' $g = 0.04$), but this effect was not statistically significant, as evidenced by the large p -value—0.58. No association was found between intervention dosage at either the child or setting level and early numeracy, with both analyses yielding an effect size of 0.00.

However, given the lack of variation in compliance by the staff attendance measure, less emphasis should be put on this result. CACE analysis is more appropriate for situations where there are larger numbers of non-compliers. Without sufficient variation in a compliance indicator, the first stage may produce unreliable standard errors or inadvertently introduce bias. Given there was only one non-compliant setting, as a robustness check, we conducted an analysis where this setting was simply dropped from the sample and we repeated the primary analysis, which shows a similarly small effect size of 0.02. This measure separates out exposure to training (a pure compliance measure) from delivery of activities, which combines compliance with dosage. However, the separation between setting-level compliance and dosage is somewhat artificial, with the undertaking of activities by staff both impacting whether or not settings comply with the intervention as intended and the degree to which children are exposed to the intervention. Given the lack of variation in the training compliance measure, the analysis of setting-level dosage may provide further context for understanding the implementation and effect of the ONE intervention in this trial.

Analysis of setting-level dosage similarly finds no additional benefit to the ONE intervention based on exposure to the maths and EF activities. There is greater variation in number of activities delivered, resulting in a less-skewed dependent variable; however, concerns over the strength of the first stage remain (see Table 9 in the Appendix). Even with the

improved distribution compared to the highly-skewed binary compliance, there remains a skew in this data, with just over 20% of the settings delivering the required 36 activities. For this reason, the linear 2SLS CACE analysis should still be interpreted with caution.

Table 14: CACE analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome EYTN Numeracy test	N (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (boot. 95% CI)	Boot. p- value
Setting-level compliance	838 (36)	32.58 (31.54; 33.62)	851 (50)	32.8 (31.72; 33.87)	1689 (838; 851)	0.04 (-0.09; 0.16)	0.58
Setting-level dosage	838 (36)	32.58 (31.54; 33.62)	851 (50)	32.8 (31.72; 33.87)	1689 (838; 851)	0.00 (-0.0042; 0.0036)	0.86
Child-level dosage	838 (36)	32.58 (31.54; 33.62)	851 (50)	32.8 (31.72; 33.87)	1689 (838; 851)	0.00 (-0.0055; 0.0042)	0.79
Non-compliant setting excluded	830 (36)	32.58 (31.54; 33.63)	851 (50)	32.8 (31.72; 33.87)	1681 (830; 851)	0.022 (-0.1; 0.14)	0.74

Missing data analysis

Further analysis was conducted to explore potential patterns to missingness as this exceeds the 5% threshold set by the EEF evaluation guidance (EEF, 2022). The approach taken by the evaluation team involved three steps: conducting a descriptive assessment of the balance of baseline characteristics, exploring the possibility of data being MAR using logistic regression models, and finally re-estimating the primary analysis model using multiple imputation (MI).

Initially we conduct a descriptive assessment of the balance of characteristics for children who undertook baseline EYT-N assessments but did not undertake endline EYT-N assessments. As previously noted, there were 169 children who fell into this category. Of these, four were missing some information on baseline characteristics, for example, only 165 or the 169 have information on their setting region. Therefore, we explore the balance of characteristics of 165 children who did not take an endline assessment but have complete characteristic information; we call this the 'missing endline sample'.

Table 15 details the number and percentage of observations in the missing endline sample and the as-analysed sample which fall into the categories of characteristics which are included as covariates in different parts of our analysis. We see that around 46% of the missing endline sample fall into the treatment group and a similar approximately 50% of the as-analysed sample fall into the treatment group. There are similar percentages of children with EYPP status in the missing endline sample (~18%) and as-analysed sample (~16%) as well as for PVI setting types (~52% and ~53% respectively). The means of both 'age in months' (both ~44 months) and baseline EYT-N scores (~22% and ~23% respectively) also appear to be similar in both the missing endline sample and the as-analysed sample.

However, we do see some differences in the balance of gender, which is ~58% male in the missing endline sample and ~49% male in the as analysed sample; this may indicate that gender is predictive of missing endline outcomes. In addition, we see some differences in the percentage of children in each region within the sample. While a similar percentage of children are from London in the missing endline sample (~28%), the missing endline sample has smaller percentages of children than the as-analysed sample in the East Midlands (~24% and ~31%) and East of England (~24% and ~28%). In contrast, the missing endline sample has a larger percentage of missing children in the Yorkshire and Humber region (~24% compared to ~14% in the as-analysed sample). This may also indicate that children who completed the baseline EYT-N assessments but did not complete endline EYT-N assessments are concentrated in the Yorkshire and Humber region.

Table 15: Balance of covariates in missing endline sample and as-analysed sample

	Missing endline sample		As-analysed sample	
	Number of observations	Percentage	Number of observations	Percentage
Treatment				
Treatment	77	46.67%	838	49.62%
Control	88	53.33%	851	50.38%
Gender				
Male	95	57.78%	824	49.14%
Female	70	42.42%	853	50.86%
EYPP status				
EYPP	29	17.58%	267	15.81%
No-EYPP	136	82.42%	1,422	84.19%
Setting type				
PVI	86	52.12%	895	52.99%
Maintained	79	47.88%	794	47.01%
Setting Region				
East Midlands	39	23.64%	516	30.55%
East of England	39	23.64%	471	27.89%
London	47	28.48%	459	27.18%
Yorkshire and Humber	40	24.24%	243	14.39%
	Missing endline sample		As-analysed sample	
	Mean	SD	Mean	SD
Age in months	43.98	3.64	43.67	3.57
Baseline EYT-N Score	22.65	15.00	23.70	14.82

Further exploration of missingness

While the descriptive analysis may indicate that missingness is concentrated among male children from the Yorkshire and the Humber, the results above do not provide causality. Therefore, we go further to conduct an analysis of missingness, exploring the possibility of data being MAR using a logistic regression model. As per the Methods section, this involved creating a dummy variable denoting missingness in the primary outcome at endline and modelling this dummy variable on baseline covariates included in Table 15 to explore whether any were significant predictors of missing endline data. The complete results of this logistic regression model are presented in Appendix Table 8 (in Appendix C).

This analysis provides some evidence that primary outcome endline data is MAR. In line with the descriptive overview presented above, the logistic regression model found an association between setting region and likelihood of missingness at endline, with children attending settings in Yorkshire and the Humber having a statistically significant higher likelihood of having primary outcome data missing at endline ($p = 0.04$). Furthermore, male children appear to

have had a higher likelihood of having missing endline data, when controlling for all other covariates. There was less statistical certainty with this finding, however, as evidenced by the higher p-value ($p = 0.08$). While this analysis suggests the presence of missingness contingent on observable characteristics, and therefore the data being MAR, the pseudo R^2 of 0.03 for fixed effects in the model is indicative of a low level of predictive power, and therefore a high incidence of missingness at endline that cannot be predicted by the baseline characteristics included in the model. In this way, a substantial portion of missing endline data cannot be deemed MAR.

As detailed in the Methods section, a small proportion of the sample had missing EYT-N baseline data, making up 5.0% ($N = 97$) of the sample as randomised. While this included 60 children who were baselined but whose data was not initially uploaded due to technical issues, it also includes a smaller number of children tested at endline who were not tested at baseline. Despite the potential inclusion of children who were not in the original randomised sample, therefore, these observations were included in the missing baseline analysis and subsequent MI results presented below to maintain the ITT approach used in the preceding analysis. To explore the possibility that baseline data was MAR, an additional logit model was run. The approach taken was the same as for the missing endline analysis, with a binary missing baseline flag modelled on baseline covariates to explore the possibility that this data was MAR. The results of this analysis are displayed in Appendix Table 9 and do not provide any evidence of MAR, as evidenced by the lack of statistical associations between any included covariates and the likelihood of EYT-N data being missing at baseline. As with the missing endline analysis, however, the low pseudo R^2 for the model's fixed effects (0.06) suggests that the included covariates have very limited predictive power, indicating that only a very small portion of the missingness in baseline can be explained by observable characteristics captured by the data.

Re-estimation of the primary analysis using multiple imputation

Given that the above analysis suggests that data cannot be assumed to be MCAR, a complete case analysis is biased. As per the Methods section, therefore, we re-estimated the effect of the ONE intervention on early numeracy using multiple imputation, allowing for us to account for possible MAR mechanisms in both baseline and endline EYT-N data. Imputation of missing baseline and endline data was conducted using chained equations using both setting- and child-level baseline characteristics, taking into account the hierarchical structure of the data. Imputations of baseline and endline scores were only possible where observations had complete cases for the covariates used to predict their values. Due to small amounts of missingness in EYPP status and gender, ten observations were excluded when re-estimating the primary effect size. Furthermore, observations where the child had both missing baseline and endline data were also excluded ($n = 11$). Twenty individual imputations were performed, with results pooled across the imputed datasets produced. The results of the MI analysis are provided in Table 16: the re-estimated pooled Hedges' g are aligned with the primary analysis results, providing no additional evidence that the intervention has had any impact on early numeracy, as indicated by the effect size of zero, wide associated confidence intervals, and extremely high p-value (0.99). It should be noted, however, that MI can only reduce biases in a complete case analysis where data is MAR; given the limited degree of both baseline and endline MAR data, this MI analysis likely only marginally improves the precision of the primary effect size estimate.

Table 16: Missing data adjusted primary outcome analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled Mean (95% CI)	n (missing)	Pooled Mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
EYTN Numeracy Test	961	32.53 (32.31; 32.74)	972	32.62 (32.40; 32.85)	1,933 (961; 972)	0.00 (-0.07; 0.07)	0.99

Subgroup analyses

We conducted subgroup analyses to explore the impact of the ONE intervention on the early numeracy of EYPP-eligible children. Initially we conducted the primary analysis, but this is restricted only to the sub-sample of children who were eligible for EYPP; this allows us to explore whether the intervention was effective within this EYPP subgroup. Next, we

investigated whether there was any additional impact of the intervention on such children by introducing an interaction term between treatment assignment and EYPP eligibility.

The results of the sub-sample analysis are presented in Table 16. While not statistically significant, the primary analysis model, when restricted to those EYPP-eligible, produces an effect size of 0.14. This suggests that, on average, children in the EYPP subgroup who received the ONE intervention (that is, were assigned to the treatment group) made approximately two months additional progress in early numeracy compared to those in the control group. However, we would caution over-interpretation of these results as the wide bootstrapped confidence intervals (-0.09; 0.37) and medium-sized p-value (0.22) do not allow us to conclude that the intervention improved early numeracy within the EYPP subgroup. However, we note that this evaluation was not powered to detect a moderate effect size in the EYPP sub-sample.

Furthermore, the results of the interaction model, displayed in Appendix Tables 11 to 13, do not suggest that there was an additional impact of the ONE intervention on children who do not have EYPP eligibility, as indicated by the marginal effect size (0.04) and wide associated bootstrapped confidence intervals which surround zero (-0.10; 0.18). Appendix Table 13 shows that the additional impact on EYPP eligible children is also insignificant at 0.11, with a wide confidence interval (-0.086; 0.30), which is broadly similar to the estimated effect size from the sub-sample model.

Table 17: EYPP sub-sample model

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled mean (95% CI)	n (missing)	Pooled mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p- value
EYTN Numeracy Test	130 (6)	26.58 (24.29; 28.88)	137 (7)	25.34 (23.27; 27.4)	267 (130; 137)	0.14 (-0.09; 0.37)	0.22

Additional analyses and robustness checks

Age adjusted analysis

In the primary analysis, raw-scores are used for the EYT-N assessment outcome, following EEF guidelines; this enables an unbiased estimate in the case where there is balance in age between the treatment and control groups. However, evidence suggests that both early numeracy and executive function are age dependent (Howard et al., 2020; Gonzales et al., 2021; Howard et al., 2022). Therefore, we conducted sensitivity analyses which included controls for age into both the primary outcome model and the secondary mixed measures model.

The results of these additional analyses are presented in Tables 18 and 19 below and follow the results from the primary and secondary analysis. That is to say, we do not find evidence of an impact of the ONE on treated children when accounting for differences in age. The effect sizes for both EYTN and HTKS are small and close to zero (0.002 and -0.024 respectively), with wide associated confidence intervals and large p-values, suggesting no direct association between assignment to treatment and early numeracy or executive function.

Table 18: Age adjusted primary outcome analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled Mean (95% CI)	n (missing)	Pooled Mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
EYTN Numeracy Test	838 (36)	32.58 (31.54; 33.62)	851 (50)	32.8 (31.72; 33.87)	1689 (838; 851)	0.002 (-0.12; 0.13)	0.98

Table 19: Age adjusted mixed measures secondary outcome analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled Mean (95% CI)	n (missing)	Pooled Mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
HTKS Executive Function Test	818 (41)	51.85 (49.72; 53.97)	805 (78)	53.3 (51.13; 55.47)	1623 (818; 805)	-0.024 (-0.11; 0.06)	0.58

Endline only analysis

During baseline data collection we regrettably became aware that test administrators failed to upload child data on the primary outcome for children in five settings, including all of the primary outcome data for one setting. As a result, the primary analysis model is restricted to only those who have valid EYT-N assessment scores at both baseline and endline, which results in observations which are missing EYT-N scores at baseline being dropped from the complete case analysis. Therefore, we conduct a sensitivity analysis which excludes baseline EYT-N assessment scores from the primary analysis model; this allows us to utilise a slightly larger sample size, but does not allow us to control for differences in baseline attainment.

The results from the endline-only sensitivity analysis (Table 20) are similar to those from the existing primary analysis, strengthening the evidence that there was no impact of the ONE intervention on early years numeracy. The effect size is very close to zero (-0.007) and the confidence intervals are large, spanning zero (-0.15; 0.13), resulting in a large p-value (0.88) and no statistically significant difference between children in the treatment and control groups.

Table 20: Endline only primary outcome analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled mean (95% CI)	n (missing)	Pooled mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
EYTN Numeracy Test	874 (0)	32.52 (31.51; 33.54)	901 (0)	32.89 (31.85; 33.93)	1775 (874; 901)	-0.007 (-0.15; 0.13)	0.88

Measurement error adjustment

During endline testing and after the publication of the SAP, an issue was discovered by RAND in the app used to implement the EYTN Numeracy test, which meant that the stopping rule was incorrectly applied. To test the effect of this

issue on the results of the primary outcome analysis, we re-run the primary model below excluding the affected test results. The results to this sensitivity analysis are outlined in Table 21. There remains no statistically significant difference in outcomes between children in the treatment and the control group (the p-value is 0.92), with an estimated effect size very close to zero (0.005) and confidence intervals spanning either side of zero (-0.12; 0.13). This means that we do not believe the technical issue affects the primary outcome and we can be more certain that our primary outcome is accurate.

As outlined under Methods, the evaluation encountered issues with measurement error in both HTKS-R and Corsi Blocks, with the appearance of item-level missingness in baseline data for a number of children. For HTKS-R, assessor reporting of baseline test completeness frequently directly contradicted the availability of valid data, consistently indicating that assessments were incomplete even when there is valid data and all stopping rules appear to have been followed. As the assessor flag for test incompleteness was inconsistently applied, children with valid HTKS-R scores with assessor-indicated incomplete tests were included in the secondary analysis; however, the inclusion of these children could inadvertently introduce measurement bias. This is chiefly restricted to the HTKS-R test and as such we re-run the results of the secondary outcome models involving HTKS-R (mixed-measures model and the HTKS-R model), excluding all children where the assessor indicated that the test was incomplete. In neither case do we find significant evidence for a non-zero effect size—the p-values are very large in both models (0.58 for the mixed measures model and 0.36 for the HTKS-R model)—indicating that, even when adjusting for potential measurement error bias, the intervention did not have a significant effect on executive function.

Table 21: Measurement error adjusted primary outcome analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled mean (95% CI)	n (missing)	Pooled mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
EYTN Numeracy Test	650 (0)	33.71 (32.5; 34.92)	663 (0)	34.22 (32.95; 35.48)	1313 (650; 663)	0.005 (-0.12; 0.13)	0.92

Table 22: Measurement error adjusted mixed-measures secondary outcome analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled mean (95% CI)	n (missing)	Pooled mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
HTKS-R Test (Corsi Blocks baseline)	630 (1)	55.49 (53.12; 57.86)	613 (8)	56.3 (53.86; 58.73)	1243 (630; 613)	-0.026 (-0.12; 0.07)	0.58

Table 23: Measurement error adjusted HTKS-R secondary outcome analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled mean (95% CI)	n (missing)	Pooled mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value

HTKS-R Test (Corsi Blocks baseline)	631 (0)	55.44 (53.08; 57.81)	621 (0)	56.1 (53.67; 58.53)	1252 (631; 621)	-0.044 (-0.14; -0.05)	0.36
-------------------------------------	------------	-------------------------	------------	------------------------	--------------------	--------------------------	------

Exclusion of unblinded settings

During the randomisation process, a switch was made to concealed randomisation. This resulted in two settings being informed of their treatment status before baseline assessments were completed. To account for any potential bias that this may have introduced, we repeated the primary analysis on a sample which excluded the children from both of the settings to which randomisation was revealed prematurely.

The results of this analysis are presented in Table 24 and, as they are similar to those from the primary analysis, provide no evidence that revealing of treatment allocation biased the initial results. The estimated effect size remains very small (0.016), with wide bootstrapped confidence intervals again suggesting no significant association between treatment assignment and early numeracy skills (as measured by EYT-N assessment scores).

Table 24: Primary outcome analysis excluding unblinded settings

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled mean (95% CI)	n (missing)	Pooled mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
EYTN Numeracy Test	828 (36)	32.572 (31.52; 33.62)	836 (48)	32.568 (31.49; 33.65)	1664 (828; 836)	0.016 (-0.11; 0.14)	0.79

Exclusion of Maths Champions settings

As discussed under Methods, a number of settings were offered Maths Champions, a similar EY intervention targeting early maths development. While these settings were encouraged to delay participation until June 2024, there may have still been a risk of a similar intervention being delivered, reducing the reliability of results of this evaluation. As such, an additional sensitivity analysis was conducted where Maths Champions settings were excluded from the sample. The results of this analysis are presented in Table 25 and provide no indication that inclusion of these settings in the initial primary analysis had an impact on the estimated effect. Again, the negligible effect size (0.01), combined with wide bootstrapped confidence intervals suggest that the intervention probably had no effect on early numeracy among settings that are not participating in Maths Champions.

Table 25: Primary outcome analysis excluding Maths Champions settings

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	N (missing)	Pooled mean (95% CI)	n (missing)	Pooled mean (95% CI)	Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
EYTN Numeracy Test	828 (36)	32.48 (31.43; 33.52)	812 (49)	32.55 (31.45; 33.65)	1640 (828; 812)	0.006 (-0.13; 0.014)	0.92

Implementation and process evaluation results

Fidelity and adaptation

The ONE appears to have largely been delivered as planned, with practitioners particularly positive about training.

Training was well-received and the number of practitioners who attended the training each week was more than what was recommended

Observations of the training suggest that it was clear and aligned with the expected programme activities in the Logic Model. Practitioners seemed to be engaged, contributing during training by offering their own examples, and demonstrating an understanding of the programme and how it should be delivered. Activities were well received and teachers were very positive about how much their children's skills, particularly executive function skills, had improved when the activities were revisited. This is supported by survey data, with practitioners reporting being overwhelmingly positive about the ONE training, with 96% (47 of 49) strongly agreeing or agreeing that 'training prepared me to deliver the ONE'. Interviews with practitioners also support this: practitioners enjoyed the training, and the fact that the training was in-person and regular over a period of time. Practitioners reported that this allowed them to develop a trusting relationship with their trainer.

