Project Title: Orchestrating Numeracy and the Executive (the ONE)
Statistical Analysis Plan

Evaluator: RAND Europe

Principal investigator(s): Elena Rosa Speciani

Template last updated: August 2019

PROJECT TITLE[1]	Orchestrating Numeracy and the Executive (the ONE) (formerly, Embedding Executive Challenge into Early Mathematics)
DEVELOPER (INSTITUTION)	University of Oxford
EVALUATOR (INSTITUTION)	RAND Europe
PRINCIPAL INVESTIGATOR(S)	Elena Rosa Speciani
SAP AUTHOR(S)	Elena Rosa Speciani, Merrilyn Groom, Rachel Hesketh, James Merewood
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at setting level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Ages 3-4 (pre-reception)
NUMBER OF SCHOOLS	150
NUMBER OF PUPILS	1859
PRIMARY OUTCOME MEASURE AND SOURCE	The Early Years Toolbox Early Numeracy (Howard et al., 2022)
SECONDARY OUTCOME MEASURE AND SOURCE	Measures of executive function: Heads Toes Knees Shoulders (HTKS-R) (Gonzales et al., 2021), Corsi blocks (Richardson, 2007).[2]

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]		*N/A*

[1] Make sure that the project title here matches the title of the document and the protocol. Please ensure that there is an identification as a randomised trial in the title as per CONSORT requirements.

[2] As one measure is composite and the other is domain-specific, they will not be combined to form a single measure of EF.

Table of contents

Introduction

The trial is being conducted to address the critical connection between early mathematics achievement and executive functions (EF) (Coolen et al., 2021). Research indicates that early mathematics skills are predictive of later performance (Verdine et al., 2014), and children who lag in mathematics often continue to do so (Purpurpa & Lonigan, 2015). EF, encompassing cognitive processes like working memory and attentional control, has been shown to correlate with mathematics skills in young children, especially in socio-economically disadvantaged groups (Blair & Raver, 2014). This suggests that improving EF could help narrow the attainment gap in mathematics.

Prior trials carried out outside of the UK have shown that integrating executive challenges into play-based activities can improve the EF of children aged 3 – 5 years, but the intervention did not significantly impact academic attainment (Howard et al., 2020). This project aims to adapt this approach to the UK Early Years context, incorporating mathematics-specific content that has been co-developed with teachers, given the evidence that executive functions are a pillar for mathematical development in early ages.

The current evaluation is structured as a two-armed, randomised waitlisted controlled trial, allowing all participating settings to eventually receive the intervention. The intervention lasts for 12 weeks. Settings in the treatment group will receive the intervention between January and May 2024,. Waitlisted settings will receive the intervention from September 2024. All settings were baseline tested in the autumn term (October – December 2023) with endline testing occurring in the summer term (April-June 2024), with all testing undertaken by assessors at Qa Research. The evaluation will measure mathematics attainment as the primary outcome and EFs as the secondary outcome.

Intervention

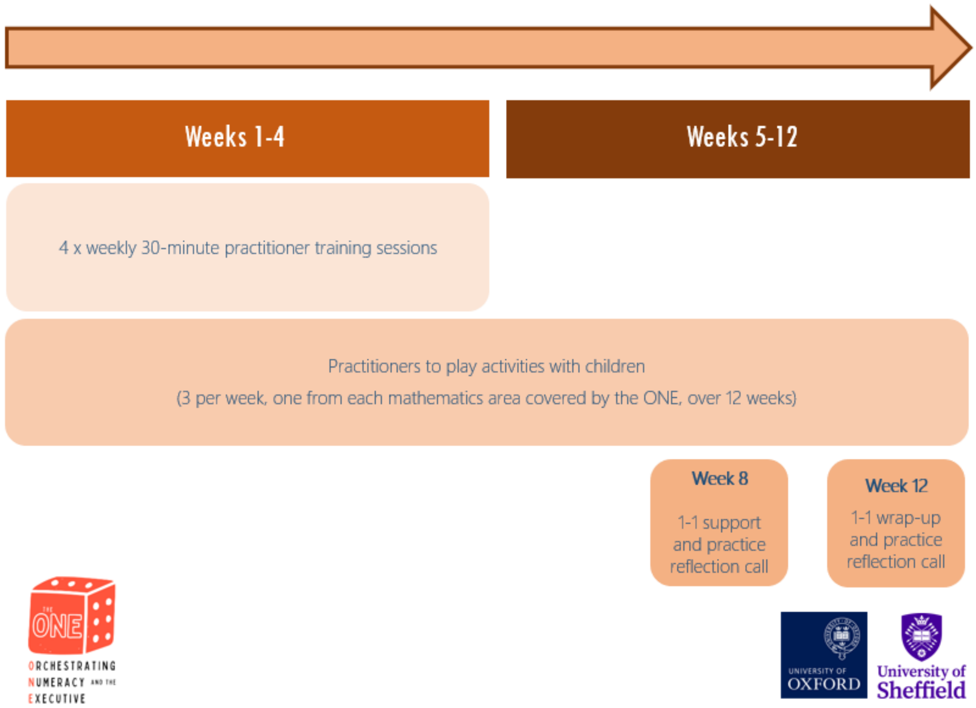
The delivery team from the University of Oxford and University of Sheffield will provide training and support to Early Years practitioners to implement play-based mathematics activities that enhance mathematics development by integrating executive functioning skills into mathematics learning. The program includes face-to-face training for educators, 25 activity cards, and resources for the activities. All participating nursery staff undergo a training program consisting of four weekly 30-minute professional development sessions for the first four weeks, with sessions spaced a week apart to provide setting staff time for self-reflection on the delivery of activities. These sessions introduce the activity cards, cater to each setting's preferences (e.g., 1-to-1, or in a group) and focus on building educators' understanding of co-development of early mathematics and executive functions (EF). The aim is to enhance practitioners' skill in conducting play-based activities that integrate executive function into mathematics learning. All professional development sessions take place within the setting.

From Week 1 to Week 12 of the programme, setting practitioners will introduce this play-based approach to children aged 3-4 years who are set to start school in the subsequent academic year. Practitioners will receive 25 activity cards covering three key areas of early years mathematics (numbers and counting; ordering and patterns; shapes, and spatial awareness), informed by expert input from the extended delivery team and educators. Each card highlights the mathematical and executive function skills it fosters and how to increase executive function demands. Some activities may be familiar, while others extend the range of mathematics skills educators can support. The activities utilise commonly available resources and a low-cost resource pack, aiming to scaffold children's mathematics learning with optimal executive challenge. Setting practitioners deliver all activities within their early years settings, during the settings' normal routines; the activities are such that they can be suitably adapted to a manner of different early years approaches to play and instruction, and the professional development provided supports this.