Interviews with the trainers suggest that the number of practitioners attending varied between settings and from session to session. The ONE requires settings to train one practitioner. In practice, only three settings consistently had just one practitioner attend training; in other settings up to thirteen practitioners attended. Only one setting failed to meet the minimum training requirements of at least one member of staff attending all training sessions. Trainers suggested that this variation was linked to factors such as setting size, staffing, budget, and the time or day when the training was delivered. An analysis of setting compliance data shows that the average number of practitioners in each training session was five, but the mode (most frequently occurring number) of practitioners per training session was three. This suggests that while there was large variation in the number of practitioners in each session, generally, three practitioners was the most frequent number to attend training.

In interviews, practitioners reported that the resources provided by the trainers were useful and detailed, and accessible to colleagues who had not attended the training. One practitioner reported not having attended the training, learning about the content of the training from their manager, but found the activities simple to pick up and incorporate, particularly thanks to the activity cards.

Practitioners report that adaptations were needed to make the ONE activities accessible and engaging for all children

Although a third of practitioners in the survey (37%, 18 of 49 respondents) disagreed or strongly disagreed with the statement 'adaptations beyond those noted in the activity sheets were needed for administering the ONE activities', an almost equal number (35%, 17 of 49) agreed or strongly agreed with the statement. When asked to describe adaptations, 73% (36 of 49) responded, suggesting that there was some need for adaptation. This process of adaptation was a feature of the delivery team's training with the practitioners.

By far the most reported reason for adaptations was to simplify the activities to make them easier for children (61%, 22 of 36 responses), with practitioners in the survey reporting that children did not have the mathematical skills to access the activities, although some (19%, 7 of 36) reported making adaptations to make the activities more challenging. Interviews with practitioners suggest that these adaptations, most often in the form of scaffolding or additional support, were needed because children in this cohort had additional needs owing to learning loss as a result of the Covid-19 pandemic. Practitioners interviewed reported that these adaptations were easy to make when they knew their children and had experience.

Adaptations were generally reported to have improved engagement with the activities: two-thirds (67%, 32 of 48 respondents) said adaptations somewhat improved child responsiveness and outcomes, while a smaller proportion (21%, 10 of 48) said they improved child responsiveness to a great extent. We were keen to understand the extent to

which executive challenge was embedded in activities throughout the programme, however, we were unable to collect this data from the survey, limiting our ability to make conclusions on this.

Dosage

Child absences prevented delivery of some sessions and may have had a negative impact on the ability of the ONE to achieve its aims

Almost half of respondents in the practitioner survey (47%, 23 of 49) reported that child absences prevented delivery of some sessions, with one setting reporting that child absence stopped the setting from delivering the ONE entirely. On average, practitioners reported an average of 3.8 child absences per week per setting over the course of the delivery period, with a median of three.¹⁵ This conceivably would reduce children's exposure to broad, high quality, maths provision that promoted executive functioning—a key output in the logic model. Regular absences would also make it difficult to expose children to increasing levels of executive functioning challenge—another key output on the logic model.

Context, moderators, and mediators

Pupils were engaged with the ONE activities, with small group sessions in particular seen to facilitate engagement. Staff absences potentially reduced the effectiveness of the ONE, though training and support from the ONE team was seen as a key factor in successful delivery.

Small group working was particularly seen as facilitator to child engagement

Overall, practitioners felt that children were engaged in the ONE activities (84%, 41 of 49 respondents); almost all (90%, 44 out of 49) agreed that small group activities facilitated or strongly facilitated child engagement. Practitioners also reported that delivering the ONE activities in a one to one format also facilitated child engagement, but less strongly than small group activities (59%, 29 of 49). Whole-group delivery was reported to have facilitated child engagement by some (45%, 22 of 49), but respondents were more likely to report neutral engagement or hindrance in this type of delivery (55%, 27 of 49).

Distractions within the classroom were seen as the biggest hindrance to engagement with the ONE activities (47%, 23 of 49 respondents). The availability of other activities was not seen as a barrier or facilitator to engagement (61%, 30 of 49 respondents were neutral) nor were specific language or learning needs seen to hinder or facilitate engagement (55%, 27 of 49 were neutral).

Competing priorities and staff absences prevented the delivery of some sessions, potentially reducing the effectiveness of the ONE

Setting managers reported that competing priorities in settings were the biggest barrier, 'slightly' or 'very' negatively impacting the delivery of the ONE (33%, 14 of 42 respondents). This was echoed by practitioners, with 37% (18 of 49) reporting that other priorities often took precedence. In similar vein, almost half of respondents in the practitioner survey reported that staff absences prevented delivery of some sessions (45%, 22 of 49) with the average number of unplanned staff absences being 0.9 in a 'typical' week.

If practitioners are limited in their capacity, it would negatively impact on their ability to achieve some of the short term outcomes in the logic model, such as including maths play-based activities in weekly planning and observing and adapting activities to embed executive challenge.

Training and materials were widely reported as being key to supporting practitioners to deliver the ONE

The training was widely reported as having been key to the successful implementation and delivery of the ONE. Almost all practitioners (92%, 45 of 49 respondents) reported that the training had had a 'slightly' or 'very' positive impact on

¹⁵ There are no national figures for child absences in early years settings for comparison.

their ability to deliver while a similar number (92%, 44 of 48) reported that support from the ONE team and the nature of the activities themselves were also beneficial (92%, 45 of 49).

Generally, children were engaged with the ONE activities, with practitioners suggesting that small group activities particularly supported engagement

Interviewees reported that children were engaged in the ONE activities. This is supported by survey findings, where a significant number of practitioners (84%, 41 out of 49 respondents) agreed or strongly agreed that children were engaged in the activities, with only a small number (12%, 6 of 49) strongly disagreeing.

Almost all (44 of 49) reported that small group activities facilitated or strongly facilitated child engagement. This was corroborated by interviews with practitioners who had found that small group activities were an effective way of engaging children. Over half (59%, 29 of 49) found delivering activities one-on-one facilitated child engagement, and less than half (45%, 22 of 49) reported that whole-group delivery facilitated child engagement, but respondents were more likely to report neutral engagement or hindrance in this type of delivery. It may be that whole-class delivery may be less successful in supporting children to be exposed to the ONE—a key outcome in the logic model.

Unintended consequences

Participating in the ONE did not appear to have negative unintended consequences, in fact, there are emerging findings that participating in the ONE was beneficial for staff retention and had positive impact on staff practice.

The ONE did not negatively impact on retention

One consideration in the logic model is the recruitment and retention crisis in early years settings, with one key research question being the extent to which engagement with the ONE impacts on staff retention. A large proportion of setting managers surveyed (83%, 35 of 42 respondents) reported that the ONE did not impact staff retention, with 16% (7 of 42) believing that it was beneficial to staff retention.

The staffing levels reported on the endline manager survey reinforce this, with no significant differences between treatment and control settings in the number of staff leaving or joining in the previous year. An average of 1.11 staff members left treatment settings over the course of the delivery period compared to 1.89 in control settings, suggesting that the ONE had no significant impact on staff retention.

The ONE did not have unintended adverse effects and had a very positive impact on staff being able to adapt what they'd learnt from the ONE in their wider practice

Many practitioners believed that they were able to carry on with their normal activities while delivering the ONE (82%, 40 of 49 respondents), with managers overwhelmingly agreeing that staff were able to continue with their normal activities (93%, 39 of 42 respondents). This is supported by the fact that the ONE was not seen as requiring significant preparation and planning, with over two thirds of practitioners (70%, 34 of 49) largely disagreeing that the ONE activities required significant time. This could be linked to reports from interviewees and trainers that the most used activities were ones that were (1) modelled by trainers, (2) low resource, or (3) built on activities they were already using or familiar with (such as 'What's the time Mr Wolf?'). There is, however, some evidence that delivering the ONE may have increased workload, with only 47% of practitioners (23 of 49 respondents) disagreeing or strongly disagreeing that the ONE intervention added additional pressure to their workload.

On questions about confidence, over a third of practitioners (69%, 34 of 49 respondents) agreed or strongly agreed that 'as a result of the ONE, I feel more confident to tailor other activities (unrelated to early maths) to a child's level of development'. This could suggest a positive unintended consequence.

Settings reported that the ONE did not interfere with, or crowd out, other structured pedagogical activities

Interviews with practitioners suggest that the ONE did not interfere with, or crowd out, other structured pedagogical activities. This is supported by survey findings, with the majority of practitioners reporting that they not feel that the ONE reduced time spent on activities in early language and literacy (69%, 34 of 49) or early science (61%, 30 of 49). This is

supported by responses from the manager survey with three-quarters of respondents (76%, 32 of 42) reporting that participating in the ONE did not increase the focus on maths and executive functioning at the expense of other subjects.

Programme differentiation

The ONE resulted in changes to practitioners' belief in the importance of executive functioning skills in maths

At baseline, practitioners across both treatment and control agreed that a number of maths skills were 'very important' for early numeracy, including knowing number facts, understanding mathematical learning, thinking flexibly, and understanding how maths is used in the real world. Practitioners across treatment and control also rated 'being able to manipulate abstract information'—a skill linked to executive challenge—as 'moderately important'.

However, after the intervention the majority in treatment settings said that all skills listed above were 'very important', with the proportion of 'very important' responses increasing for many skills, as can be seen in Table 26. Changes were particularly noted in key skills linked with executive functioning—'being able to store and manipulate information in their head' and 'thinking flexibly'—with a 15 and 21 percentage point difference, respectively, between baseline and endline.

Table 26: Practitioners in treatment settings rating early years maths skills as 'very important' at baseline and endline

	Baseline	Endline
Knowing number facts	25/56 (44%)	26/49 (53%)
Understanding mathematical learning	29/56 (51%)	33/49 (67%)
Understanding how mathematics is used in the real world	45/59 (76%)	35/49 (71%)
Being able to store and manipulate information in their head	24/59 (40%)	32/49 (65%)
Focusing on relevant information and ignore distractions	25/59 (42%)	30/49 (61%)
Thinking flexibly	30/59 (50%)	35/49 (71%)
Having good language skills	27/59 (45%)	26/49 (53%)
Thinking creatively	37/59 (62%)	36/49 (71%)
Having good spatial skills	29/59 (49%)	27/49 (55%)

This suggests that the ONE was successful in communicating the importance of executive functioning and different types of skills in early years maths. This is especially true when compared to endline practitioner survey results from the control group where 'moderately important' responses became comparatively more likely, with a plurality of practitioners now saying that a several maths skills were moderately important, including 'being able to store and manipulate information in their head' and 'thinking flexibly'.

Views on the role of the ONE in supporting practitioners' to adapt activities to appropriate levels of challenge and support maths and executive functioning were mixed

Interviews with practitioners in the treatment group suggest that the ONE did not particularly help them to adapt activities to suit the needs of their children—that they had these skills already. This is supported by survey data, with practitioners across treatment and control groups reporting that they were broadly comfortable planning and doing maths activities at their setting, increasing for both groups at endline.

However, this is slightly contradicted by managers in treatment settings, who agree or strongly agree that they had seen an improvement in the ability of the staff to support children with early numeracy (81%, 34 of 42 respondents) and executive function (90%, 38 of 42).

Business as usual

Control settings reported delivering executive functioning activities multiple times a week, but seemed to face greater barriers compared to treatment settings.

Maths activities were delivered in small groups across treatment and control, but it is unclear if there is significant difference in the time spent delivering maths in small groups over the trial period

Like the ONE, practitioners in control settings were most likely to report doing maths and executive functioning activities in small groups with 81% (30 of 37 respondents) delivering maths activities in small groups compared to 68% (25 of 37) delivering them one-on-one and in the whole class.

There were some differences in the time spent doing small group and whole-class activities across the two groups, but these were not consistent nor confined to one trial arm. Treatment settings were 12 percentage points more likely to report spending one to two hours per day in small group activities than control settings (31%, compared to 19%), an increase of 9 percentage points compared to baseline (22%). However, at endline, control settings were 12 percentage points more likely to spend more than two hours in small groups compared to baseline (23%, 7 of 31 respondents, compared to 11%, 5 of 46). There was a slight increase in the number of treatment practitioners reporting spending more than two hours in small groups, but this was minimal.

Practitioners in control settings were just as likely as those in the ONE to report supporting executive challenge multiple times a week

At baseline, most practitioners reported supporting executive challenge multiple times a week (doing activities that involve multiple steps or long explanations, waiting and taking turns, coming up with new solutions). This is broadly the same across treatment and control, especially for activities taking turns, reported by a majority of practitioners to be practiced more than once a day at baseline (treatment: 73%, 43 of 59; control: 74%, 34 of 46). Asking children to come up with new ways to solve problems was also frequently reported, with a majority of practitioners across both groups reporting doing activities that involved this at least four or five times per week at baseline (treatment: 69%, 40 of 58; control: 71%, 32 of 45).

At endline, there are similar increases in practitioner-reported frequency of taking turns activities happening every day (treatment: 80%, 39 of 49; control: 81%, 25 of 31). At endline, control group practitioners were only moderately less likely to do activities involving asking children to come up with new solutions every day than those in the treatment group (30%, 9 of 30 compared to 37%, 18 of 49, respectively).

However, these results should be interpreted with caution as it is difficult to understand the underlying degree of executive challenge across both arms when examples of activities which could have executive challenge embedded are ubiquitous—for example, impulse control is a key pillar of executive function but activities that practice impulse control are also those that are developmentally appropriate and widespread for this age group—and difficult to capture in self-report data such as a survey. For example, it could be that practitioners included routine turn-taking and circle time in their answer to these survey questions, even though these routine activities are not necessarily scaffolding executive function or embedding executive challenge. It remains plausible that, while the reported number of times these activities are undertaken is similar across treatment and control, the level of executive challenge being embedded into these activities may be higher within the treatment group as a result of the ONE programme. Such a distinction is difficult to capture in a survey.

Barriers to delivering maths activities were similar across control and treatment but were more significant for control practitioners

Interestingly, practitioners in control settings were more likely to report that staff absence had a slightly or very negative impact on delivery of maths activities (67.74%, 21 of 31 respondents, compared to 45%, 22 of 49). Those in control settings were also more likely to report that child absence had a negative impact on maths activities: 74% (23 of 31) compared to 47% (23 out of 49). In similar vein, practitioners in control settings were more likely to report that competing priorities had a negative impact on the delivery of maths activities (control: 52%, 23 of 31; treatment: 37%, 18 of 49). Over half of practitioners in control settings (55%, 17 of 31) reported that difficulty engaging children had a negative

impact on their maths activity delivery, compared to only 12% (6 of 49) in treatment settings. Taken together this suggests that practitioners in control settings faced more significant barriers delivering maths activities.

Cost

The cost estimates presented in Table 27 and Table 28 are derived from data reported by the delivery team regarding costs for settings to implement the ONE in the trial. We also collected costs from the settings involved in this trial. As this is a whole-class intervention, the per child costs are based on the average number of children in the preschool class (32.5, on average, across all treatment and control settings), rather than the average number of children tested.

The cost of programme administration and support was estimated to be £3,070 for the first year of delivery to a school; this include recruiting schools, project management, training-the-trainer, and travel. The recurring costs related to the above are lower, estimated to be £1,535 per year, which is 50% of the start-up costs per setting. The recurring costs do not include setting recruitment or training-the-trainers costs as it is assumed these are incurred only once, but does still include project management and travel to allow for face to face delivery of the ONE to new staff at the setting each year.

The minimum number of practitioners that need to be trained is one per setting, but in practice an average of two per setting were trained, with some settings choosing up to five. It is important to note that the staff costs for training represent the total amount required for training in this trial as delivered by graduate researchers in the University of Oxford (£19.41 per hour) delivering three hours of training (PD sessions: 4 x 30 minutes per setting), time supporting activity delivery in Week 8 (30 minutes), and time for Week 12 closure and reflection call (30 minutes)). No travel costs were included in this estimate but this would need to be considered for future roll-out given that the delivery team delivered face to face training in each setting.

Furthermore, each setting was re-imbursed a flat amount of £75 for EY practitioners to attend training ('backfill' is the term used by the EEF for this cost) as opposed to doing it in their own time. Due to the high rate of staff turnover across EY settings, the evaluation team assumed a re-training rate of one practitioner a year per settings, therefore leading to recurring costs for both training delivery and staff cover.

Additionally, there was the cost of materials, which included laminated activity cards and complementary materials that were offered to support settings but not required to deliver activities (for example, folder, plastic box to contain resources, art and craft resources). The delivery team estimated that, of these materials, around £8 would be needed to be spent in subsequent years to replace the arts and craft materials (for example, glue, Blu Tack, coloured foam materials). Our manager survey suggests that settings needed to purchase additional resources beyond those provided by the delivery team (for example, unifex cubes, shapes, number pegs, dice, relational rods, balance mats) ranging between £10 and £200 (average £76.25, median = £55). We have included this in the Table 27. The majority of these resources were relatively substantial so were considered a one-off set-up cost.

Overall, the cost of delivering the ONE for the first year is £3,389.80; over three years this equates to £6,740.88 per setting. Over three years, the cost per child would be £69.14.

Table 27: Cost of delivering the ONE

Item	Type of cost	Cost (for Year 1)	Total cost over 3 years	Total cost per child per year over 3 years
Programme support (including recruiting schools, project management, training trainers and travel costs)	Start up and recurring cost per setting	£3,069.68	£6,139.36	£62.97
Training (4 x 30 minutes sessions and 1 x 12 week call) per setting	Start-up and reoccurring cost per setting	£58.23	£174.69	£1.79

Cost of covering staff attending training (average 2 practitioners trained per setting, range 1 – 5 practitioners)	Start-up and reoccurring cost per setting	£150 (£75-£375)	£300	£3.08
Materials for delivery provided by the delivery team (inc laminated activity cards)	Start-up and reoccurring cost per setting	£35.64	£50.58	£0.52
Additional materials purchased by setting for delivery	Start-up	£76.25 (£10 – £200)	£76.25	£0.78
Total		£3,389.80	£6,740.88	£69.14

Table 28: Cumulative costs of delivering the ONE—assuming delivery over three years

	Year 1	Year 2	Year 3
The ONE	£3,389.80	£5,065.34	£6,740.88

Managers also reported having a number of existing resources that they used to deliver the ONE, including sport equipment, whiteboards, coloured blocks, beanbags, computers and tablets, number cards, chalks, and chalk boards. They reported that these resources cost an average of £500 (range: £40 to £2,100; median: £90). These were not included in the cost table above as managers had these resources in settings prior to the intervention.

While not typically considered a 'direct' cost for settings, the time practitioners are expected to spend preparing and delivering the ONE is an important factor when evaluating broader resources needed. According to the delivery team, practitioners need time to:

- select three activities per week;
- decide when to integrate them into the weekly routine—for example, run them as circle time activities or smaller group activities;
- familiarise themselves with activity cards; and
- gather materials.

The team estimates that this amounts to 20 planning minutes per week (with support provided by the delivery team for this in Weeks 1 to 4); activities last five to ten minutes (average estimate 7.5 minutes per activity) leading to an average estimate of 22.5 minutes per week. Therefore, the team estimates that the total time spent per week on planning and delivery is 45 minutes. In contrast, managers reported that practitioners spent, on average, 3.63 hours per week delivering the ONE (range: zero to 20 hours; median: two hours).

Finally, we sought to use the survey to assess the costs of delivering similar interventions in the control settings. Typically, they reported buying general professional development and subscriptions (for example, Hamilton, Twinkl) to support their maths and executive functioning teaching and learning. This amounted to an average over the year of £6,126.67 (£100 to £35,000; median: £1,450). Managers also reported the time spent weekly on maths activities to be 9.84 hours (zero to 75 hours; median: 5) and time spent on activities elevated to include executive function challenge to be 12.72 hours (zero to 60 hours; median: 6). Both costs per year and time spent per week are particularly higher in control settings compared to the ONE. However, it is worth noting that the general PD and subscriptions that control groups reported using may be relevant to all domains of the EYFSP: it is not clear to what extent settings would stop using these PD resources when receiving the ONE.

Conclusion

Table 29: Key conclusions

Key conclusions	
1.	Children in the ONE settings made no additional progress in maths, on average, compared to children in control settings. This result has a high security rating.
2.	Children in the ONE settings made no additional progress in executive functioning, on average, compared to children in control settings.
3.	Among children receiving Early Years Pupil Premium (EYPP), those in the ONE settings made two additional months' progress in maths, on average, compared to children in control settings. These results may have a lower security than the overall findings because of the smaller number of children.
4.	There is evidence to suggest that the training and support offered by the ONE team were well received and led to changes in practitioners' understanding of the importance of executive functioning to mathematical attainment.

Impact evaluation and IPE integration

Evidence to support the logic model

The impact evaluation suggests that the ONE does not result in improved maths attainment or executive functioning as hypothesised in the logic model. There is evidence from the IPE that the staff-level inputs—training, materials, and support from the ONE team—were very well received by practitioners who engaged with the training materials as outlined in the logic model. There is also evidence that the ONE resulted in changes to practitioners' belief in the importance of executive functioning skills in maths—a key short-term outcome in the logic model.

Practitioners report delivering between two and three activities per week, with children more engaged in the ONE activities than the corresponding activities in control settings. Over 20% of all settings delivered 36 activities, with settings, on average, delivering 28.67 activities over the 12 week periods (equivalent to 2.375 activities per week). There is emerging evidence from the IPE that children in the ONE were more engaged in these activities compared to children exposed to business-as-usual activities in control settings. Jointly, these findings are in line with the logic model outputs: 'Practitioners deliver activities in the classroom, gauging children's engagement and implementing instructions to scaffold executive functioning.'

At the child level, the logic model hypothesises that children are exposed to a regular programme of adult-led activities, with activities increasing in difficulty of executive functioning challenge. The activities seem to be delivered as intended. Practitioners reporting that adaptations were needed to make the ONE activities accessible and engaging for all children, with the majority needing to simplify the activities as children did not have the mathematical skills to access them, rather than having to increase the challenge level.

Looking closer at IPE data, there is some evidence to suggest that child absences may have also led to lack of exposure to the ONE, thus diluting the impact of the programme on intended outcomes in a twelve week intervention. There is no nationally available data on average absences at the child level in early years settings, however, for a relatively short-term intervention, absences would reduce children's exposure to broad, high quality, maths provision, embedded with executive functioning—a key output in the logic model. Regular absences would also make it difficult to expose children to increasing levels of executive functioning challenge—another key output on the logic model.

Interpretation

The impact evaluation suggests that the ONE likely does not result in improved maths attainment or executive function. This does not change once a CACE analysis is applied on those who implemented the intervention as intended, suggesting that even when delivered with fidelity the ONE likely does not impact on the outcomes measured in this trial. There was a relatively high degree of error with the administration of the measurements in the baseline and, to a lesser

extent, endline, which may have impacted on the results. However, the analyses we conducted that attempt to account for measurement error also did not find an impact of the ONE on the measures selected for this trial.

There is evidence in the IPE that the ONE training was well received by practitioners and lead to real change in practitioners awareness and understanding. It is clear that the training, materials, and support provided by the ONE team were received very positively by practitioners and that the ONE increased their understanding of the importance of executive function in early maths development.

Looking at previous studies, where there was evidence that integrating executive challenge into play-based activities resulted in improvements in executive functions for the intervention settings (Howard et al., 2020), there are some differences in the interventions. The original intervention was six months long, with two months of training for practitioners, monthly one-hour professional development teleconferences, and an optional formative self-regulation assessment tool for practitioners to use. This is considerably more training and support for practitioners and longer time for the intervention to be embedded and delivered. This increased duration in the original study may have further supported practitioners in implementing the ONE activities and allowed for children to receive a higher dosage. Another previous study (Scerif et al., 2023), almost exactly replicating the delivery in this trial, found a greater impact on executive function and numeracy skills (as measured by Corsi Blocks and EYTN) in settings classed as having high intervention adherence. However, these findings are limited by the fact that the trial consisted of fifteen settings and was not independently evaluated.

Overall, findings suggest that the ONE as delivered in this trial likely did not have an impact on the desired child-level outcomes.

Limitations and lessons learned

By far the greatest threat to the validity of the findings presented here is the measurement error. We are aware that test administrators faced some difficulties in administering the executive functioning tests at baseline and that this resulted in high attrition from HTKS-R compared to both EYTN and Corsi Blocks, as well as concern over validity of the data, particularly at baseline. Additional training was implemented alongside quality assurance checks to improve administration at endline, and additional statistical tests introduced to understand the extent to which these errors may have impacted on findings. However, as much as actions were taken to mitigate the potential effects to validity and reliability, we do acknowledge that measurement error at baseline and floor effects in secondary outcomes could impact on the robustness of the findings presented in this report. However, the sensitivity and additional analyses that we conducted would suggest that these did not introduce bias. Nevertheless, we would still recommend caution in interpreting results for the secondary outcomes

The lack of actual child-level participation data is also an important limitation. While the evaluation team would have preferred to use actual child attendance records at each activity, it is not feasible for settings to collect attendance data to this level of granularity. As such, the team opted instead to use attendance patterns reported by settings. Being unable to map dosage in this way does limit the ability of this evaluation to robustly estimate the impact of the ONE on children as it risks underestimating the number of the ONE activities children were exposed to.

We are also aware of the role that quality plays in early years provision, with high process quality (for example, delivery of teaching and learning) and high structural quality (for example, staffing requirement, physical environment) facilitating positive child experiences and interaction (Janta et al., 2016). However, we were unable to explore how quality may have interacted with the ONE to impact results. Future studies would benefit from exploring these important contextual factors.

Future research and publications

While results from the impact evaluation suggests that the ONE does not result in improved maths attainment or executive functioning, evidence from the IPE suggests that the ONE resulted in changes to practitioners' belief in the importance of executive function skills in maths, a relatively high level of compliance with delivery (though not with the executive functioning elements of the activities), and high child-level engagement. This is positive as it suggests that the ONE has high 'buy-in' from practitioners and is engaging for children. Furthermore, the IPE and comparison to

previous studies suggest areas for improvement (for example, increasing duration of training and activity delivery duration of the intervention to increase level of exposure to the intervention for both practitioners and children). Given the relatively low-cost of the intervention this feels like a potential option for future iterations (though 'buy-in' and costs may be impacted as a result of increasing the length of the intervention).

In light of the difficulties experienced with baselining children in this trial, it is recommended that future early years evaluations increase resources devoted to recruiting and training assessors to ensure assessors have the experience to baseline appropriately. Future evaluations of the ONE may wish to choose a different combination of outcomes as some measures, such as HTKS-R, are more complicated for assessors to deliver and more prone to floor effects.

Similarly, there were difficulties in measuring the prevalence and quality of executive function scaffolding and support via survey. There were concerns that staff would neither recognise the term 'executive function' nor understand what activities might support its development, particularly in the baseline and control arm surveys. Further refinement of IPE survey tools to understand what executive function support already exists in settings, and how this support changes as a result of the intervention, is needed in future analysis to ensure the findings from the impact evaluation can be better understood.

In terms of wider representativeness, this evaluation involved both PVI and SBS settings, so findings can be applied more widely across settings in England. However, it is also important to underline the fact that this evaluation took place as part of the Stronger Practice Hubs initiative in England, whereby early practitioners were supported in varying degrees by local early years hubs. This evaluation did not explore the role of Stronger Practice Hubs in the evaluation, though anecdotally, there was a high degree of regional variation in SPH activity. Future evaluations could benefit from exploring wider regional initiatives and how they may (or may not) interact with programmes being evaluated.

References

- Angrist, J. D. (2006) 'Instrumental Variables Methods in Experimental Criminological Research: What, Why and How', *Experimental Criminology*, 2 (1), pp. 23–44.
- Angrist, J. D. and Krueger, A. B. (1991) 'Does Compulsory School Attendance Affect Schooling and Earnings?', *Quarterly Economics*, 106 (4), pp. 979–1014.
- Blair, C. and Raver, C. C. (2014) 'Closing the Achievement Gap through Modification of Neurocognitive and Neuroendocrine Function: Results from a Cluster Randomized Controlled Trial of an Innovative Approach to the Education of Children in Kindergarten', *PLoS ONE*, 9 (11).
- Blair, C. and Razza, R. P. (2007) 'Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten', *Child Development*, 78 (2), pp. 647–663.
- Blakey, E., Matthews, D., Cragg, L., Buck, J., Cameron, D., Higgins, B., Pepper, L., Ridley, E., Sullivan, E. and Carroll, D. J. (2020) 'The Role of Executive Functions in Socioeconomic Attainment Gaps: Results from a Randomized Controlled Trial', *Child Development*, 91, pp. 1594–614.
- Bonetti, S. and Blanden, J. (2020) 'Early Years Workforce Qualifications and Children's Outcomes: An Analysis Using Administrative Data', Education Policy Institute.
- Clarke, D. (2021) 'Rwolf2 Implementation and Flexible Syntax', <https://www.damianclarke.net/computation/rwolf2.pdf>
- Clarke, D., Romano, J. and Wolf, M. (2019) 'The Romano-Wolf Multiple Hypothesis Correction in Stata', IZA Institute of Labor Economics.
- Coolen, I., Merkley, R., Ansari, D., Dove, E., Dowker, A., Mills, A., Murphy, V., von Spreckelsen, M. and Scerif, G. (2021) 'Domain-General and Domain-Specific Influences on Emerging Numerical Cognition: Contrasting Uni- and Bidirectional Prediction Models', *Cognition*, 215.
- Corsi, P. M. (1972) Human Memory and the Medial Temporal Region of the Brain'. *Dissertation Abstracts International*, 34(2-B), 891.
- Dong, N. and Maynard, R. (2013) 'PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies', *Research on Educational Effectiveness*, 6 (1), pp. 24–67.
- EEF (2022a) 'Implementation and Process Evaluation Guidance for EEF Evaluations', London: Education Endowment Foundation.
- EEF (2022b) 'Statistical Analysis Guidance for EEF Evaluations', London: Education Endowment Foundation.
- EEF (2023) 'Cost Evaluation Guidance', London: Education Endowment Foundation.
- Glennerster, R. and Takavarasha, K. (2013) *Running Randomized Evaluations: A Practical Guide*, London: Princeton University Press.
- 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.
- Grund, S., Lüdtke, O. and Robitzsch, A. (2018) 'Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations', *Organizational Research Methods*, 21 (1), pp. 111–149.
- Hedges, L. V. (2007) 'Effect Sizes in Cluster-Randomized Designs', *Educational and Behavioral Statistics*, 32 (4), pp. 341–370. <https://doi.org/10.3102/1076998606298043>
- Howard, S. J., Vasseleu, E., Batterham, M. and Neilsen-Hewett, C. (2020) 'Everyday Practices and Activities to Improve Pre-school Self-Regulation: Cluster RCT Evaluation of the PRSIST Program', *Frontier Psychology*, 11 (137).
- Howard, S. J., Vasseleu, E., Neilsen-Hewett, C. and Cliff, K. (2018) 'Evaluation of the Preschool Situational Self-Regulation Toolkit (PRSIST) Program for Supporting Children's Early Self-Regulation Development: Study Protocol for a Cluster Randomized Controlled Trial', *Trials*, 19 (1), p. 64. <https://doi.org/10.1186/s13063-018-2455-4>
- Howard, S. J., Neilsen-Hewett, C., de Rosnay, M., Melhuish, E. C. and Buckley-Walker, K. (2022) 'Validity, Reliability and Viability of Pre-School Educators' Use of Early Years Toolbox Early Numeracy', *Australasian Early Childhood*, 47 (2), pp. 92–106

- Howard, S., Melhuish, E. and Chadwick, S. (2023) Early Years Toolbox website: 'Norms'.
<http://www.eytoolbox.com.au/toolbox-norms>
- Hutchison, D. and Styles, B. (2010) 'A Guide to Running Randomised Controlled Trials for Educational Researchers', Slough: NFER.
- Jakobsen, J. C., Gluud, C., Wetterslev, J. and Winkel, P. (2017) 'When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomised Clinical Trials: A Practical Guide with Flowcharts', *BMC Medical Research Methodology*, 17 (1), pp. 1–10.
- Janta, B., van Belle, J. and Stewart, K. (2016) Quality and Impact of Centre-Based Early Education and Care', RAND Europe.
- Lachowitz, M., Preacher, K. and Kelley, K. (2018) 'A Novel Measure of Effect Size for Mediation Analysis', *Psychological Methods*, 22 (2), p. 244.
- Lee, T. and Shi, D. (2021) 'A Comparison of Full Information Maximum Likelihood and Multiple Imputation in Structural Equation Modeling with Missing Data', *Psychological Methods*, 26 (4), pp. 466–485
- Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W. and Wilson, S. J. (2017) 'Learning-Related Cognitive Self-Regulation Measures for Prekindergarten Children: A Comparative Evaluation of the Educational Relevance of Selected Measures', *Educational Psychology*, 109 (8), p. 1084.
- MacKinnon, D. (2012) *Introduction to Statistical Mediation Analysis (1st edn)*, Routledge.
- McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A. and Pratt, M. E. (2014) 'Predictors of Early Growth in Academic Achievement: The Head-Toes-Knees-Shoulders Task', *Frontiers in Psychology*, 5, p. 599.
- McClelland, M. M., Gonzales, C. R., Cameron, C. E., Geldhof, G. J., Bowles, R. P., Nancarrow, A. F., Mercurief, A. and Tracy, A. (2021) 'The Head-Toes-Knees-Shoulders Revised: Links to Academic Outcomes and Measures of EF in Young Children', *Frontiers in Psychology*, 12, p. 721846.
- Moss, J., Bruce, C. D., Caswell, B., Flynn, T. and Hawes, Z. (2016) *Taking Shape: Classroom Activities to Improve Young Children's Geometric and Spatial Thinking*, Toronto, ON: Pearson.
<https://wordpress.oise.utoronto.ca/robertson/>
- Paull, G. and Popov, D. (2019) 'The Role and Contribution of Maintained Nursery Schools in the Early Years Sector in England', London: Department for Education.
- Pieters, R. (2017) 'Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication', *Consumer Research*, 44 (3), pp. 692–716.
- Preacher, K. J. and Hayes, A. F. (2004) 'SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models', *Behavior Research Methods, Instruments and Computers*, 36, pp. 717–731.
- Purpura, D. J. and Lonigan, C. J. (2015) 'Early Numeracy Assessment: The Development of the Preschool Early Numeracy Scales', *Early Education and Development*, 26, pp. 286–313.
<https://files.eric.ed.gov/fulltext/EJ1050571.pdf>
- Raudenbush, S. W. and Bloom, H. S. (2015) 'Learning About and from a Distribution of Program Impacts Using Multisite Trials', *American Evaluation*, 36 (4), pp. 475–499.
- Scalise, N. R., Daubert, E. N. and Ramani, G. B. (2017) 'Narrowing the Early Mathematics Gap: A Play-Based Intervention to Promote Low-Income Preschoolers' Number Skills', *Numerical Cognition*, 3 (3), pp. 559–581.
<https://doi.org/10.5964/jnc.v3i3.72>
- Scerif, G., Gattas, S., Hawes, Z., Howard, S., Merkley, R. and O'Connor, R. (2023) 'Orchestrating Numeracy and The Executive: The One Programme', *PsyArXiv*. (7 March). <https://doi.org/10.31234/osf.io/2gxzv>
- Scerif, G., Sučević, J., Andrews, H., Blakey, E., Gattas, S. U., Godfrey, A., Hawes, Z., Howard, S. J., Kent, L., Merkley, R., O'Connor, R., O'Reilly, F. and Simms, V. (2025) 'Enhancing Children's Numeracy and Executive Functions Via Their Explicit Integration', *NPJ Science of Learning*, 10 (1), p. 8. <https://doi.org/10.1038/s41539-025-00302-9>
- Senn, S. (1994) 'Testing for Baseline Balance in Clinical Trials', *Statistics in Medicine*, 13, pp. 1715–1726.
- Speciani, E. R., Groom, M., Hesketh, R. and Merewood, J. (2024) 'The ONE Statistical Analysis Plan', London: Education Endowment Foundation.

- Speciani, E. R., Groom, M. and Angell, S. (2023) 'The ONE Evaluation Protocol', London: Education Endowment Foundation.
- Speciani, E. R., Groom, M., Hesketh, R. and Merewood, J. (2024) 'The Orchestrating Numeracy and the Executive (the ONE) Statistical Analysis Plan', London: Education Endowment Foundation.
- StataCorp. (2023) *Stata 18 Structural Equation Modeling Reference Manual*, College Station, TX: Stata Press.
- Torgerson, C. and Torgerson, D. (2013) 'Randomised Controlled Trials in Education: An Introductory Handbook', London: Education Endowment Foundation.
- Vallis, D., Singh, A., Uwimpuhwe, G., Higgins, S., Xiao, Z., De Troyer, E. and Kasim, A. (2022) 'EEFANALYTICS: Stata Module for Evaluating Educational Interventions using Randomised Controlled Trial Designs. <https://EconPapers.repec.org/RePEc:boc:bocode:s458904>.
- Verdine, B. N., Irwin, C. M., Golinkoff, R. M. and Hirsh-Pasek, K. (2014) 'Contributions of Executive Function and Spatial Skills to Preschool Mathematics Achievement', *Experimental Child Psychology*, 126, pp. 37–51. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4107032/>
- Watson, S. (2024) 'Package 'crctStepdown': Univariate Analysis of Cluster Trials with Multiple Outcomes'. <https://cran.r-project.org/web/packages/crctStepdown/crctStepdown.pdf>
- Yu, Q. and Bin, L. (2022) *Statistical Methods for Mediation, Confounding and Moderation Analysis Using r and SAS*, Chapman and Hall/CRC. <https://www.routledge.com/Statistical-Methods-for-Mediation-Confounding-and-Moderation-Analysis/Yu-Li/p/book/9780367365479>

Appendix A: EEF cost rating

Figure 3: Cost rating

Cost rating	Description
£ £ £ £ £	<i>Very low:</i> less than £80 per child per year.
£ £ £ £ £	<i>Low:</i> up to about £200 per child per year.
£ £ £ £ £	<i>Moderate:</i> up to about £700 per child per year.
£ £ £ £ £	<i>High:</i> up to £1,200 per child per year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per child per year.

Appendix B: Security classification of trial findings

OUTCOME:

Please use this template to assign a separate security rating for each primary outcome.