Practitioners are asked to implement at least three activities per week, including one from each area of mathematics. These 5–10-minute activities can be incorporated into nursery routines such as small group activities, larger group activities, outdoor play, and free play. As a whole-class intervention, these activities can be applied to the entire class or playroom. In larger settings, a specific classroom or playroom was selected, with all staff and children in that room participating in the intervention. Practitioners start running these activities from the first week of the intervention, alongside the four-week professional development program. They have the flexibility to choose the size of implementation group, as long as staff participating in professional development and children in the year preceding the move into Reception are included. During the first four weeks of the delivery period, practitioners were provided with a list of children randomly selected for baseline assessment, and encouraged to makethe activities available to these children on days they attend; however, the whole-class nature of the intervention means that other children are not excluded from participation.

The program additionally offers opportunities for practitioners to reflect on their implementation of the activities. After the initial four-week program, one representative from each setting will have one-to-one follow-up sessions with the delivery team in the eighth and twelfth weeks to receive support, check fidelity, and encourage reflection. These sessions will aim to promote conscious observation of children's engagement and adjustment of the level of embedded executive challenge as needed. The delivery timeline, inclusive of professional development, activity delivery and practitioner support, is outlined below in Figure 1.

Figure 1: The ONE delivery timeline


Weeks 1-4
Weeks 5-12
4 x weekly 30-minute practitioner training sessions
Practitioners to play activities with children
(3 per week, one from each mathematics area covered by the ONE, over 12 weeks)
Week 8
1-1 support and practice reflection call
Week 12
1-1 wrap-up and practice reflection call
THE ONE
ORCHESTRATING NUMERACY AND THE EXECUTIVE
UNIVERSITY OF OXFORD
University of Sheffield

A final refinement phase was conducted following the completion of the feasibility study (Scerif et al., 2023) to inform the implementation of the intervention in this trial. This phase specifically sought feedback from educators in settings serving low-income communities. Activity cards were revised to emphasize differentiation strategies for children with a lower initial knowledge of mathematics, special educational needs (SEND), or English as an Additional Language (EAL). Conceptual clarifications were also made to individual activity cards to help educators understand the key elements to retain and how to differentiate the activities. In the refinement phase prior to this efficacy evaluation, practitioners utilized some executive functions and mathematics adaptations in activities. Based on the feedback from the refinement phase, no further adaptations are planned for this intervention. Any further adaptations made during the course of the delivery for this evaluation will be documented.

The project is delivered within a larger initiative by the Department for Education's Stronger Practice Hubs (SPH),[3] which focuses on evidence-based development in early years education. The goal is to generate evidence on effective interventions in early years settings.

Participant Selection

Setting level

The trial was open to both maintained settings and private, voluntary, and independent (PVI) early years settings. The recruitment goal for SPH trials was to include at least 30% of settings from both the maintained and PVI sectors.

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

1. Sending out direct emails to all eligible educational establishments within the specified Local Authorities (LA), whose contact information is publicly accessible.

2. Engaging LA Early Years specialists to assist with recruitment.

3. Actively reaching out to educational establishments in low-income areas (identified by an Index of Multiple Deprivation (IMD) score of less than 5) to increase participation from settings eligible for Early Years Pupil Premium (EYPP).

4. Collaborating with Stronger Practice Hubs to extend recruitment efforts through their networks.

Settings could only take part in one Stronger Practice Hub programme and could not be involved in another trial that included the same children and the same outcomes of interest (i.e., 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 will take place in most settings during April and May 2024, delaying participation until June 2024 will mitigate the possible contamination effects of Maths Champions participation on the ONE trial. Nevertheless, RAND

[3] More information is available using the following link: Early years stronger practice hubs - GOV.UK (www.gov.uk)

Europe will be provided a list of all settings participating in Maths Champions by the EEF so that additional sensitivity analysis can be included to allow for possible spill-over effects.

Child level

Parents or carers were provided 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. Whilst classrooms may have included children who would not be attending school the following year, younger children were not in scope for this evaluation. 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.

To remove bias introduced by testers 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. Whilst many of the baseline assessors were able to test according to the randomised class list, others appeared to not follow the randomised class list.[4]

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.

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 appears to be truly random at the child-level, when examining balance at randomisation of the sample, we will additionally examine whether the baselined sample are significantly different from the sample available for baseline (as outlined in the imbalance at baseline section), with regard to available setting-provided covariates, such as gender, age, EYPP status[5] and EAL status.

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 mimics the design of the ONE intervention, as a whole-class intervention suitable for all children within the age range. However, if children refused to 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. As outlined above, we will assess whether the prevalence of EAL and EYPP status differs between the sample available for baseline and the final baselined sample, to further understand this. Unfortunately, there are no accurate and readily available indicators of SEND, 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

[4] We do not expect this to be an issue at endline, as the need to follow the randomised class list was to ensure random selection into the trial at the child-level. The focus at endline is to test as many children assessed at baseline as possible.

[5] Whilst we collected EYPP data at baseline, we are aware that collecting this data at baseline can lead to incomplete data given that settings cannot confirm EYPP until Spring term. Therefore, the evaluation team will collect 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 will therefore be used during the analysis.

status may make analysis under-powered. Therefore, we will not include SEND status in the evaluation of The ONE.

The number of hours a pupil attended was also not an exclusion criterion at the setting level, as the intervention is a whole-class intervention that is delivered to all children regardless of attendance pattern. However, given the attendance patterns of nursery children varies, not all children who attended the setting were in attendance during baseline assessment. Those who were not in attendance and were not assessed on baseline assessment days will not be included in endline assessment. Baseline assessment was conducted on at least two different days[6] of the week to increase the sample of eligible children for assessment. Endline assessment will be conducted on the same days of the week as baseline assessment to ensure patterns of attendance do not bias the endline sample, although it is recognised that patterns of attendance in early years settings can change over time for individual children. In this case, individual level changes in patterns of attendance may nevertheless lead to attrition at endline, which may further harm the power of the analysis; if endline data is not collected for children for whom we have baseline data, they will not be able to be included in the primary (or secondary) analysis.

Design overview

Table 1: Trial Design

Trial design, including number of arms		Two-armed, clustered randomised waitlisted-control trial
Unit of randomisation		Setting
Stratification variables (if applicable)		Region (West and South London, East of England, East Midlands, and Yorkshire and Humber); setting type (Private, Voluntary, Independent (PVI) or Maintained)
Primary outcome	variable	Mathematics attainment related to acquisition of mathematics concepts
	measure (instrument, scale, source)	Early Years Toolbox (EYT) numeracy measure, 0 – 120, Howard et al., 2022[7]
Secondary outcome(s)	variable(s)	Executive functioning (composite measure) Executive functioning (visuo-spatial)
	measure(s) (instrument, scale, source)	HTKS-R (composite measure), 0 – 118, Gonzales et al., (2021) Corsi blocks (visuo-spatial measure), 0 – 15, as used in Blakey et al., (2020) As one measure is composite and the other is domain-specific, they will not be combined to form a single measure of EF.
Baseline for primary outcome	variable	Mathematics attainment
	measure (instrument, scale, source)	Early Years Toolbox (EYT) Numeracy measure, 0 – 120, Howard et al., 2022
	variable	Executive functioning (composite measure) Executive functioning (visuo-spatial)

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

[7] Whilst 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 UK 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.