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	MDES	Attrition					
5	Randomised design	<= 0.2	0-10%	5				
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%					4
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%					
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	Adequate allocation sequence. Concealment was compromised in two settings – sensitivity analysis showed no impact of this on the estimate of the effect size. Balanced groups at baseline in terms of background characteristics. Neither peer reviewer had concerns about confounding.
Threat 2: Concurrent Interventions	Low	Sensitivity analysis indicated no evidence of bias for those settings that were offered the Maths Champions intervention. Neither peer reviewer had concerns about concurrent interventions but one commented that the level of EF activity in BAU settings suggests The ONE is not boosting teachers EF related teaching enough.
Threat 3: Experimental effects	High	One peer reviewer rated this as high and the other as low risk. Given that the IPE indicated that >50% of control settings were implementing similar activities to the treatment group, this is rated as high risk following EEF guidance. It is noted that the control settings were perhaps spending less time on these activities and reporting more barriers to implementing these activities.
Threat 4: Implementation fidelity	Moderate	Both peer reviewers rated this as a moderate risk. Implementation fidelity was well defined by three different measures at both the setting and child level. Varying degrees of compliance were recorded (depending on the definition of compliance) and compliance was very high for practitioners attending professional development sessions, with only one setting not meeting the minimum requirement. However, settling-level dosage analysis reveals that only just over 20% of settings delivered the required 36 activities (although no additional benefit to the ONE intervention based on exposure to activities was found). The IPE suggests that there were some issues with implementation in treatment

		settings, particularly around child absences, competing priorities and distractions within settings, and to a lesser extent staff absence, which may have contributed to dilution of intervention effects.
Threat 5: Missing Data	Low	Both reviewers rated this as low risk. Total missing data is low, and the characteristics of missing data is largely balanced between intervention and control groups (with a couple of exceptions). Endline sensitivity analysis showed no impact of missing baseline data. Response rate for primary outcome is very good.
Threat 6: Measurement of Outcomes	Low	Both reviewers rated this as low risk. There were some issues with data collection, but they mostly affected one of the secondary outcome measures, and didn't raise substantial concerns about validity of the measures themselves. Additional sensitivity analyses to account for data collection issues indicate that it was unlikely to have impacted the estimated effect of the intervention.
Threat 7: Selective reporting	Low	Study is registered and protocol and SAP published and followed. No evidence of selective reporting.

- **Initial padlock score:** 5 Padlocks – RCT, MDES at randomisation and analysis = 0.20, attrition between randomisation and analysis = 9.1% overall.
- **Reason for adjustment for threats to validity:** One padlock dropped – there is one moderate risk in relation to implementation fidelity and one high risk in relation to experimental effects owing to a large proportion of the control group implementing similar activities to the treatment group. According to EEF's security rating system, one padlock should therefore be dropped for risks to internal validity.
- **Final padlock score:** initial score adjusted for threats to validity = 4 Padlocks

Appendix C: Effect size estimation and additional tables

Estimation of effect sizes

Appendix table 1: Effect size estimation (primary analysis)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	N (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled Variance
EYTN	-0.22	0.11	838 (36)	234.56	851 (50)	255.05	244.89

Appendix table 2: Effect size estimation (secondary analysis)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	N (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
HTKS (Corsi Baseline)	-1.37	-0.93	818 (41)	957.26	811 (70)	983.18	970.16
HTKS (HTKS baseline)	-0.89	-1.62	842 (17)	955.95	843 (38)	1001.95	978.96
Corsi Blocks (Corsi Blocks baseline)	0.14	0.12	825 (42)	7.15	818 (73)	6.89	7.02

Appendix table 3: Effect size estimation (CACE analysis)

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	N (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance
EYTN (Setting-level compliance)	-0.22	0.57	838 (36)	234.56	851 (50)	255.05	244.89
EYTN (Setting-level dosage)	-0.22	-0.005	838 (36)	234.56	851 (50)	255.05	244.89
EYTN (Child-level dosage)	-0.22	-0.009	838 (36)	234.56	851 (50)	255.05	244.89

EYTN (Non-compliant setting excluded)	-0.21	0.35	830 (36)	236.64	851 (50)	255.05	245.82
--	-------	------	----------	--------	----------	--------	--------

Appendix table 4: Effect size estimation (missing data analysis)

Outcome			Intervention group		Control group		Pooled variance
	Unadjusted differences in means	Adjusted differences in means	N (missing)	Variance of outcome	n (missing)	Variance of outcome	
EYTN	-0.16	-0.17	0	231.95	0	253.54	242.79

Appendix Table 6: Effect size estimation (EYPP sub-group analysis)

Outcome				Intervention group		Control group		Pooled variance
	Unadjusted differences in means	Adjusted difference in means in non-EYPP sub-group	Adjusted difference in means in EYPP sub-group	N (missing)	Variance of outcome	n (missing)	Variance of outcome	
EYTN (EYPP subgroup)	1.25	NA	1.8	130 (6)	174.76	137 (7)	149.77	161.93
EYTN (accounting for EYPP status)	-0.22	-0.18	1.71	838 (36)	234.56	851 (50)	255.05	244.89

Appendix Table 7: Effect size estimation (sensitivity analysis)

Outcome			Intervention group		Control group		Pooled variance
	Unadjusted differences in means	Adjusted differences in means	N (missing)	Variance of outcome	n (missing)	Variance of outcome	
EYTN (accounting for age)	-0.22	0.03	838 (36)	234.56	851 (50)	255.05	244.89
HTKS (Corsi Blocks Baseline; Accounting for Age)	-1.46	-1.24	818 (41)	957.26	805 (76)	983.19	970.12
EYTN (endline only)	-0.37	-0.11	874 (0)	233.21	901 (0)	251.17	242.33

EYTN (measurement error adjusted)	-0.51	0.07	650 (0)	246.68	663 (0)	276.57	261.77
HTKS (Corsi Blocks Baseline; measurement error adjusted)	-1.41	-0.92	813 (37)	949.69	803 (66)	979.64	964.58
HTKS (HTKS Baseline; measurement error adjusted)	-0.95	-1.55	835 (15)	946.43	833 (36)	995.06	970.71
EYTN (excluding unblinded settings)	0.004	0.25	828 (36)	235.86	836 (48)	252.29	244.11
EYTN (excluding maths champions settings)	-0.08	0.1	828 (36)	235.45	812 (49)	255.8	245.53

Estimation of ICC

Appendix Table 8: Estimation of ICC

Outcome	Follow-up (95% CI)
ICC (primary model)	0.188 (0.13- 0.215) (n=1689)
ICC (EYPP sub-group)	0.098 (1.652e-19- 0.215) (n=267)
ICC (empty model)	0.141 (0.094- 0.183) (n=1775)

Results of first stage regression in CACE model

Appendix Table 9: first stage results

Practitioner constantly present at training	Fixed effects:					
		Estimate	Std. Error	df	t value	Pr(> t)
	(Intercept)	-2.081e-02	4.063e-03	2.715e-02	-5.120	0.894
	treatment	9.865e-01	3.600e-03	2.779e-02	274.005	0.799
	SettingRegionEast Midlands	2.428e-02	4.756e-03	2.778e-02	5.105	0.892
	SettingRegionEast of England	2.034e-02	4.848e-03	2.778e-02	4.195	0.897
	SettingRegionLondon	-8.857e-15	1.823e-08	2.787e-02	0.000	1.000
	SettingRegionYorkshire and Humber	2.037e-02	5.802e-03	2.774e-02	3.511	0.902
	SettingtypeMaintained	1.213e-02	3.789e-03	2.779e-02	3.201	0.904

Setting completed 32 the ONE activities	Fixed effects:						
		Estimate	Std. Error	df	t value	Pr(> t)	
	(Intercept)	-1.296e+00	2.944e-01	4.994e-01	-4.403	0.304807	
	treatment	2.855e+01	2.679e-01	1.417e+00	106.563	0.000973	***
	SettingRegionEast Midlands	4.698e-01	3.499e-01	1.034e+01	1.343	0.208143	
	SettingRegionEast of England	2.169e+00	3.495e-01	2.609e-01	6.207	0.445136	
	SettingRegionLondon	3.283e-12	5.583e-07	1.136e+01	0.000	0.999995	
	SettingRegionYorkshire and Humber	-1.770e-01	4.345e-01	1.388e+00	-0.407	0.737664	
	SettingtypeMaintained	1.310e+00	2.802e-01	9.188e+00	4.677	0.001095	**

	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						
Child attends for at least 15 hours a week	Fixed effects:						
		Estimate	Std. Error	df	t value	Pr(> t)	
	(Intercept)	-3.48559	1.14842	150.45531	-3.035	0.00283	**
	treatment	24.51659	0.85639	140.97673	28.628	< 2e-16	***
	B_score	0.02451	0.01134	1828.61936	2.161	0.03085	*
	SettingRegionEast of England	1.43327	1.18625	139.59557	1.208	0.22900	
	SettingRegionLondon	3.90382	1.13549	141.40649	3.438	0.00077	***
	SettingRegionYorkshire and Humber	1.05018	1.40117	139.08234	0.750	0.45482	
	SettingtypePVI	2.21049	0.89795	140.70174	2.462	0.01504	*

	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Baseline endline correlations

Appendix table 10: correlations between baseline and endline outcomes by model

Model	Pearson correlation coefficient	Spearman Rank correlation coefficient
Primary model (EYT-N)	0.655	0.671
Secondary model (Corsi-HTKS)	0.312	0.337
Secondary model (HTKS-HTKS)	0.429	0.448
Secondary model (Corsi-Corsi)	0.315	0.349

Interaction EYPP model

Appendix table 11: EYPP interaction model raw regression outputs

Outcome	Variable	Raw coefficient	Standard error	95% confidence interval	p-value
EYTN Numeracy Test	Treatment	-0.183	1.005	-2.162 - 1.801	0.859
	EYPP status	-3.914	1.076	-6.015 - -1.728	<0.001
	Treatment EYPP interaction	1.714	1.533	-1.353 - 4.723	0.277

Appendix table 12: EYPP interaction model—overall effect size

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group		Total n (intervention; control)	Pooled Hedges g (95% CI)	p-value
	N (missing)	Pooled Mean (95% CI)	n (missing)	Pooled Mean (95% CI)			

EYTN Numeracy Test	838 (36)	32.58 (31.54 - 33.62)	851 (50)	32.8 (31.72 - 33.87)	1689 (838;851)	0.042 (-0.1 -- 0.18)	0.86
---------------------------	----------	-----------------------	----------	----------------------	----------------	----------------------	------

Appendix table 13: Additional impact of The ONE on EYPP eligible children

	Effect size		
	Total n (intervention; control)	Estimate	p-value
EYTN Numeracy Test	1689 (838;851)	0.11 (-0.086-0.302)	0.277

Appendix D: Outcome Distributions

Figure 4: Baseline distribution of early numeracy outcome in analytical sample by treatment arm

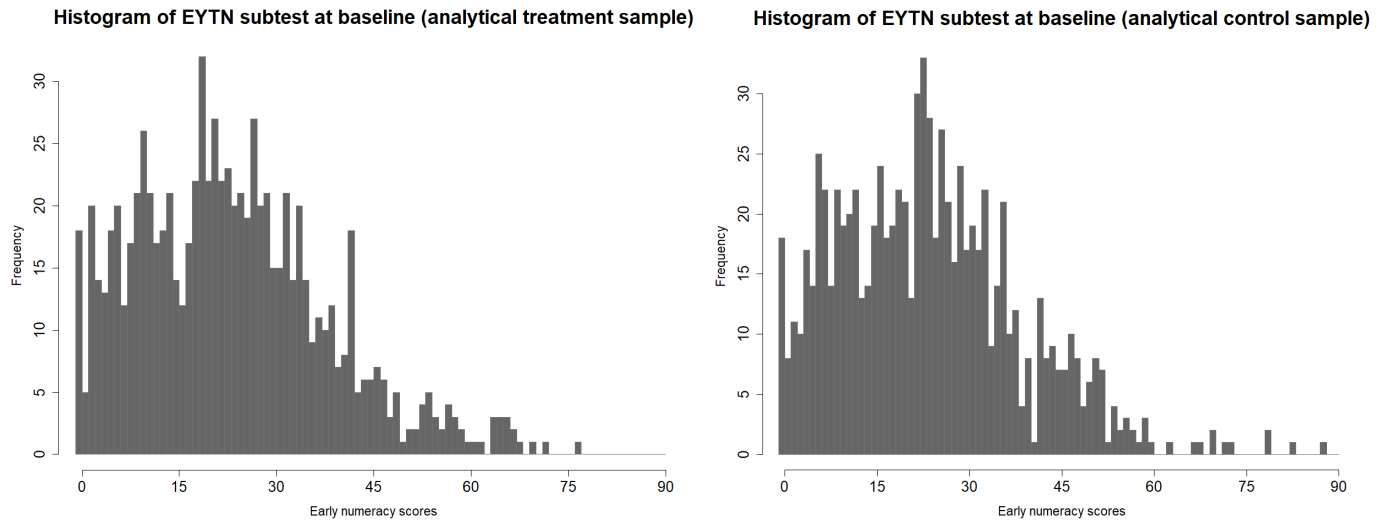


Figure 5: Endline distribution of early numeracy outcome in analytical sample by treatment arm

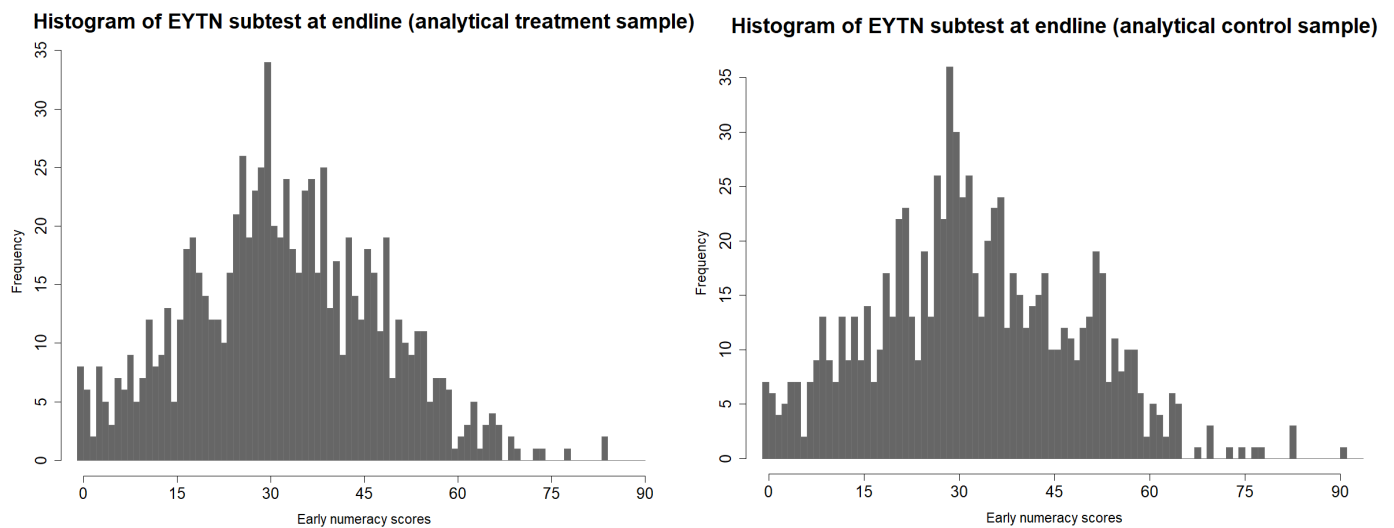


Figure 6: Baseline distribution of early numeracy outcome in EYPP subsample by treatment arm

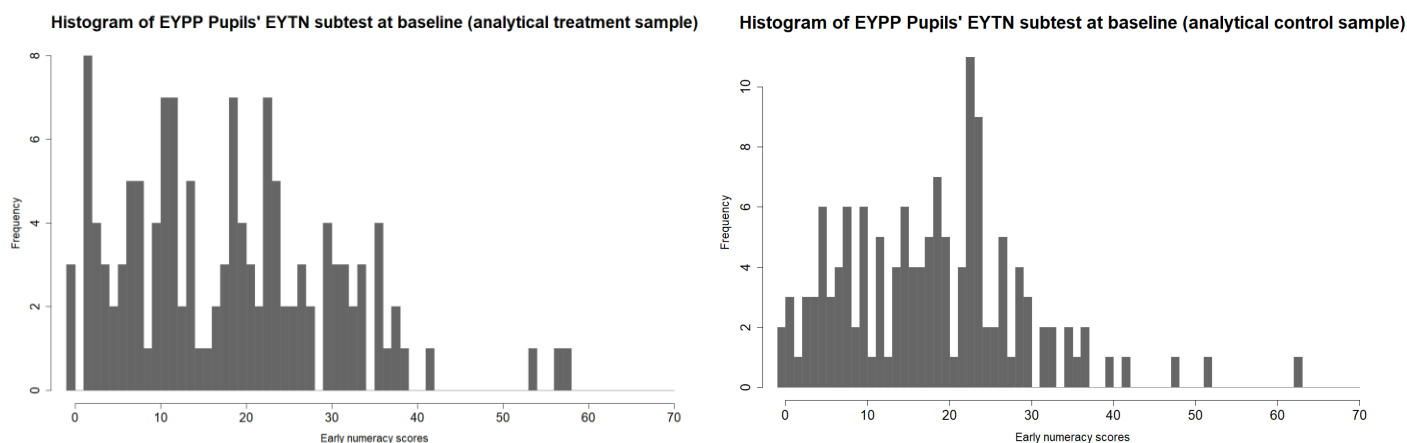


Figure 7: Endline distribution of early numeracy outcome in EYPP subsample by treatment arm

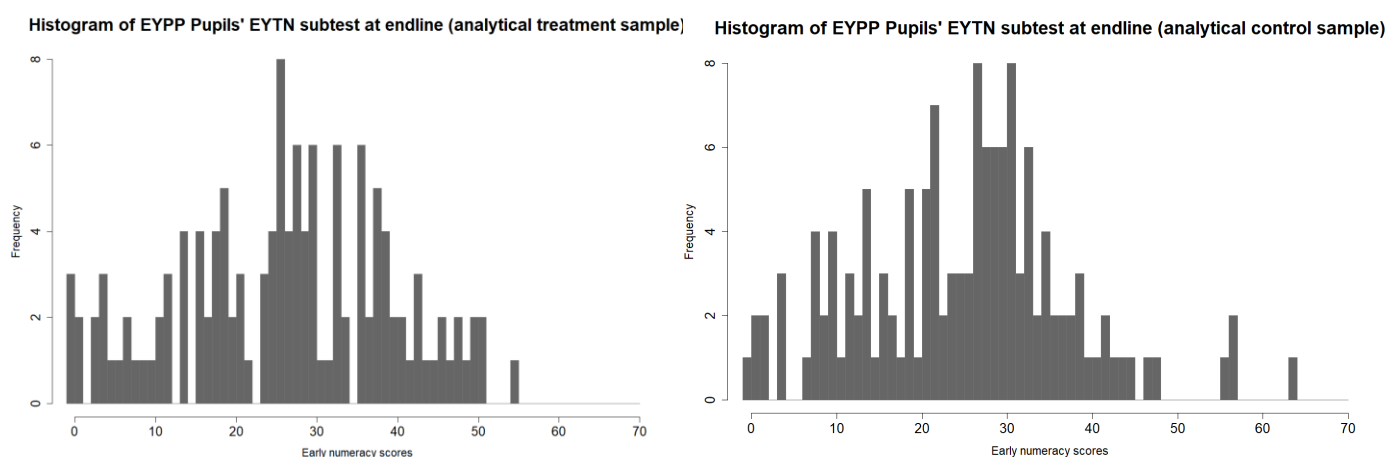


Figure 8: Baseline distribution of HTKS outcome in analytical sample by treatment arm

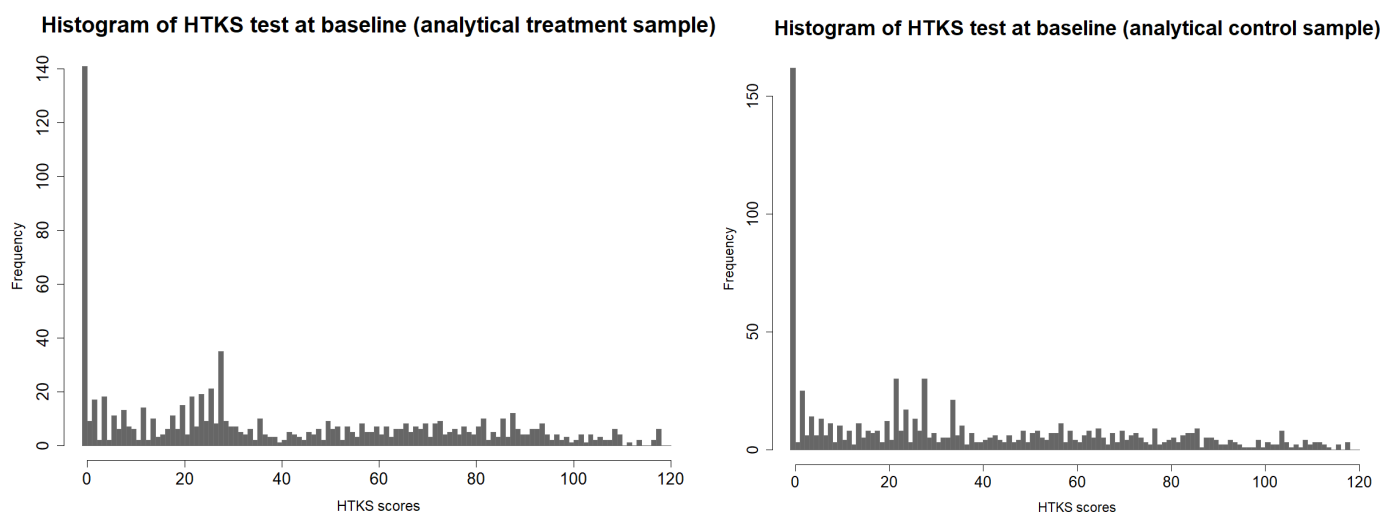


Figure 9: Endline distribution of HTKS outcome in analytical sample by treatment arm

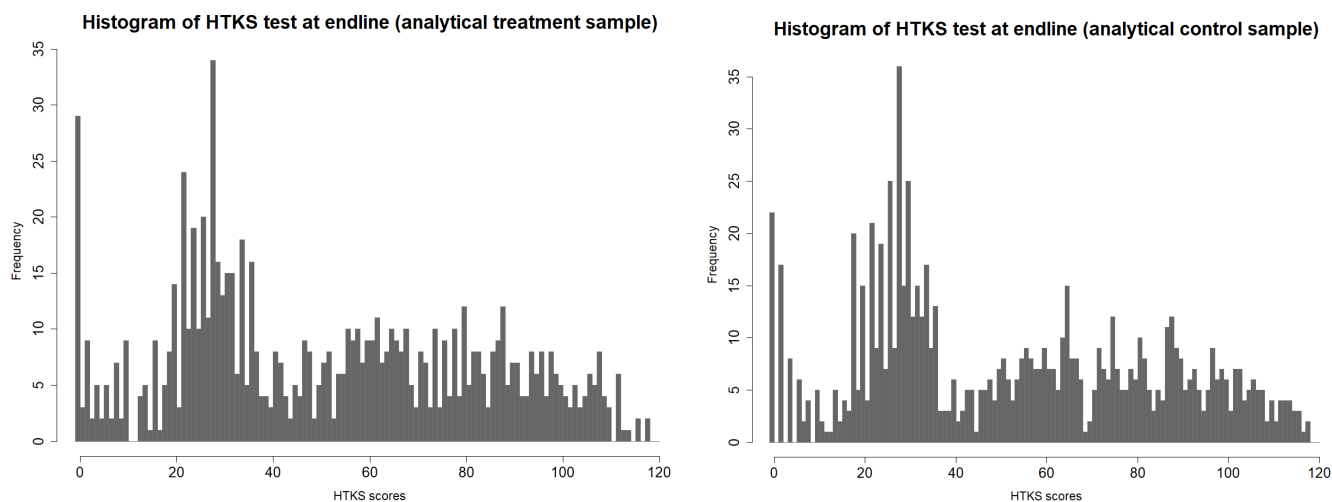


Figure 10: Baseline distribution of Corsi Blocks outcome in analytical sample by treatment arm

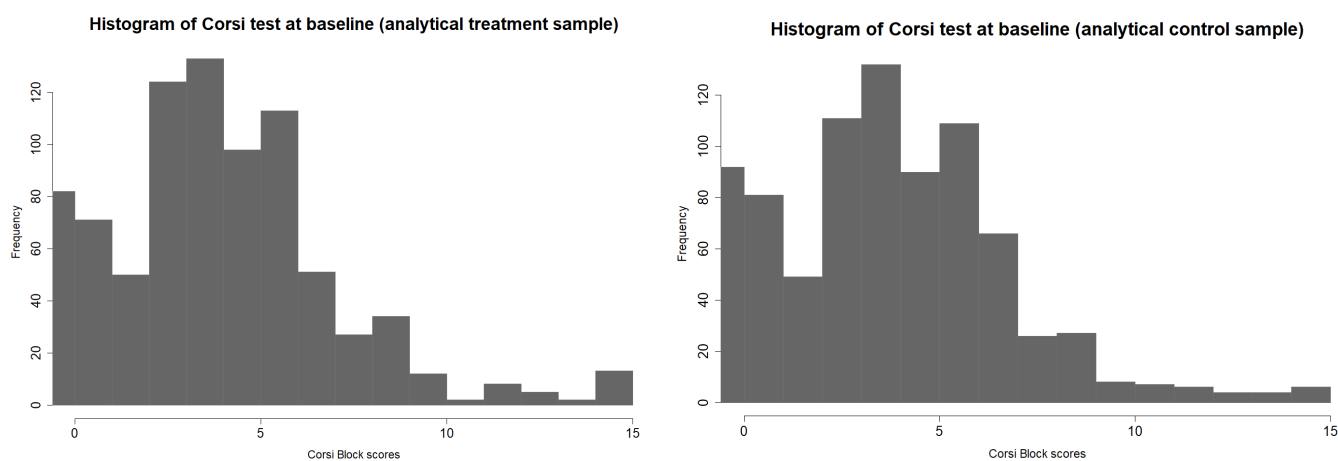
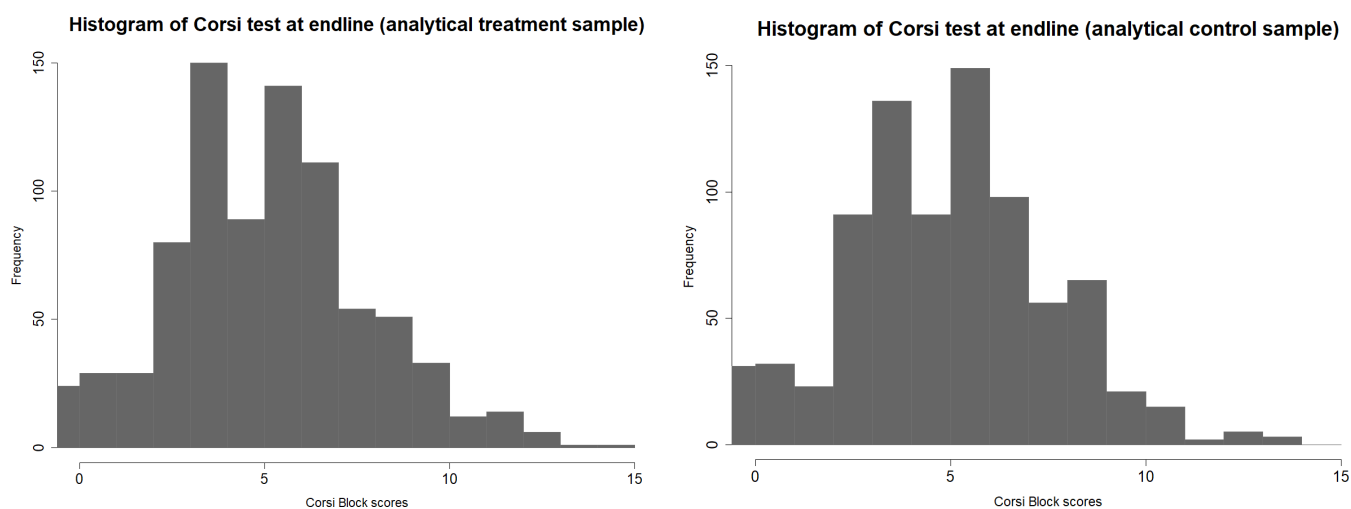


Figure 11: Endline distribution of Corsi Blocks outcome in analytical sample by treatment arm



Appendix E: Residual plots from analysis models

Figure 12: Primary analysis residual density plot (EYTN)

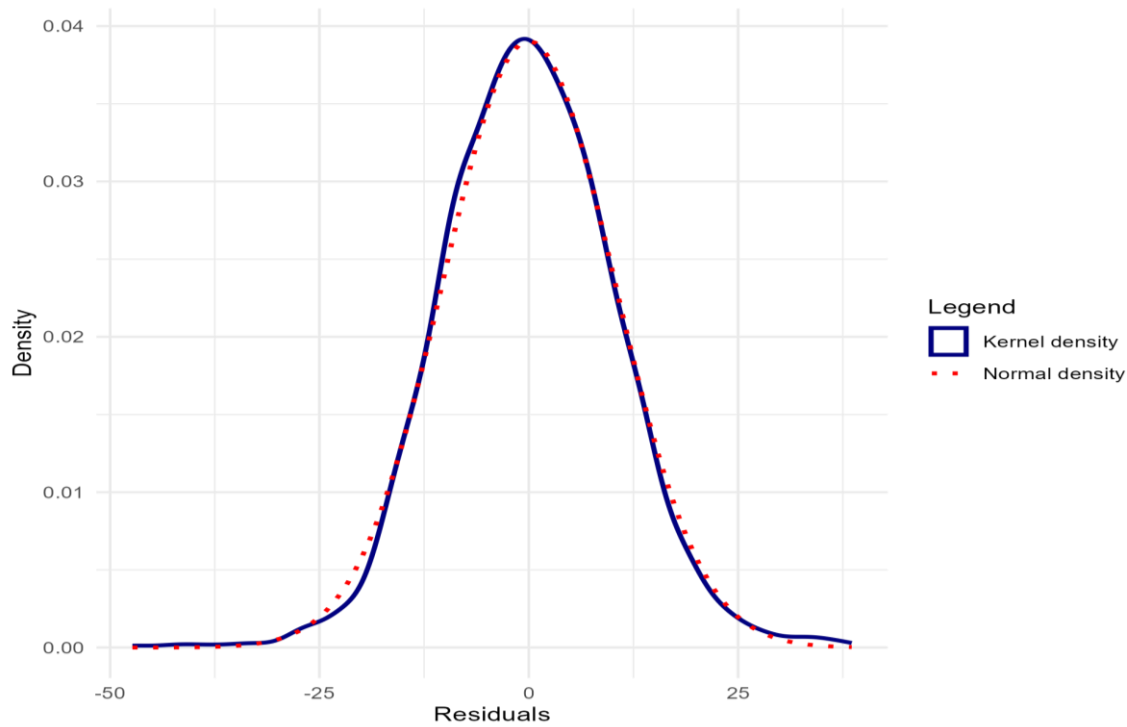


Figure 13: Primary analysis residual Q-Q plot (EYTN)

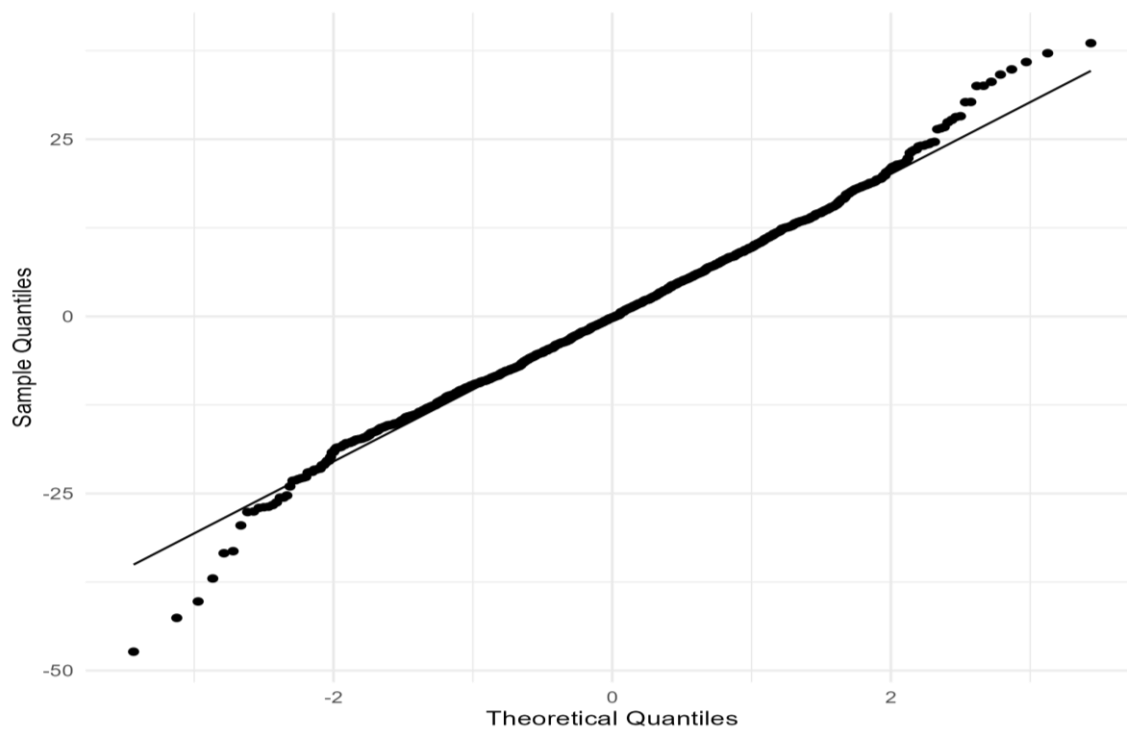


Figure 14: Secondary analysis residual density plot (mixed measures model)

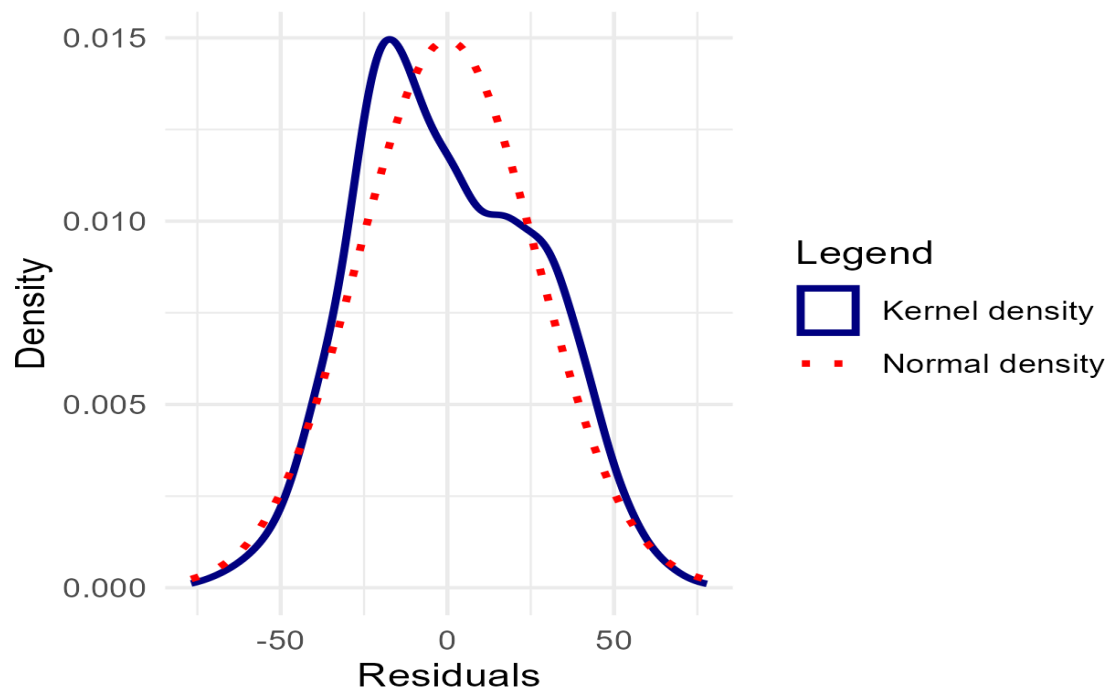


Figure 15: Secondary analysis residual Q-Q plot (mixed measures model)

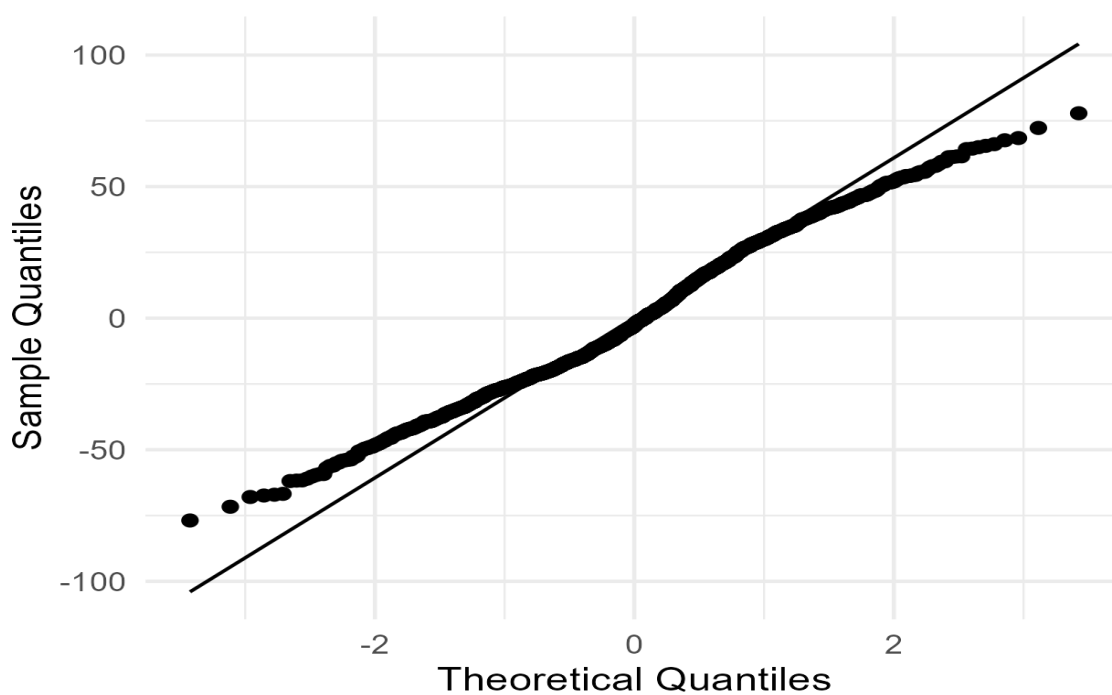


Figure 16: Secondary analysis residual density plot (HTKS)