Baseline for secondary outcome	measure (instrument, scale, source)	HTKS-R (composite measure), 0 – 118, Gonzales et al., (2021) Corsi blocks (visuo-spatial measure), 0 – 15, as used in Blakey et al., (2020). As one measure is composite and the other is domain-specific, they will not be combined to form a single measure of EF.

This evaluation is a two-arm, clustered randomised effectiveness trial, as outlined in Table 1. The trial operates on a waitlist design. Half of the settings were randomly selected to form the treatment group and receive the intervention in the academic year 2023-24. The remaining settings have been placed on a waitlist to receive the intervention in the subsequent academic year (2024-2025). These waitlisted settings will serve as the control group for the evaluation. Children in these settings will continue with their usual early learning and care during the academic year 2023-24, forming a business-as-usual control. For further information on the randomisation approach, please see the Randomisation section below.

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 (Brown et al., 2023).

The delivery team will collect training attendance logs for practitioners, while the evaluation team and subcontractors will gather pupil attendance patterns from settings as a proxy for pupil-level exposure to the intervention. Practitioners are asked to record all completed activities using either a printed poster or electronically. Practitioners have not been requested to provide daily attendance lists or participation lists for each activity. This was considered, in order to get an accurate measure of dosage, but deemed to impose too substantial a burden on settings and participating staff, given the many simultaneous and competing demands on practitioners' time when with children. However, we acknowledge that attendance patterns do not accurately capture pupil-level dosage, as many settings have an open-play environment where children can choose to participate in activities, which limits the dosage analysis to some extent.

Randomisation

Randomisation of the schools 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 to treatment and control, resulting in 75 settings randomised to intervention and 75 settings randomised 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 and South London, East of England, East Midlands, and Yorkshire and Humber) and setting type (Private, Voluntary, Independent (PVI) or Maintained). A stratification by region helps ensure balance between control and treatment group since key covariates (such as EYPP/FEEE 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 settings as well as all relevant variables will be included in the final Evaluation Report at the conclusion of the study. 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 have been receiving training and will be expected to deliver the intervention in the 2023-24 academic year. Settings allocated to control will be expected to carry on with business as usual in the 2023-24 academic year and will receive training in the 2024-25 academic year given the study's waitlisted design. As these children are expected to transition to primary school by the start of the academic year 2024-25, 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 Brown 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 (e.g., illness in settings, Ofsted visits) and test administrator availability. To increase time for baseline testing, whilst 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:

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

Randomisation was initially concealed from all settings who had not completed baseline assessment. However, for two settings, 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 settings was a treatment setting, who had started but not completed baseline assessment prior to their allocation being revealed, and the other a control setting, who 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 indicates that inclusion of these children significantly alters the estimated treatment effects, primary outcome estimates will be reported without these children included.

Table 2 below 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 settings in West and South 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%) settings classed as maintained settings, meeting the EEF's target of at least 30% of the sample classified as maintained and at least 30% of the sample classified as PVI.

Table 2: The ONE randomisation results

	Control group	Intervention group	Total schools
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

Outcome measures

Baseline and endline assessment will consist of the following:

1. **Mathematics Attainment** using the Early Years Toolbox's numeracy (EYTN) subset (EYT Numbers 2 app)
2. **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 will be conducted by Qa Research with trained, blind-to-allocation administrators, on a one-on-one basis in person. 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 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 HTKS-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 will be 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 UK-specific norms for the EYTN, its use is justified by its validation on Australian children, with TN 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 straight-forward 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 pupil's age, as entered by administrators, and ends after five consecutive incorrect responses, typically lasting 7 minutes (Howard et al. 2022). The administrators' role is limited to resolving technical issues 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. For further details on EYT Numbers 2 (the EYT numeracy subtest), please see the protocol (Brown et al., 2023)

Secondary outcome measures

EF will be 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 EF suitable for 4-year-olds, assessing all three EF components (working memory, inhibitory control, cognitive flexibility). Corsi blocks is a domain-specific measure, testing visuo-spatial working memory, and is well validated for 3- and 4-year-olds and predictive of early maths outcomes (Blakey et al., 2020).

HTKS-R integrates multiple EF domains into a single game-like measure (McClelland et al., 2021). The game introduces behavioural rules, where children are asked to do the opposite (e.g., “when I say touch your head, you touch your toes”). It includes four parts with increasing complexity through the changing or introduction 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 4 points to continue onto 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. The approach taken by the creators of HTKS-R aggregates both the practice and test scores for analysis. Therefore, ffor this evaluation, we will aggregate 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 0, self-corrected responses are scored as 1 and correct responses as 2 for each item in practice and test rounds, with aggregated scores ranging from 0 to 118 for HKTS-R. HTKS-R is short, taking just 5-7 minutes to complete, and straight-forward 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 will include a full distribution of HTKS-R at baseline and endline in the evaluation report, 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 EF in the intervention and theory-of-change, rather than focussing on a single domain. Moreover, it is strongly correlated with other measures of EF (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 HKTS due to fewer floor effects in younger, socio-economically diverse children, (Gonzales et al., 2021; McClelland et al., 2021). However, whilst it is well-validated in 4-year-olds, HTKS-R has not been validated in 3-year-olds, so concerns over possible floor effects for the youngest children at baseline remain.

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) is also used. 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 EF and math ability, and remains predictive of math performance in the nursery years (Blakey et al., 2020).

The measures of EF will not be combined in this evaluation. Generating a single latent factor model of EF from the two measures was considered. However, previous examination of the factor structure for HTKS-R and EF in the early years suggests that the best model fit is a 1-factor solution (Gonzalez et al., 2021), and that HTKS-R is the only measure available which is a consistent independent predictor of early maths achievement (McClelland et al., 2021). Evidence suggests that as a single measure, the original HTKS measure can perform similarly or more strongly than individual measures of EF (McClelland et al., 2014; Lipsey et al., 2017), and provides an efficient composite measure in terms of the predictive relationship between EF 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. Given this, we propose principally measuring EF at endline through HTKS-R alone. The inclusion of Corsi blocks is to guard against possible floor effects in the youngest or most disadvantaged children, particularly at baseline, given previous evidence suggests HTKS and, to a lesser extent, HTKS-R may suffer from these issues (McClelland et al., 2021; Gonzalez et al., 2021).

An examination of baseline data suggests there was high attrition from HTKS-R compared to both EYTN and Corsi blocks. This may be due to a number of factors: it is a longer assessment, especially when compared to Corsi blocks; it is the last assessment faced by the children; and it is more complex for assessors to administer. Given there is lower attrition and less likelihood of floor effects at baseline 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 will use Corsi blocks as the baseline measure of EF in the headline model for secondary outcome analysis. We will additionally report two measure-specific models alongside this headline model to examine the degree to which estimated effect sizes depend on how EF is measured. The pre-post 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.