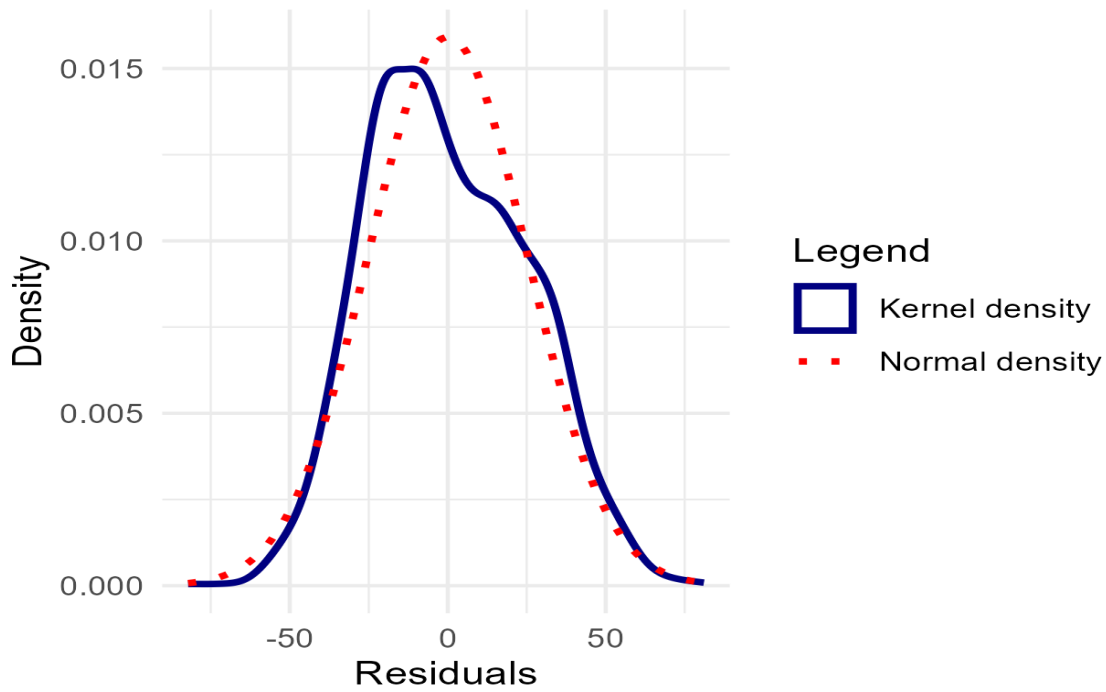


Figure 17: Secondary analysis residual Q-Q plot (HTKS)

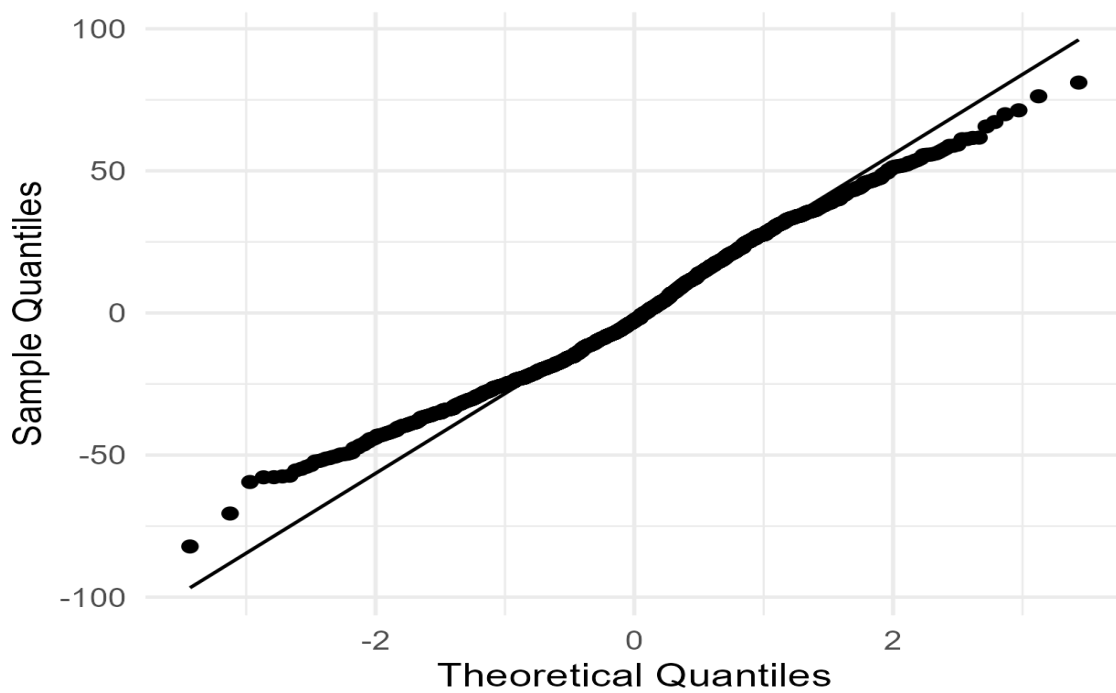


Figure 18: Secondary analysis residual density plot (Corsi Bocks)

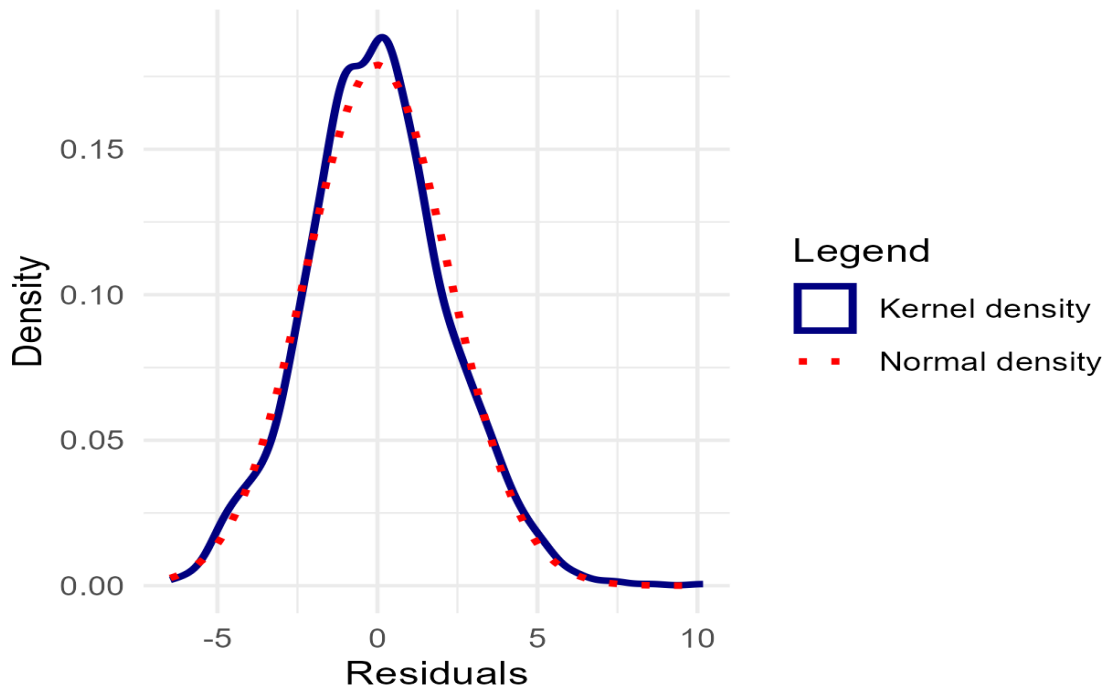


Figure 19: Secondary analysis residual Q-Q plot (Corsi Blocks)

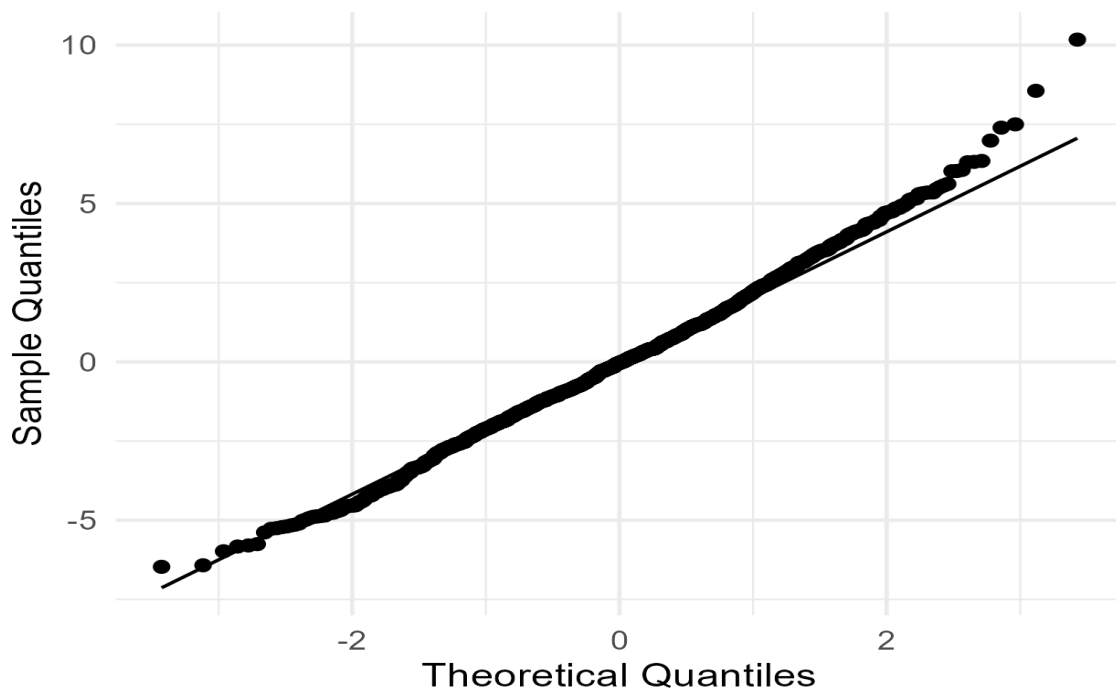


Figure 20: CACE analysis residual density plot (binary compliance)

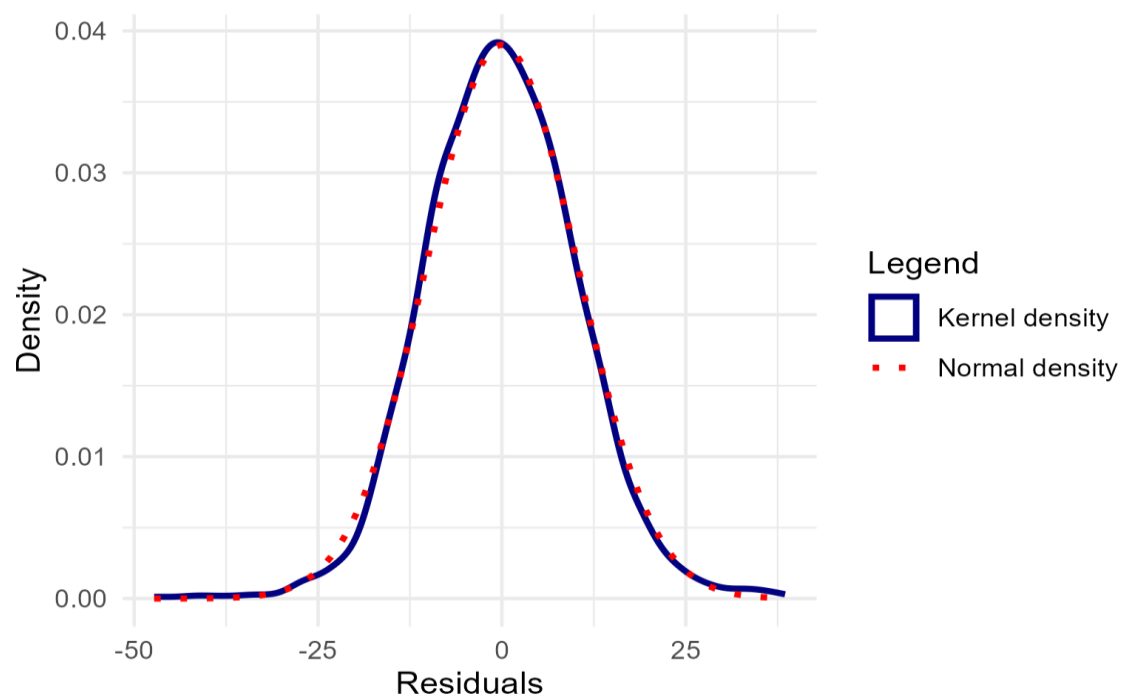


Figure 21: CACE analysis residual Q-Q plot (binary compliance)

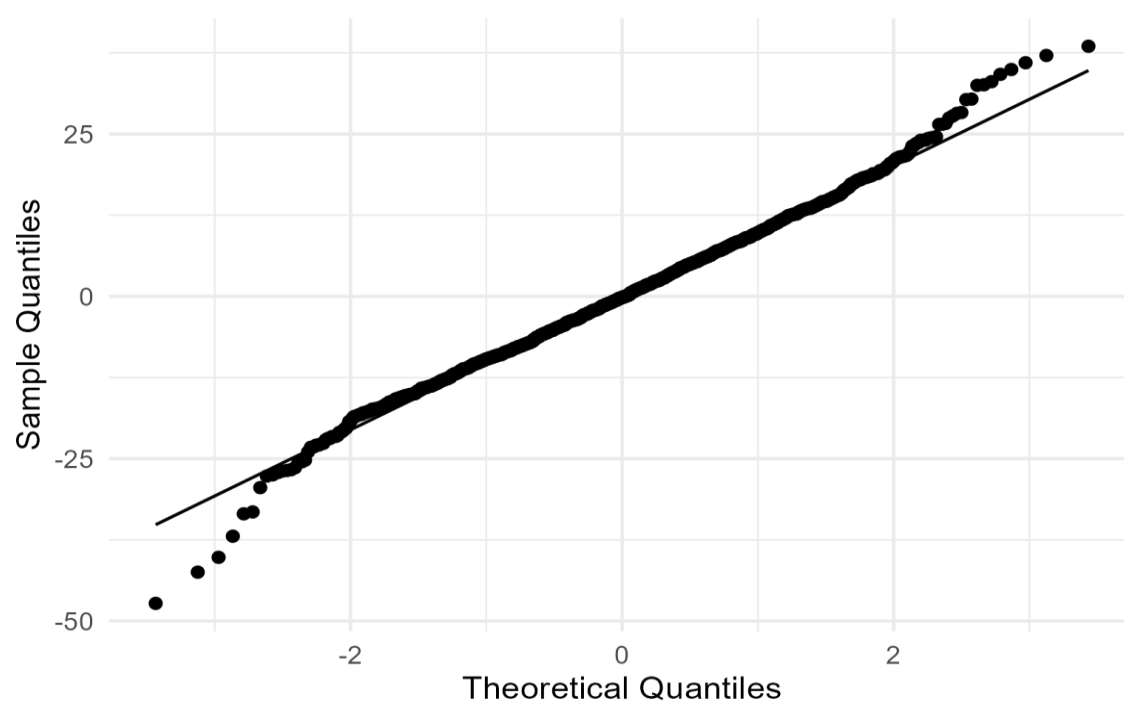


Figure 22: CACE analysis residual density plot (setting-level dosage)

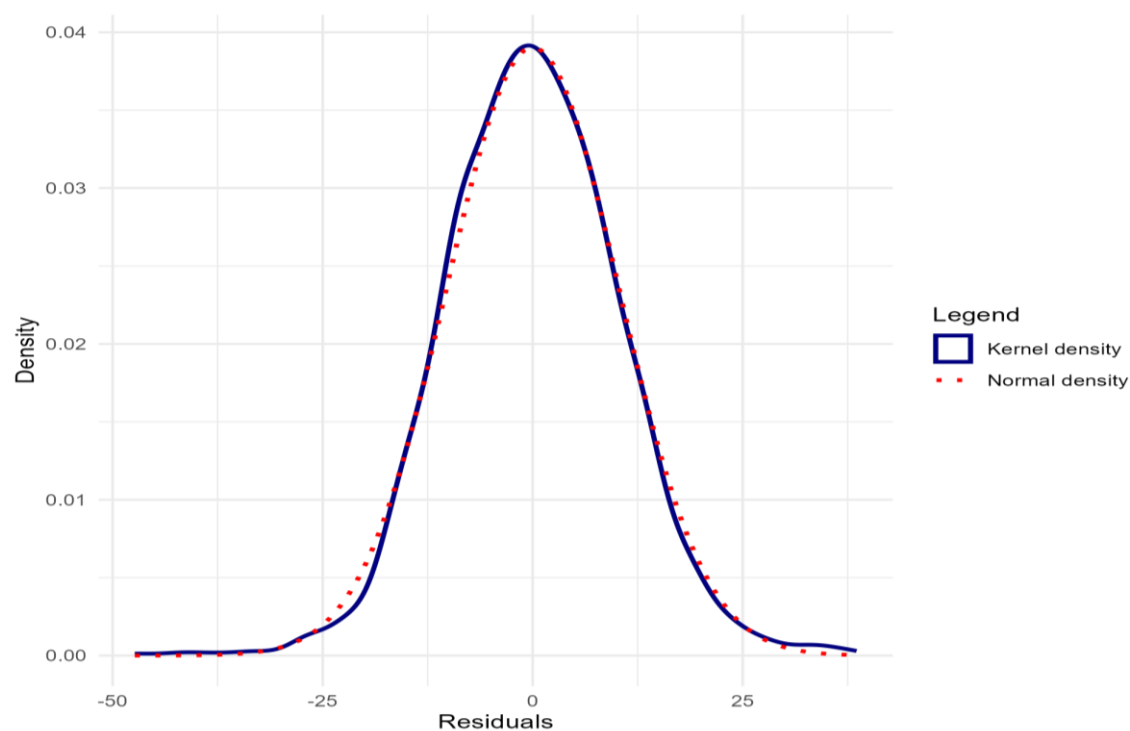


Figure 23: CACE analysis residual Q-Q plot (setting-level dosage)

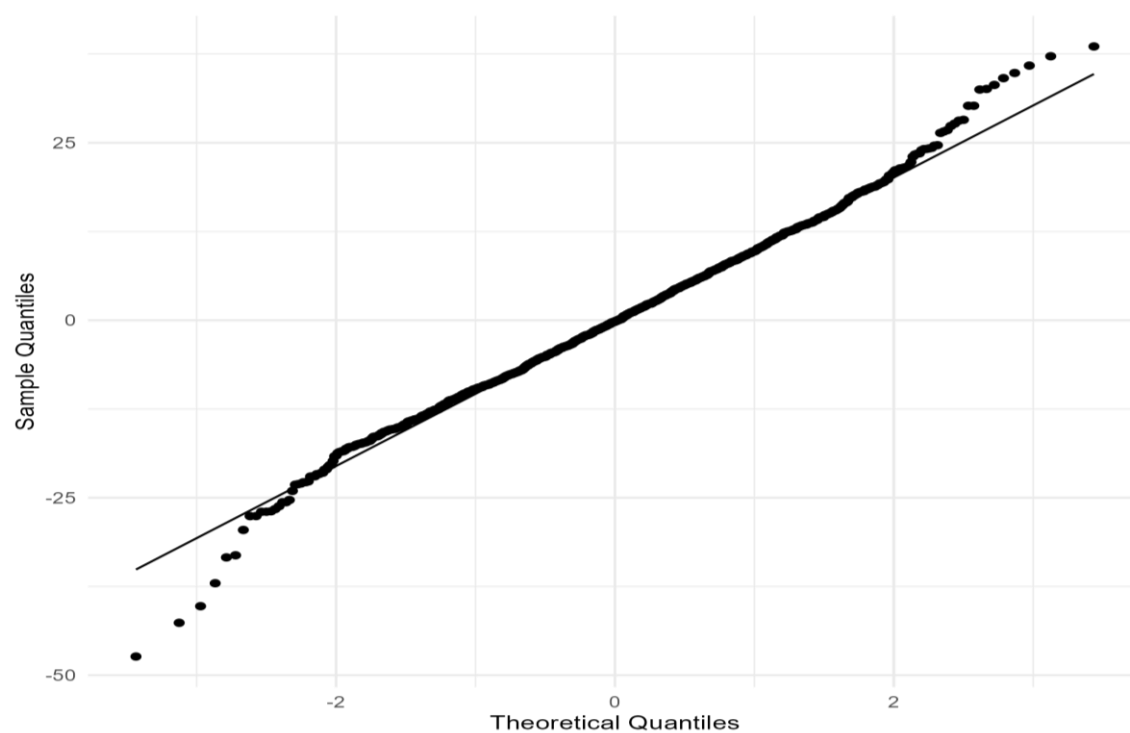


Figure 24: CACE analysis residual density plot (child-level dosage)

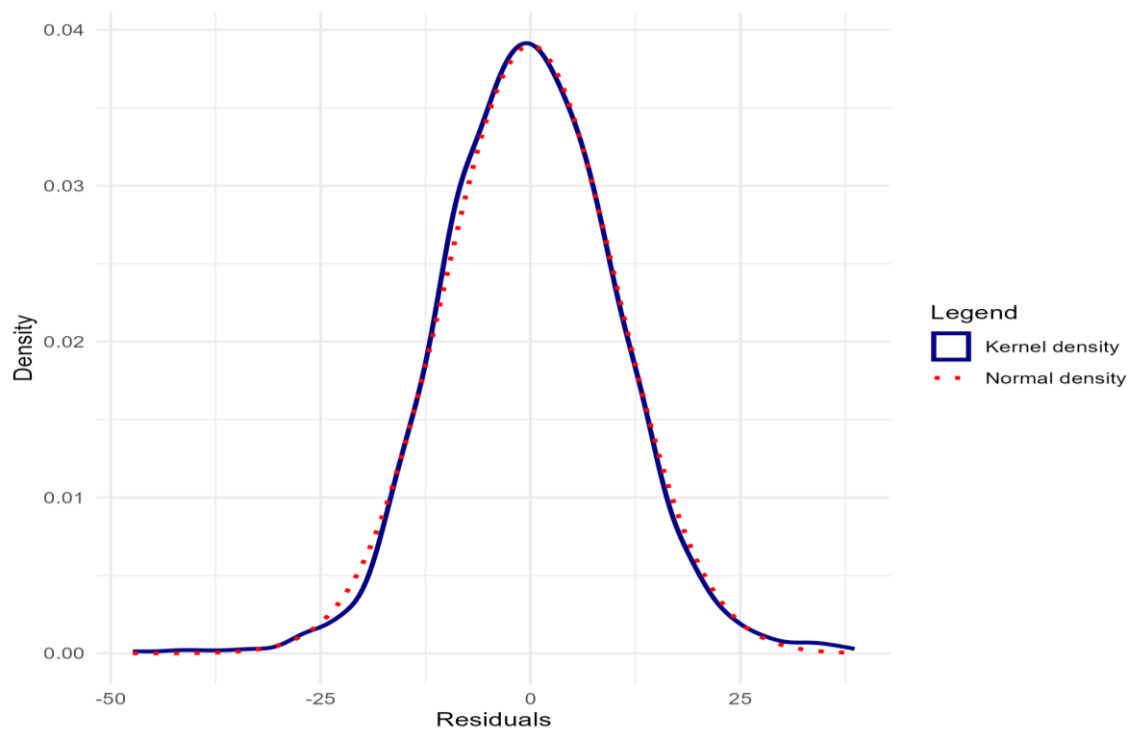


Figure 25: CACE analysis residual Q-Q plot (child-level dosage)

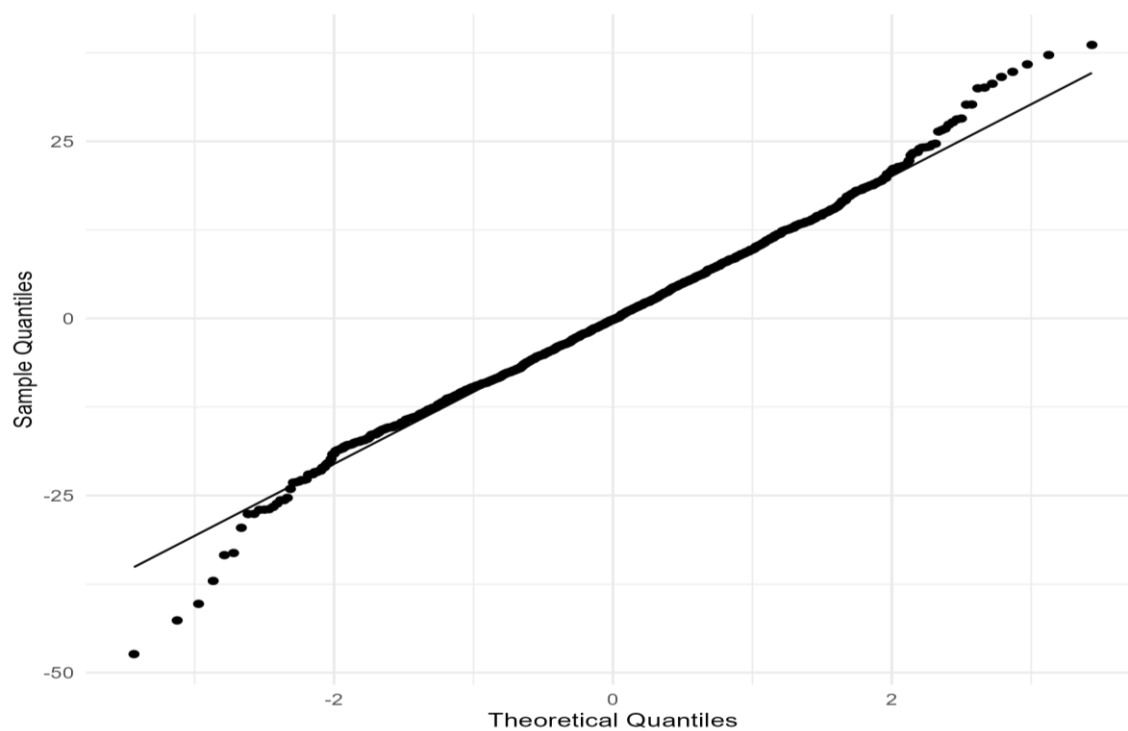


Figure 24: EYPP interaction model residual density plot (EYTN)

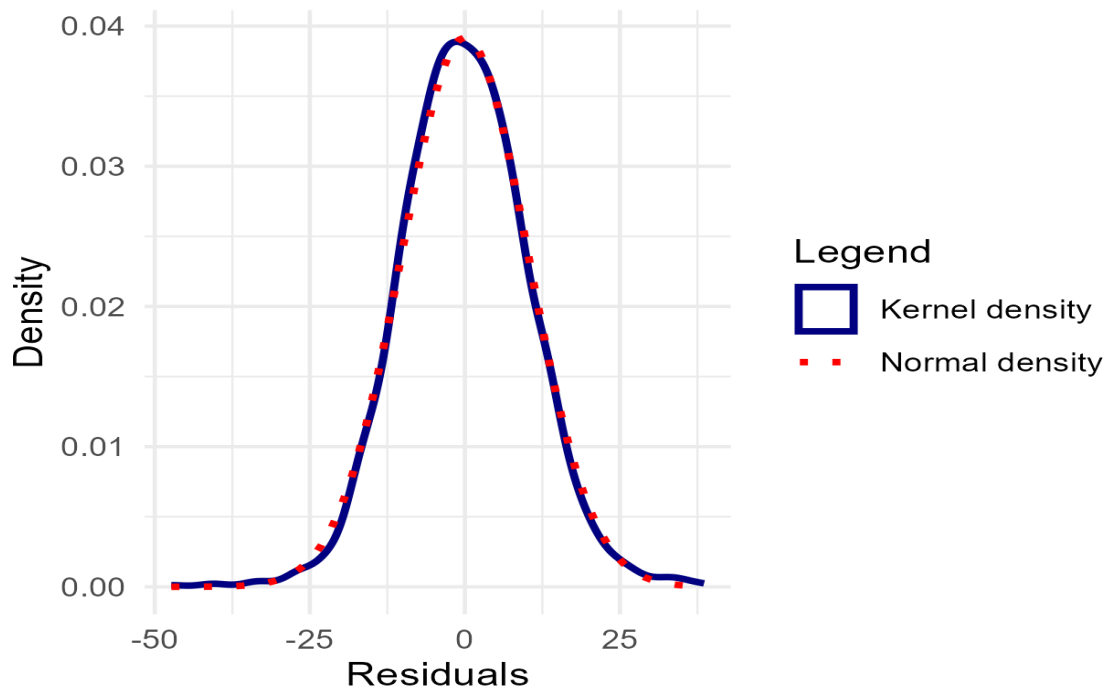


Figure 25: EYPP interaction model residual Q-Q plot (EYTN)

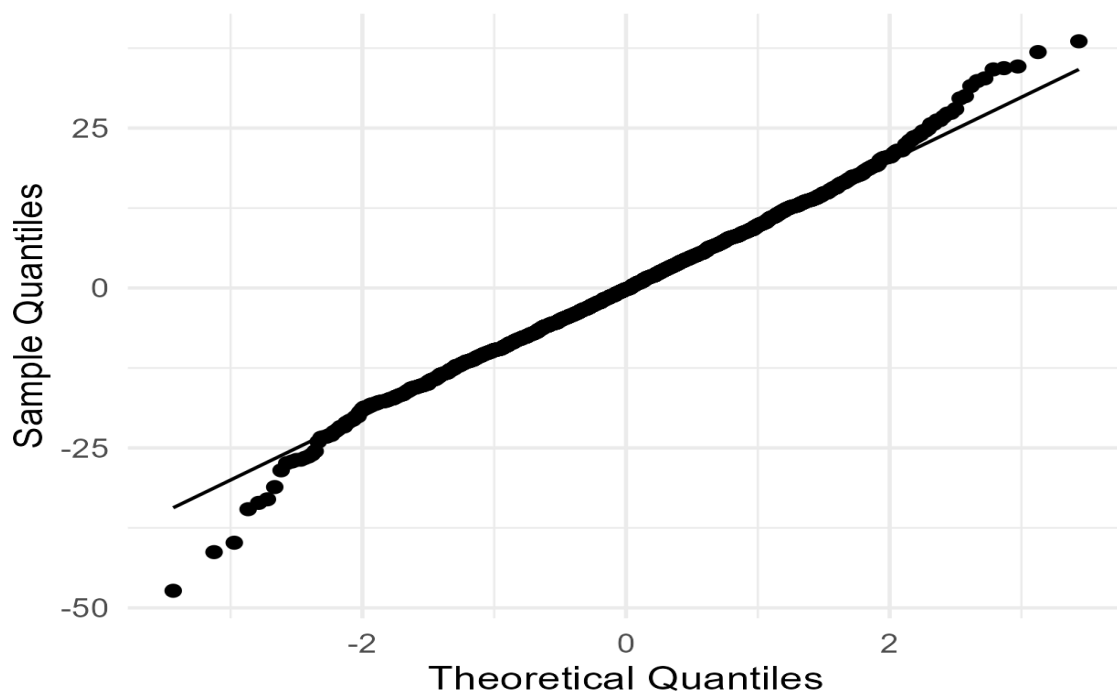


Figure 26: EYPP sub-sample model residual density plot (EYTN)

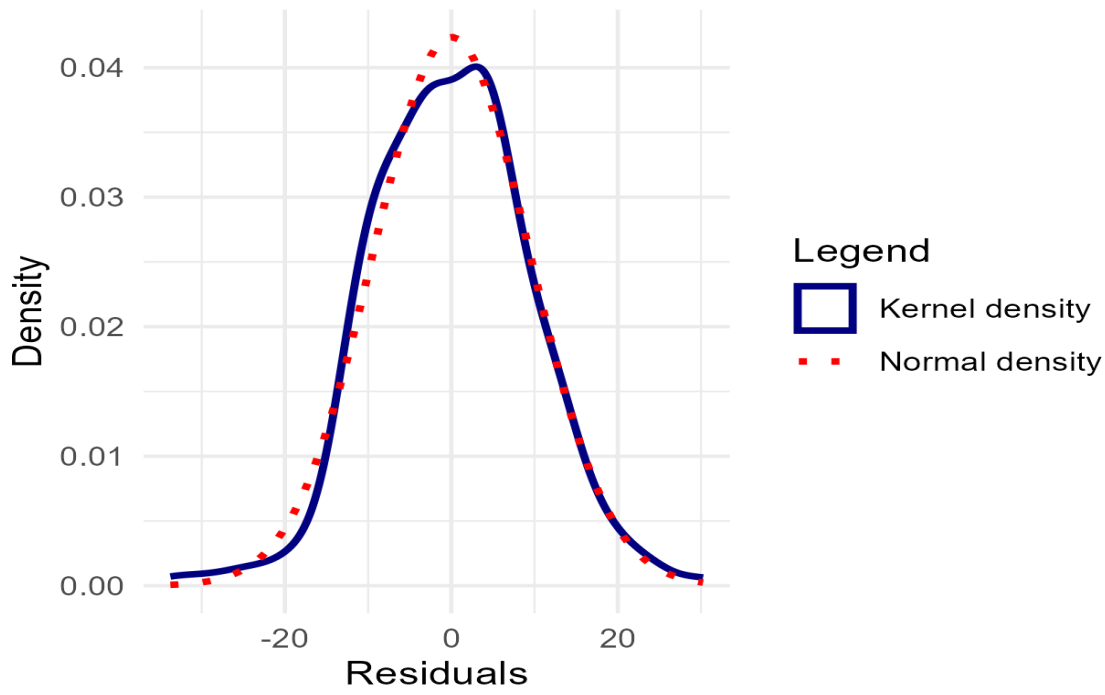


Figure 27: EYPP sub-sample model residual Q-Q plot (EYTN)

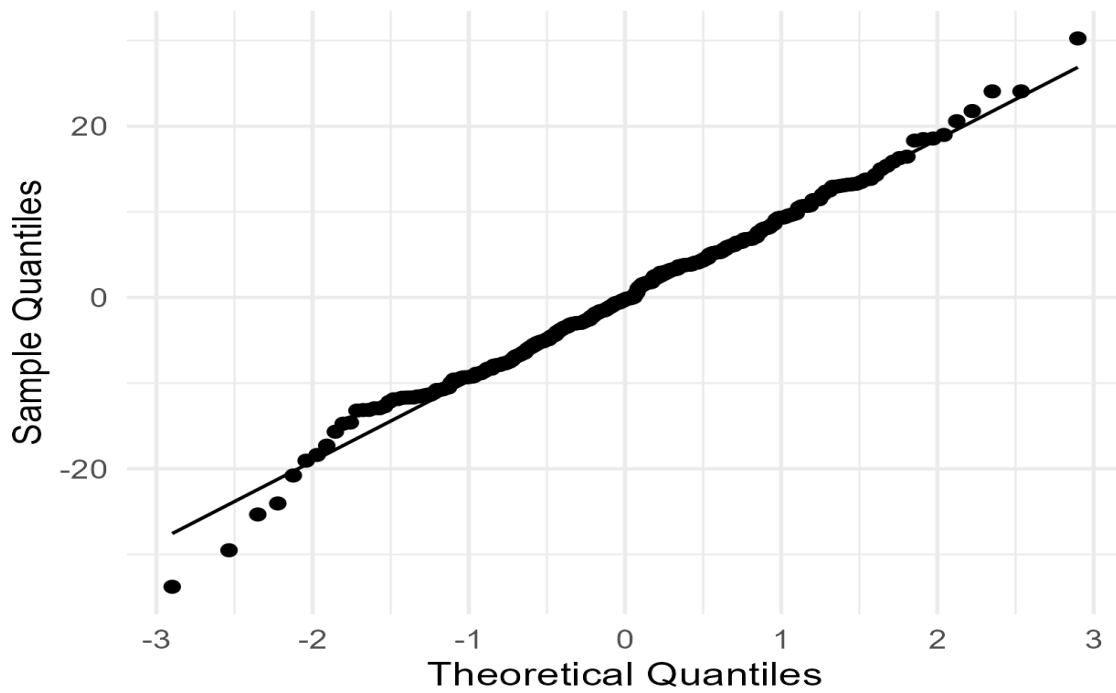


Figure 26: Sensitivity analysis residual density plot, excluding unblinded setting (EYTN)

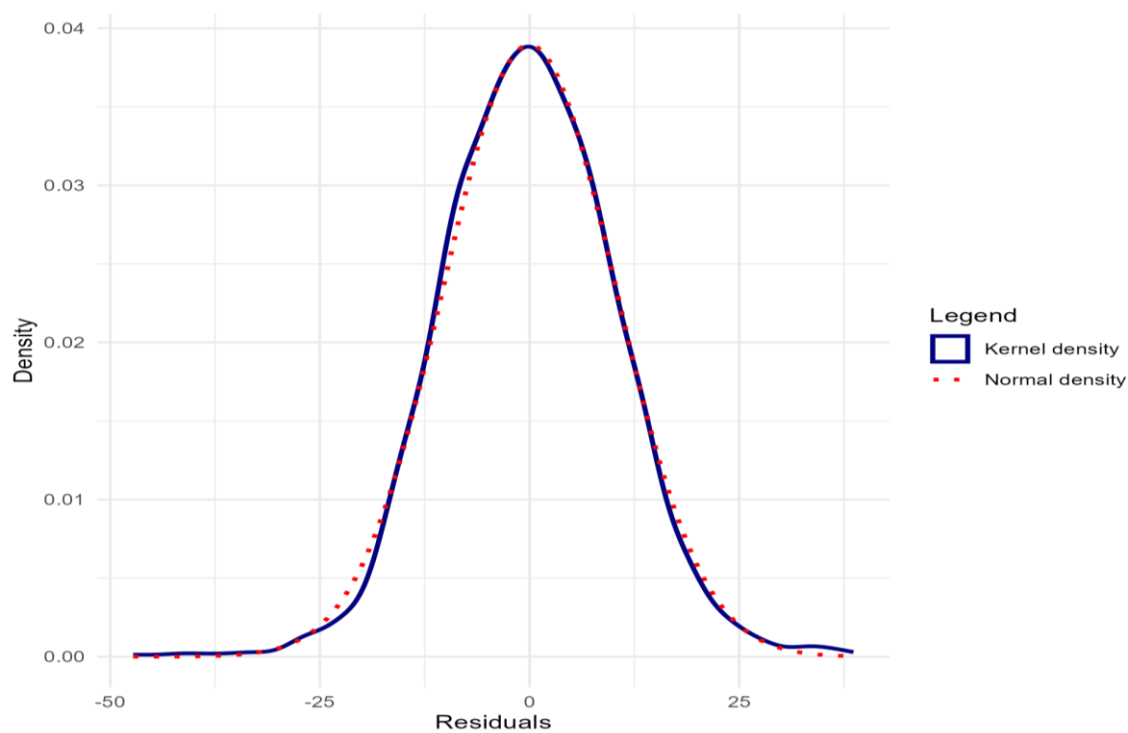


Figure 27: Sensitivity analysis residual Q-Q plot, excluding unblinded setting (EYTN)