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	EYPP/FEEE	OVERALL	EYPP
Minimum Detectable Effect Size (MDES)		0.204	0.250	0.204	0.280
Pre-test/ post-test correlations	level 1 (pupil)	0.8	0.8	0.8	0.8
	level 2 (setting)	0.2	0.2	0.2	0.2
Intracluster correlations (ICCs)	level 2 (setting)	0.18	0.18	0.18	0.18
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided		Two-sided	
Average cluster size		12.5[8]	2.4	13	1.5
Number of settings	intervention	75	75	75	75
	control	75	75	75	75
	total	150	150	150	150
Number of children	intervention	1125	180	974	111
	control	1125	180	981	114
	total	1875	360	1955	225

At the protocol stage, 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 MDES, we made a number of assumptions: randomisation at the setting level with 50:50 allocation, alpha of 0.05 and power at 0.8, between 10 – 15 children per setting (to give an average of 12.5 per setting), and pre-test/post-test correlations of 0.8[9]. In line with other early years trials, we assumed an intra-cluster correlation of 0.18. As is standard practice, we assumed an alpha of 0.05 and a power of 0.8. All MDES calculations were made using PowerUp! (Dong & Maynard, 2013). At randomisation the MDES remained almost the same. Using the same assumption as the protocol stage, and updating the cluster size based on baselined figures (13.3 per setting baselined on at least one baseline assessment) the MDES is 0.204 overall. We additionally calculated ICC of baseline EYTN scores ($\rho = 0.27$) which was substantially higher than assumed in the protocol. If endline scores displayed a similarly high ICC, this would drive MDES up to 0.243 on the full sample.

The calculations given above in the table do not take attrition into account. We know that to date, one setting (allocated to the treatment arm) has withdrawn from the study and that the independent test administrators failed to upload all primary outcome data for one setting and partial primary outcome data for another four settings. However regrettable, these are within our assumptions for attrition at protocol. If one assumes attrition at setting level to be 23%

[9] The EYT has pre- post-test correlations of 0.89, but we have chosen to be more conservative and assumed 0.8 given the re-test was administered one week after the original.

(based on the findings of a synthesis of EEF's early years trials)[10] and at the pupil level at 20%, we obtain a range of potential MDES for the sample at randomisation as outlined below[11]:

Table 3: Attrition assumptions at randomisation

	N settings	N children	MDES
At randomisation	150	1955	0.204
Setting attrition 23%	116	1511	0.232
Pupil level attrition 20%	150	1564	0.207
Setting level attrition 23% and pupil level attrition 20%	116	1209	0.235

As is standard in EEF trials, we plan to run a subgroup analysis on children from disadvantaged backgrounds. In early years interventions, disadvantage can either be operationalised by the number of 3 and 4-year-olds in receipt of Early Years Pupil Premium (EYPP) or the number of 2-year-olds eligible for the Free Early Educations Entitlement (FEEE). Given take up of EYPP is lower for 3 and 4-year-olds than take up of FEEE amongst 2-year-olds[12], using EYPP as the basis for power calculations provides a more conservative estimate of MDES. We estimated that the average number of 3 and 4-year-olds registered for EYPP in each setting across England is 2.4[13]; assuming the intervention settings are representative of settings across England, we thus estimated that 360 pupils in the sample will be in receipt of EYPP within the intervention. At the randomisation, the average number of children reported as eligible for EYPP was substantially below expectations, at just 1.5 per setting. Updated EYPP eligibility information will be sought by RAND Europe at endline. The stated MDES for the baseline EYPP subsample is thus 0.280.

[10] Source: https://d2tic4wvo1iusb.cloudfront.net/production/documents/Early-Years-Lessons-learnt-from-EEF-trials.pdf?v=1690972141

[11] Using the same assumptions at randomisation.

[12] The Department for Education reports that 135,400 2-year-olds were registered for FEEE in 2022, whereas only 116,500 3 and 4-year-olds were in receipt of the EYPP in 2022. Source: https://explore-education-statistics.service.gov.uk/find-statistics/education-provision-children-under-5

[13] The Department for Education reports that 116,500 3- and 4-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

Analysis

Primary outcome analysis

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

RQ1: What is the difference in mathematics 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 will estimate 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 will be assessed by the Early Years Toolbox Numeracy subtest (Howard et al., 2022). Under an ITT approach, analysis will include 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's guidance (see EEF 2022).

More specifically, using multi-level modelling (MLM), we will estimate a two-level random-intercept model. The two-level model, with the first level the unit of analysis (the child) and the second level the unit of randomisation (the setting), reflects the trial design and nested nature of the data, as recommended by the (Education Endowment Foundation, 2022). 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, whilst 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 will be 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* ($\beta_{0\mathrm{j}}$=β_0+u_j) vary randomly with the setting-level residual ($\beta_{0j} \sim_{i.i.d} N(\beta_0, \sigma_u^2)$). The model will additionally control 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.

$$\text{(1) } Y_{ij} = \beta_0 + \mathrm{ONE}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij}$$

where:

Y_{ij} 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, i.e., 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 will then calculate 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 (Education Endowment Foundation, 2022), 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).[14] 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. In the case of imbalance, we will conduct a sensitivity analysis including age in equation (1) (see sensitivity analysis section).

All analyses will be done in R or Stata, version 17 and above, using the *eefanalytics* package (Vallis et al., 2021)[15].

Secondary outcome analysis

As outlined in the protocol (Brown et al., 2023), this study will answer 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 pupil's executive functioning (HTKS-R Score) while the other is the measure of visual spatial ability of the pupil (Corsi Block Score). Whilst the HTKS-R, as a composite measure, is better suited to the theory of change, a possibly larger incidence of floor effects, particularly at baseline, prompted the collection of a domain-specific measure, Corsi Blocks, as well. 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 will use the same multi-level modelling approach as in the primary analysis. That is, more specifically, we will estimate 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 will estimate fixed effects for variables used in stratification (region and setting type) at the level of each setting. In all models, raw scores will be used for both baseline and endline EF tests.

$$(1)\ Y_{ij} = \beta_0 + \mathrm{ONE}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij}$$

where:

[14] Howard et al. (2022) found a correlation of r=0.97 in a sample of 126 children aged 3 – 5 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.

[15] See here for more information https://econpapers.repec.org/software/bocbocode/s458904.htm

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, i.e., 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 will then calculate the effect size using Hedges' g for τ (more in Calculation of Effect Sizes).

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

i) **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}.

ii) **A HTKS-R model**, which uses raw HTKS-R scores both at baseline (X_{ij}) and endline (Y_{ij}).

iii) **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. Whilst the second model will likely have higher correlation between endline and baseline test scores, given it uses the same measure at both time points, the presence of higher rates of attrition in HTKS-R at baseline will likely reduce the available sample size and power of the HTKS-R model (see discussion in Secondary outcome measures section). For this reason, we propose using 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[16]. This ensures our headline secondary outcome model has the highest possible sample size, to improve the power of the analysis.