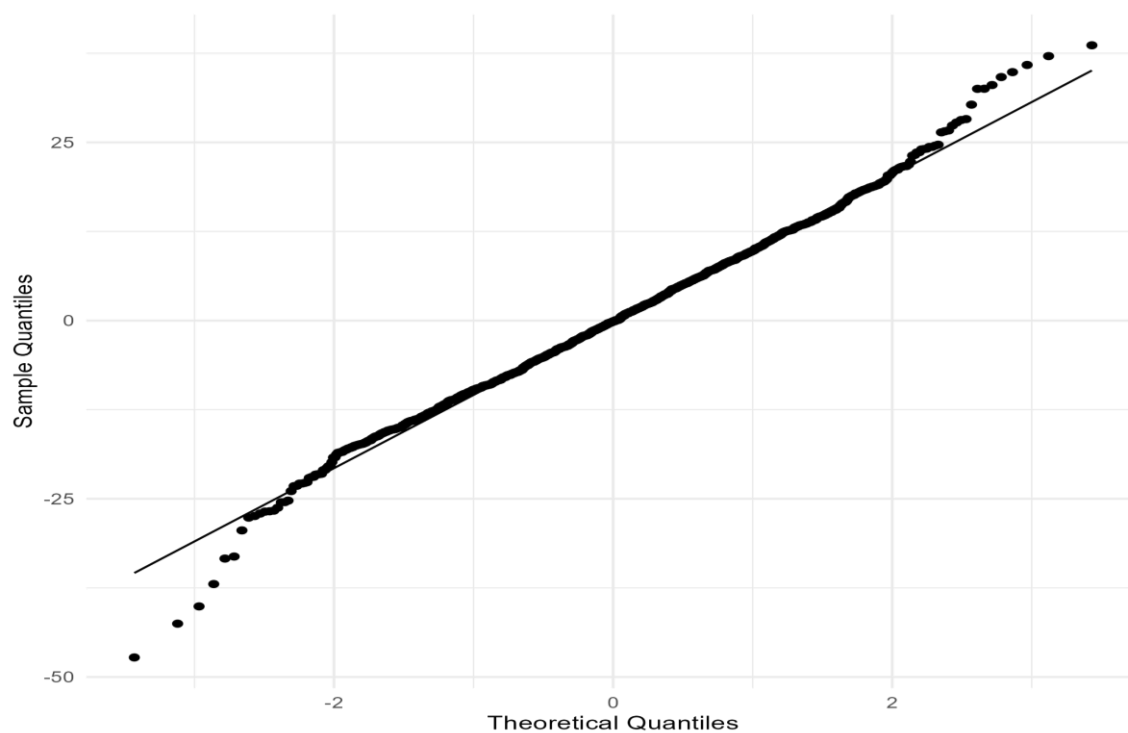


Figure 28: Sensitivity analysis residual density plot, primary analysis adjusted for age (EYTN)

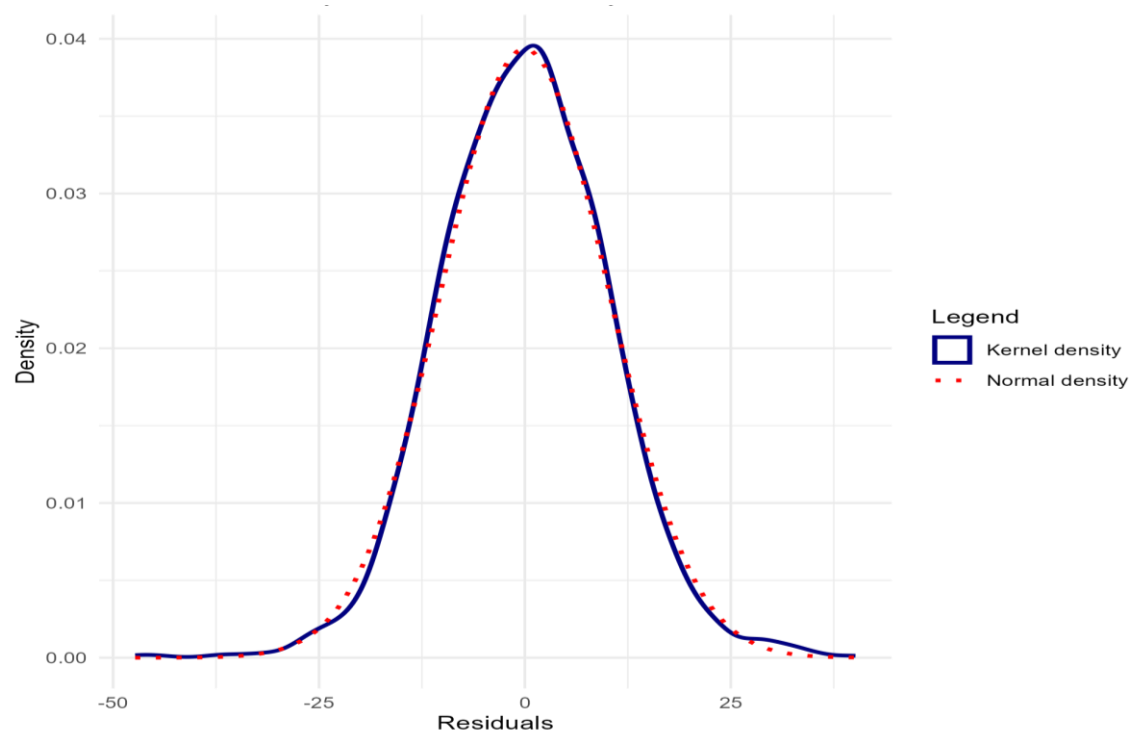


Figure 29: Sensitivity analysis residual density plot, primary analysis adjusted for age (EYTN):

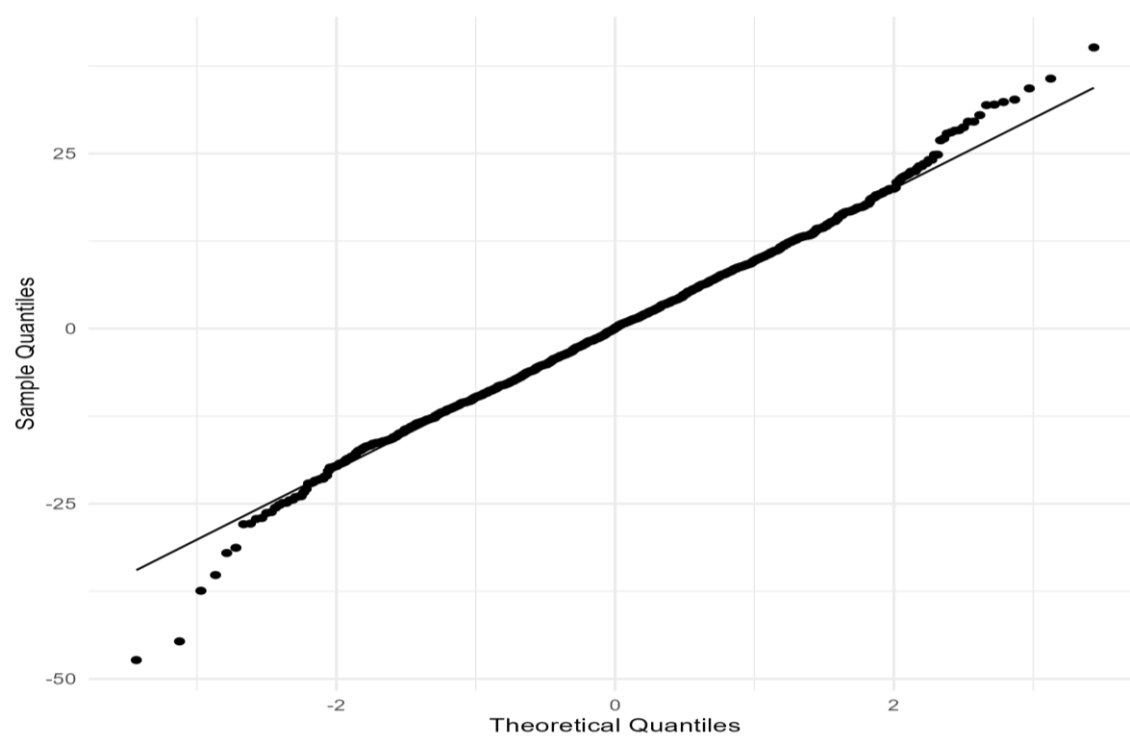


Figure 28: Sensitivity analysis residual density plot, primary analysis excluding Maths Champions participating settings (EYTN)

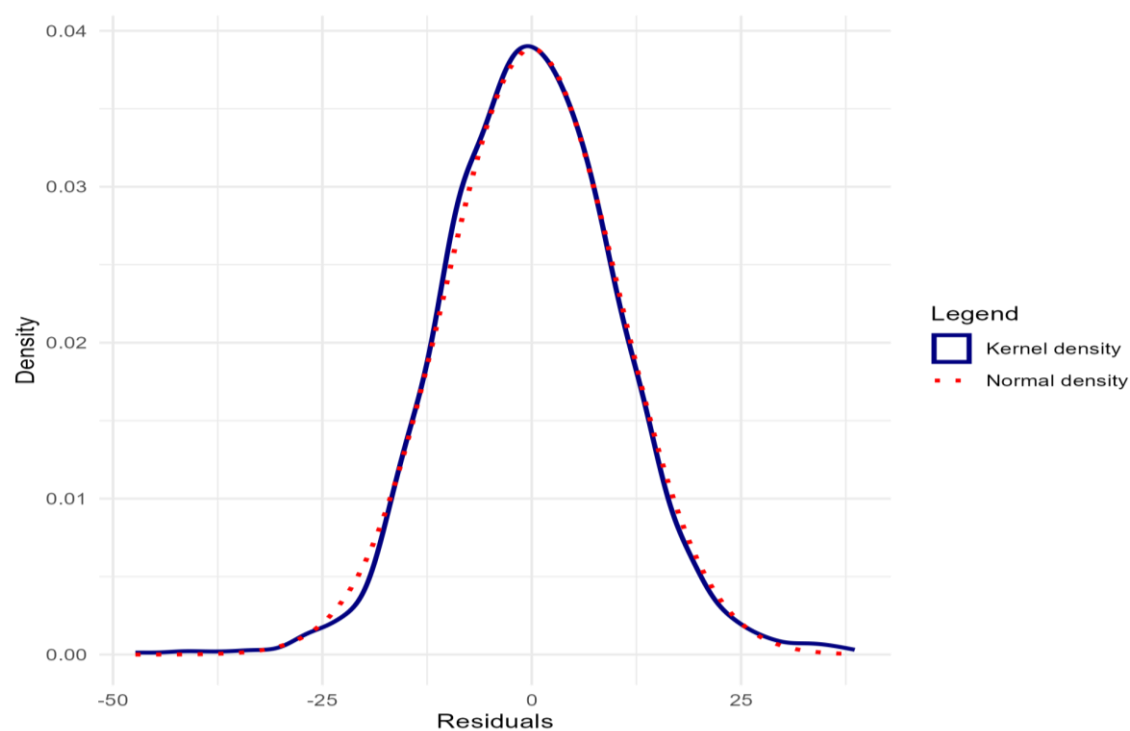


Figure 29: Sensitivity analysis residual Q-Q plot, primary analysis excluding Maths Champions participating settings (EYTN)

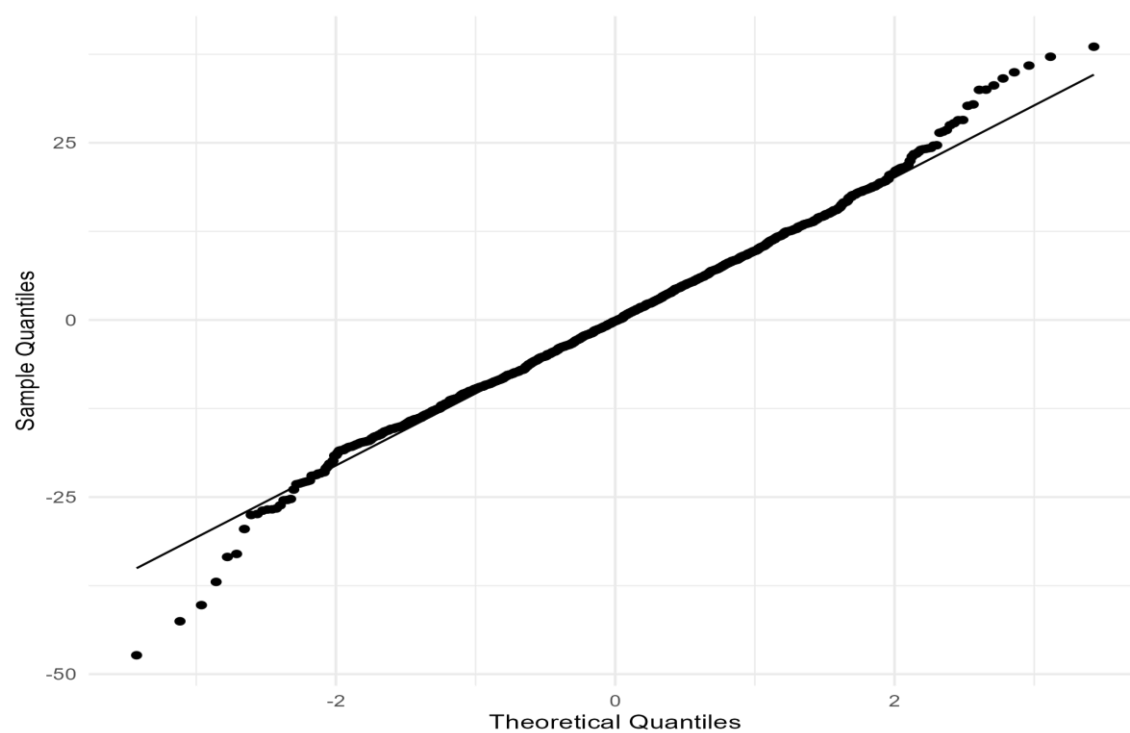


Figure 30: Distribution of activities done in treatment settings (setting dosage measure)

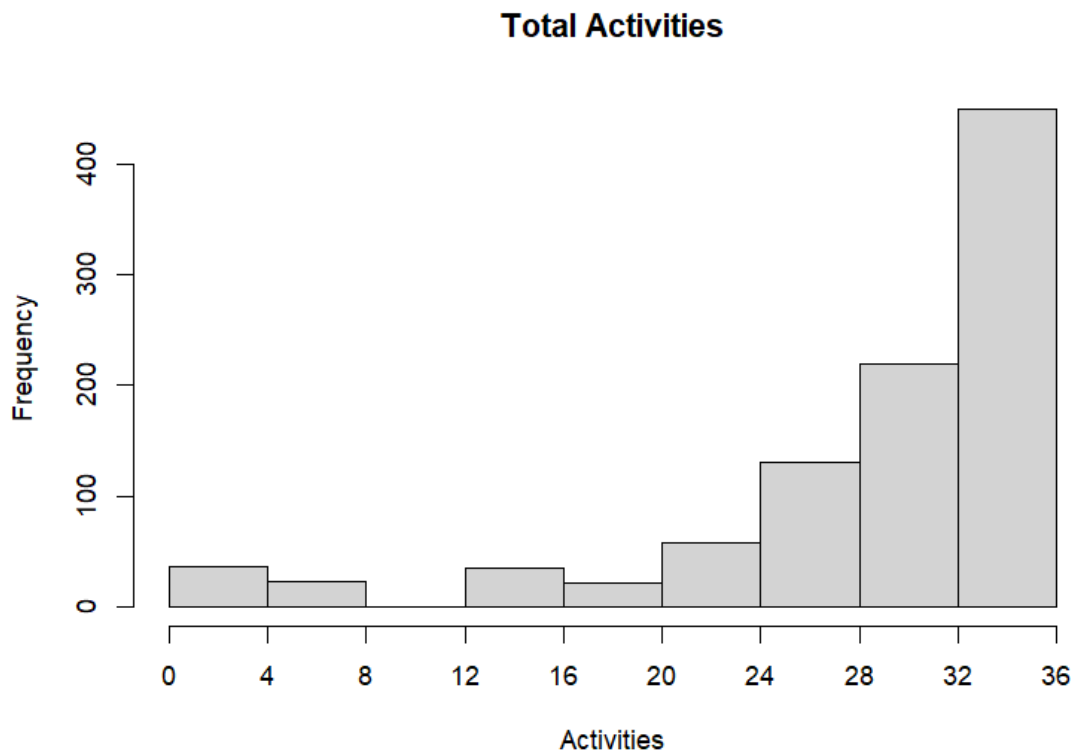
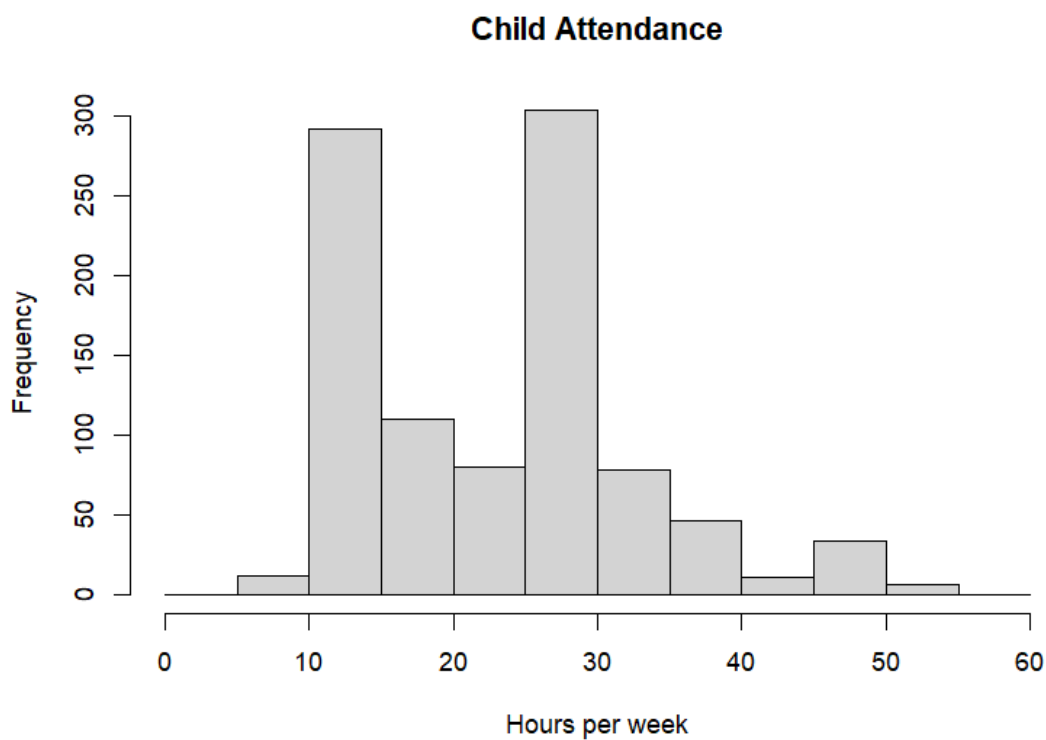


Figure 31: Distribution of hours attending setting of children in treatment settings (child dosage measure)



Further appendices:

You can find the further appendices as a separate document published on the project page.

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

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


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

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



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)