As outlined in the Secondary outcome measures, we are additionally concerned about potential floor effects in HTKS-R, particularly at baseline. The Corsi blocks model allows for a comparative analysis (albeit domain-specific) to gauge how large the floor effects might be and how they might influence the estimated effect size. We will transparently report these

[16] Correlation at baseline is 0.35

effects and will interpret our findings with an understanding of their potential impact on the efficacy assessment of the intervention.

We have noted above (see Secondary Outcome section) that missingness and attrition is higher in the secondary outcomes than for the primary outcome at baseline. Whilst we are working with the assessors to ensure this doesn't happen at endline, we cannot rule out the possibility that attrition in the secondary outcomes will be higher at endline as well. If this is concentrated in HTKS-R again, the mixed-measures and HTKS-R models will likely have lower power than the Corsi Blocks model. In this instance, we will caveat all findings appropriately and ensure we present the Corsi Blocks model alongside the other two models to ensure the effect of the intervention on EF is treated robustly. This secondary outcome analysis will be accompanied by a series of sensitivity analyses outlined below.

Given we have multiple secondary outcomes, we will 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 correction will take into account the dependent nature of the test statistics and will provide a strong control against the family-wise error rate (Clarke et al., 2019), as recommended in the EEF's evaluation guidance (Education Endowment Foundation, 2022).

All analyses will be done in R or Stata, 17 or above, making use of the *eefanalytics* package where possible. Multiple hypothesis testing corrections will be made using the *rwolf2* package in Stata (Clarke, 2021) or *crctStepdown* in R (Watson, 2024) to correct for this.

Subgroup analyses

This trial is not powered to detect the effect size of treatment on children from disadvantaged backgrounds. However, a subgroup analysis will be undertaken given the EEF's focus on this group. We have two indicators of disadvantage in the Early Years: i) pupils' eligibility for Early Years Pupil Premium eligibility (EYPP), an additional top-up for eligible 3-4 year olds, and; ii) Free Early Education Entitlement (FEEE) at the age of two. Since the target population is 3 and 4-year-olds, the more relevant measure of disadvantage is receipt of the Early Years Pupil Premium (EYPP). During baseline pupil data collection, some settings also 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 in the NPD. However, since the take up of EYPP is known to be lower than FEEE, we also captured FEEE eligibility at the age of 2 for those children where that information is known by the setting. For the purpose of analysis, however, we will use EYPP, despite the lower take-up, as this is both more relevant to the age group and likely to be more accurately reported by settings.

The analysis will be done in two stages, as recommended in the EEF evaluation guidance (Education Endowment Foundation, 2022). Firstly, we will complete the primary outcome analysis, outlined above in, on the EYPP subsample. The effect sizes and statistical uncertainty will be calculated based and reported as per the procedure outlined in Primary Outcome, using equation 1.

Secondly, we will estimate 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 will add 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 2 below:

$$(2)\quad Y_{ij} = \beta_0 + \mathrm{ONE}_j\tau + EYPP_{ij}\beta_1 + (EYPP_{ij} * ONE_j)\beta_2 + Z_j\beta_3 + X_{ij}\beta_4 + u_j + e_{ij}$$

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 will 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 sub-sample of EYPP pupils, the effect size will be calculated and outlined in the same way outlined under the Primary Outcomes analysis section. As a sensitivity check, we will additionally calculate the effect size of treatment for EYPP children using the interaction model outlined in equation 2, run on the full sample. This will be done according to the following formula:

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

The coefficients in the numerator come directly from equation 2, and the standard deviation used in the denominator is the unconditional standard deviation of the EYPP sub-sample (both treatment and control).

While any subgroup analysis that hasn't been prespecified will not be undertaken, in an unforeseen event when such an analysis is undertaken, it will be reported under *post-hoc* exploratory analysis.

Additional analyses

All analyses will be conducted using R or Stata, version 17 or later, as outlined in the primary outcome and secondary outcome analysis sections.

Sensitivity Analysis

1. **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 EYTN follows EEF guidance (Education Endowment Foundation, 2022), as age-standardised scores are not recommended by the developer. While using raw scores instead of age-standardised scores will keep 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 propose, as a sensitivity analysis, calculating the effect size of the intervention conditional on child's age.

This analysis will repeat the analysis described in the Primary Outcome analysis section (equation repeated here for ease), but will now include age in the child-level characteristics matrix X_{ij}:

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

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, i.e., the stratifying variables of geographical location and setting-type (as used for randomisation);

X_{ij} = characteristics of child *i* in setting *j*, i.e., 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.

We know that EF is also age-dependent (McClelland et al, 2021; Gonzalez et al, 2021), so we propose repeating the secondary outcome analysis, as outlined above, as well. As with the primary outcome analysis, we will use an identical model, but with the addition of age to the child-level characteristics matrix, $X_{ij.}$ We propose doing this only for the mixed-measures model, as we expect this will have the largest sample size and power.

2. Accounting for possible bias introduced by the randomisation process

As outlined above in the Randomisation section, 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 the revelation of treatment status, we will repeat the primary analysis, excluding all children at these settings baselined after allocation was revealed. If the sensitivity analysis indicates that inclusion of these children significantly alters the estimated treatment effects, headline primary and secondary outcome estimates will be reported without these children included.

3. Accounting for possible measurement error in EF assessment

We are aware that test administrators faced some difficulties in administering the EF tests at baseline, particularly HTKS-R (for further detail, see the Secondary Outcome measures section above). All secondary outcome analysis outlined above will only use raw scores, including only children where valid data exists. Valid data is defined in this evaluation by all stopping rules for the assessments being followed and sufficient data recorded to allow for raw scores to be calculated. For HTKS-R and, to a lesser degree, Corsi blocks, there were some concerns over how incomplete tests were recorded.

For Corsi blocks, there were a higher-than-expected number of incomplete responses, with some suggestion in the data that incomplete responses may have been incorrect responses (e.g., where incomplete responses were then followed by correct responses, suggesting children were still willing to engage and participate). To avoid introducing additional measurement error, on the basis of researcher interpretation of data, all observations with incomplete responses at baseline were dropped due to item-level missingness, as it is impossible to determine ex-post in a robust and non-biased way which responses are truly incomplete responses and which are more accurately incorrect responses. However, we acknowledge that this could introduce non-random missingness into the data if these incomplete tests should have been recorded as incorrect, as missingness will likely be higher at the lower end of the distribution. As a result, we propose repeating the mixed-measure model outlined above in the Secondary outcome analysis section, scoring all incomplete responses at baseline as incorrect instead, and increasing the power of the analysis at the risk of introducing measurement error. We will compare these results to understand the possible magnitude of measurement error introduced in baseline assessment. This is separate to the sensitivity analyses on the HTKS-R measurement errors outlined below.

For HTKS-R, the magnitude of the problem is somewhat greater, with assessor reporting directly contradicted the data. In over 300 cases, assessors indicated that a child did not complete the assessment; however, assessors consistently indicated that assessments were incomplete even when there is valid data and all stopping rules appear to have been followed. For the vast majority of these children, the HTKS-R raw score, should we have included this data, would have been zero; when combined with the assessor flag that the child did not complete the assessment, it is most likely in these instances assessors entered in “incorrect” for all items to allow the assessment survey to end, but that these are not a reflection of the child’s EF. However, for some of these children, valid non-zero 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 test administrators were inconsistent in their use of this flag, we will rely on the properties of the data recorded rather than the assessor-entered indication of administration. As such, the analysis outlined in the Secondary outcome analysis section will include all children who despite being flagged as not completing by assessors have valid, non-zero raw scores in all analysis outlined above.

However, we recognise that the inclusion of these children could inadvertently introduce measurement bias through this inconsistency in the reported test administration data. As this issue at baseline is largely confined to HTKS-R, our current proposed sensitivity analysis is to repeat secondary outcome models involving HTKS-R (mixed-measures model and the HTKS-R model). We will repeat the secondary outcome analysis for these two models, outlined above, excluding any child where assessors indicated the child did not complete regardless of whether valid, non-zero HTKS-R raw scores exist for these children.

Given the above issues with HTKS-R and Corsi blocks assessment, we did a further analysis of EYTN data to check for inconsistencies in the data. The tablet-based assessment mode of EYTN largely protected against these issues. However, there are a limited number of assessments in EYTN that could be a result of administration error: approximately 20 children were flagged as possible test administration error, due to the combination of very low test duration (less than a minute to get through the minimum of 10 questions) and a baseline score of zero. Combined, these may not be indicative of these children truly being at the floor, but could be indicative of assessors allowing children to tap through the ten questions quickly without listening to or processing the instructions for each item. However, given there is no clearly-stated expected minimum duration for this age group, dropping these children would be on an ad-hoc basis and risks introducing additional possible sources of researcher bias. For this reason, no sensitivity analysis is recommended.

The evaluation team is working closely with the test administrators, Qa Research, to ensure similar assessor-introduced measurement error does not occur at endline. Should these issues re-occur at endline, we will undertake similar sensitivity analyses on endline scores as well.

4. Analysis with endline data only

We are regrettably aware that the independent test administrators failed to upload child data on the primary outcome for five settings, including all primary outcome pupil data for one setting and partial data for the remaining four settings. As a result, despite having 1955 children baselined across one of the three assessments, we only have valid primary data available for 1859. We will still endline all children who were baselined in at least one of the three assessment tasks.

Given we are likely to have primary outcome data on more children at endline than at baseline, we propose running the primary outcome model outlined above in equation 1 excluding the baseline EYT numeracy variable. Assuming these data collection issues are not repeated at endline, the endline-only analysis will have a higher sample size but will not allow for conditioning on baseline attainment.

5. Analysis excluding settings participating in Maths Champions

As outlined above 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 will repeat the primary outcome analysis excluding those settings who participate in Maths Champions. We will run the primary outcome model outlined above in equation 1, excluding all settings who the EEF indicate will be participating in Maths Champions. If this sensitivity analysis suggests bias has been introduced into the primary outcome model from the inclusion of Maths Champions settings, we will report headline results excluding these settings.

Mediation Analysis

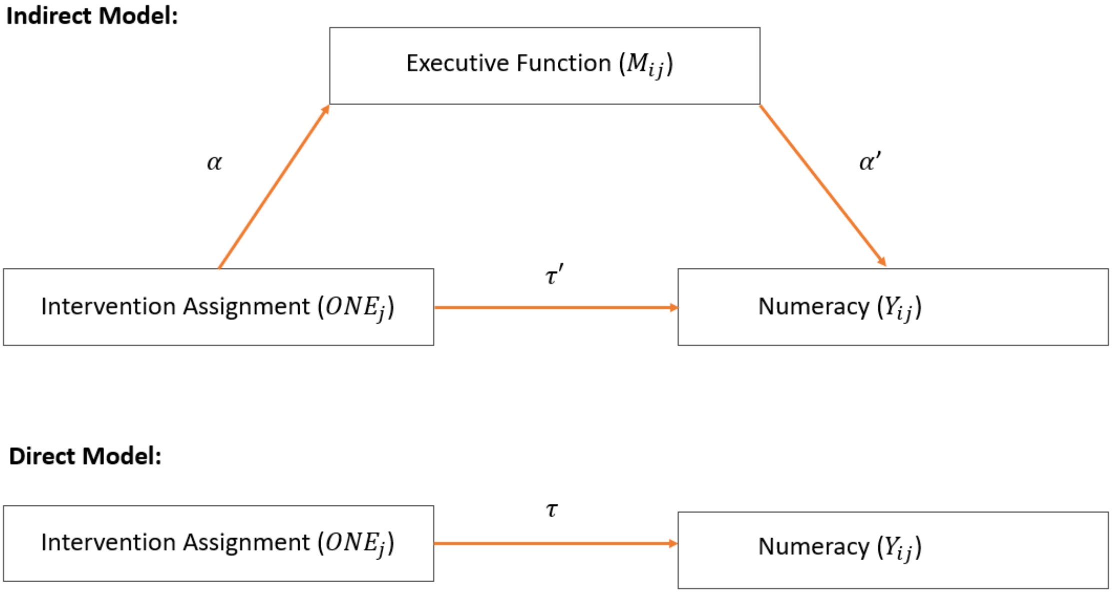
To further understand the theory of change, we will conduct a mediation analysis to explore how executive functioning (EF) potentially influences the impact of our intervention on numeracy learning outcomes. Our theory of change identifies EF as a key intermediary factor in the development of math skills.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

The mediation analysis is targeted towards answering the following question from the protocol:

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 (i.e., can the intervention logic model for EF as a mediator be validated)?

We will evaluate a multilevel model of mediated effects to determine if EF mediates the intervention's impact on early years maths attainment. Specifically, EF would be considered a mediating factor if: a) the intervention markedly improves math success, b) the intervention has a significant predictive effect on EF, c) EF substantially predicts math success, taking into account the influence of the intervention and d) the intervention predicts improvement in math success less strongly when controlling for EF (Field, 2017; Preacher & Hayes, 2004). This is illustrated in the diagram depicted in Figure 1.

Figure 1: Mediation model diagram



A multi-level mediation model can be estimated by hand, as outlined in Krull and MacKinnon (2010):

1. Effect of intervention on outcome (direct model, which is identical to the primary outcome model outlined by equation 1):

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

   The total effect of intervention on the outcome variable is given by this direct model, captured by the coefficient τ.

2. Effect of intervention on mediator (first arm of the indirect model):

$$M_{ij} = \beta_0 + \mathrm{ONE}_j\alpha + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij}$$

   This mimics the two-level random intercepts model used in the secondary and primary outcome analysis outlined above, with the dependent variable now the mediator, M_{ij}. We will use HTKS-R endline scores as the mediation variable, M_{ij}, and will be controlling in this model for baseline Corsi blocks as part of the child-level characteristics matrix, X_{ij}. The coefficient of interest in this model is α, the effect of the intervention on the mediator variable.

3. Effect of intervention on outcome, controlling for mediator (remaining arms of the indirect model):

$$Y_{ij} = \beta_0 + \mathrm{ONE}_j\tau' + M_{ij}\alpha' + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij}$$

This mimics the two-level random intercepts model outlined in the secondary and primary outcome above, with the additional introduction of the EF mediator, M_{ij}, measured by endline HTKS-R. The dependent variable is the primary outcome, EYT Numeracy. The child-level characteristics matrix would, in this instance, contain both baseline EYT Numeracy raw scores and baseline Corsi blocks scores. The coefficients of interest are the effect of EF on maths attainment (α') and the direct effect, conditional on the mediator, of the intervention on numeracy (τ').

The total effect can be measured by τ in the direct model. The direct effect of the intervention on early maths attainment can be measured by τ'in the third model, controlling for EF. The indirect effect of EF as a mediator of early maths attainment can be, as proposed in Krull and MacKinnon, 2010), measured by the product of the two indirect effect coefficients above, $\alpha\alpha'$

These effects can be readily interpreted. For EF to be a mediator, we need:

i) A positive total effect (positive τ)

ii) A positive indirect effect (positive $\alpha\alpha'$, or equivalently, positive α and positive α')

iii) A direct effect that is smaller than the total effect ($\tau' < \tau$)

Combined, the above conditions would allow EF to be interpreted as a partial mediator of the interventions effect on early maths attainment, as is hypothesised in the theory of change. For EF to be interpreted as a full mediator, the direct effect, once we control for EF, would need to be statistically insignificant. If EF is found not to be a mediator of the ONE, we would examine the individual coefficients to understand if it is due to failure of the intervention to effect EF (an “action theory failure”, as characterised by Krull and MacKinnon, 2010) or a failure of EF to mediate the change in early maths attainment (a “conceptual theory failure”, as characterised by Krull and MacKinnon, 2010).

Meaningful mediation analysis as outlined above 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. Much of the primary and secondary outcome analysis proposed above will help establish whether a mediation relationship may exist. If there is insufficient evidence that the intervention effects both early maths (the primary outcome) or EF (the secondary outcome), we will not proceed with mediation analysis, as we are likely to be facing an ‘action theory failure’ as outlined by Krull and MacKinnon (2010). We will additionally report correlations between EF, as measure by HTKS-R, and early maths, as measured by EYTN, to verify that the proposed mediator and outcome of interest are correlated. Should no significant correlation be present, we will similarly not proceed with mediation analysis.

Even if prior analysis and reported correlations indicate a mediation relationship may exist, we additionally propose, as a first stage of the analysis, to establish the following conditions, as recommended in Pieters (2017):

i) Directionality: We will primarily rely on theory of change developed in the protocol and the extensive literature on EF and early maths, to establish plausible causal directionality. A positive estimated effect for primary and secondary outcomes, combined with a significant correlation between HTKS-R and EYTN, does not establish causality, as reverse causation or causation by a third variable could be equally statistically likely.

ii) Reliability: Whilst we have concerns about measure error, particularly in the mediator (HTKSR) at baseline, we have worked with the test administrators to minimise measurement error in the endline data collection. Should concerns over measurement error remain in endline data, we will consider performing mediation analysis using the alternative EF measure, Corsi Blocks, instead.

iii) Unconfoundedness: Extensive balance at baseline testing has been proposed, due to issues outlined above. If there remains concerns, based on this balance analysis, that the model may be affected by unobserved variables, we will control for all variables that exhibit imbalance in the mediation model.

iv) Distinctiveness: we largely rely on the existing literature indicating that EF, as measured by HTKSR, and early maths performance are moderately correlated, but theoretically and empirically distinctive (Gonzales et al., 2021; McClelland et al., 2021). However, we will be reporting correlations between HTKSR and EYTN in this evaluation as well, with very high correlations giving cause for concern with regards to distinctiveness.

v) Power: The evaluation will report the MDES of the sample as analysed, to establish sufficient statistical power, with the primary and secondary outcome analysis establishing whether there is evidence to suggest a statistically significant effect between the intervention and both the potential mediator and final outcome. We note that the power of the mediation model will depend principally on the power of the secondary outcome analysis, where we know missingness is higher. To help preserve power, we will use Corsi Blocks as a baseline control for EF. If there are concerns that the mediation analysis is insufficiently powered to identify true non-null effects, such as due to higher endline attrition in HTKS-R than in other outcomes, we will consider using Corsi Blocks instead of HTKS-R as the hypothesised mediator, and any results will be appropriately caveated to reflect concerns over power.

We will proceed to the second stage, estimating the mediation model outlined above, if we can plausibly conclude that a meaningful mediation relationship exists according to the above outlined criteria. In practice, we do not propose estimating mediated effects by hand. We will use available statistical packages to estimate multilevel mediation using structural equation modelling, as outlined by Preacher et al. (2010), or other commonly-used available statistical packages, as outlined in Qingzhao and Li (2022). This will ensure we appropriately account for the two-level structure of the evaluation (with randomisation at setting level and analysis at individual level), as outlined in the analysis above and recommended in EEF evaluation guidance (Education Endowment Foundation, 2022). We will report total effects, direct effects and indirect effects, alongside their standard deviations and measures of statistical uncertainty (p-values and 95% confidence intervals). We will additionally calculate the proportion mediated, which is simply the indirect effect as a proportion of the total effect:

$$Proportion\ mediated = \frac{indirect\ effect}{total\ effect} = \frac{\alpha\alpha'}{\tau}$$

The full model, and all associated coefficients and standard deviations, will be reported in the appendix.

Given the complexities of producing effect size calculations for the indirect effect within a mediation model, since it is a non-standard regression model in which standardised differences do not sufficiently capture the entire indirect effect (Lachowicz *et al*, 2018), we will present the results of the mediation model as coefficients, rather than effect sizes. The reporting of the indirect, direct and total effects, as coefficients in the full model, should be sufficient to allow for a robust analysis of the assumptions underpinning the Theory of Change, with regard to the hypothesised mediating effect of EF on early maths, which is the primary purpose of the mediation analysis. The total effect of the intervention on pupil outcomes, as analysed in the primary and secondary outcome models, will have already been presented in terms of effect size and months progress in previous stages of the analysis, as is standard in EEF evaluations.

We also propose a suite of possible sensitivity analyses. To further explore the directionality assumption, we will test a mediation model in which the mediator and outcomes are switched. Here, this will mean testing a model in which executive function is treated as the outcome in the direct model, and numeracy ability is treated as the mediator through which the intervention impacts the outcome (EF).As briefly outlined above, if the imbalance at baseline analysis, the primary or secondary analyses (and their subsequent sensitivity and robustness analyses) indicate the presence of confounders (e.g., the age of the child), then the impact of these confounders will also be considered during the mediation analysis proposed above. This would require the inclusion of the confounder variables into the direct and indirect models outlined earlier in this section, to account for differences in this variable.

The analysis will be done in Stata 17 or higher, using *gsem*, as outlined in StataCorp (2023), or alternatively in R using the *mlma* package (Qingzhao and Li, 2022).

Longitudinal follow-up analyses

While longitudinal analysis for this study is not in scope, data-collection during evaluation will enable long-term follow-up using the National Pupil Database (NPD), despite the lack of Unique Pupil Numbers (UPNs). RAND will collect 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 would then be archived in the EEF's data archive. In addition, the delivery team will approach parents via settings in 2024 to gather permission to re-contact them later, and thereby collect further information on children while in Reception. As part of this, they will 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 pupils. 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

Imbalance at baseline

A well-conducted randomisation should create groups that are equivalent on observables at baseline, with any imbalance at baseline occurring by chance (Glennerster & Takavarasha, 2013). To check for imbalance at baseline after randomisation, baseline equivalence testing will be conducted at the setting and child level. At the setting level, we will examine the balance over setting type, Ofsted ratings and proportion eligible for EYPP. These will evaluate the distribution of each characteristic between the control and intervention groups. At the child level, balance will be assessed over: gender, EYPP status and attainment in all baseline tests. As recommended in the EEF guidance (Education Endowment Foundation, 2022), we will report differences in child-level pre-tests as effect sizes with accompanying confidence intervals. The balance at baseline will be on the sample as randomised, which produces high levels of missingness for the baseline secondary outcomes in particular, as discussed above.

At the time of writing this SAP, data on setting type and proportion eligible for EYPP is available at the setting level and data on all baseline variables are available for most children at the child level on the sample as randomised. Table 4 documents the balance between settings and children on these variables at the time of writing. At both the setting and child level, there appears to be balance across all available variables at baseline. Further data cleaning and liaison with settings should reduce missingness on some child-level variables (particularly gender and EYPP status)

Table 4. Baseline characteristics of groups as randomised

Setting-level (categorical)	Control group		Intervention group		
	n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Setting Type					
1. Maintained	33/75 (0)	44%	34/75 (0)	45.33%	
2. PVI	42/75 (0)	56%	41/75 (0)	54.67%	
Setting-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**	
Proportion of EYPP-eligible children	66/75 (9)	0.127 (0.147)	69/75 (6)	0.116 (0.165)	
Child-level (categorical)	**n/N (missing)**	**Count (%)**	**n/N (missing)**	**Count (%)**	
Gender					
1. Female	471/935 (39)	50.37%	490/955 (19)	51.31%	
2. Male	464/935 (39)	49.63%	465/955 (19)	48.69%	
EYPP status					
1. Yes	114/912 (72)	12.5%	111/923 (51)	12.03%	
2. No	798/912 (72)	87.5%	812/923 (51)	87.97%	
Child-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**	**Effect size [95% CI]**
EYT Numeracy at Baseline	929/984 (55)	23.66 (15.07)	930/974 (44)	23.61 (14.66)	0.03 (-0.16, 0.22)

HTKS-R at Baseline	769/984 (215)	39.12 (32.76)	803/974 (171)	42.78 (32.79)	0.13 (-0.06, 0.31)
Corsi Blocks at Baseline	815/984 (169)	4.33 (2.98)	846/974 (128)	4.44 (3.08)	0.05 (-0.12, 0.23)

In the final report, as above, we will not conduct statistical significance tests to evaluate balance at baseline, as any statistical differences are by pure chance and the premise does not hold (Torgerson & Torgerson, 2013). Instead, we present in Table 4 means along with distributions (for continuous variables) or counts with percentages (for categorical variables), as suggested by Senn (1994)[17]. Should concerns over imbalance arise, we will add an additional sensitivity analysis that controls for covariates imbalanced at baseline.

As flagged in the Randomisation section above, there were some concerns over the possibility that some assessors were not following protocols for child-level random selection into the trial. We collected some variables (gender, age, EYPP and EAL status) from settings on all children in the setting eligible for baseline, which we refer to as the sample available for baseline. To help understand risks to the trial introduced by some assessors not following child-level random selection protocols, we will compare the sample as randomised with the sample available for baseline on available covariates. Significant differences in observed covariates across the two samples would indicate possible bias in the selection of baselined sample. However, given only have a small selection of child-level covariates for the sample available for baseline, a lack of differences between the two samples on these covariates should not be taken as conclusive evidence that the baselined sample is unbiased in its selection.

Missing data

Missingness may occur due to attrition at setting or pupil level, pupil non-response to primary or secondary outcome testing (e.g., refusing to participate in one test), or test administration errors. Unfortunately, non-random missingness can introduce bias into the intention-to-treat approach outlined in the analysis above. To better understand the impacts of missingness on the analysis, we will report the extent of and sources of missing data and whether there is a pattern in the missingness. For all primary, secondary and subgroup analysis, we will report the extent of missing data through cross-tabulations. For the primary outcome measure, we will additionally analyse the pattern of missingness and perform a multiple imputation analysis, as recommended in the EEF evaluation guidance (Education Endowment Foundation, 2022).

To assess whether there are systematic differences or a clear pattern in missingness, we will model missingness at follow-up (defined as pupils with missing primary outcome data at endline) as a function of covariates available at baseline[18], 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 will mirror the multilevel level model specified in Equation 1, with children clustered at the setting level. In this instance, however, given the outcome will be a binary variable identifying missingness (where 1=missing; 0=complete), we will use a

[17] In some fields, it's commonly accepted that a difference of 10 percentage points or more in baseline means between treatment and control groups indicates an 'imbalance.' This is often used as a rationale to include these measures in sensitivity analyses. However, there are arguments challenging this notion. See Roberts, C. and Torgerson, D. (1999) 'Baseline imbalance in randomised controlled trials', *BMJ*, 319:185; de Boer et al. (2015) 'Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate', *International Journal of Behavioral Nutrition and Physical Activity*, 12:4.

[18] 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 baseline, we could exclude these baseline measures from this logit model.

multilevel mixed-effects logistic regression model (using Stata's melogit command, or an R equivalent). Given issues with baseline assessment, albeit concentrated largely in the secondary outcomes, we will also model missingness at baseline using this same approach to understand patterns of missingness at baseline on the sample as randomised.

We will follow the protocol suggested for missing data suggested by EEF guidance (Education Endowment Foundation, 2022). 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 are 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 are 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 multi-level 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 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 & Shi, 2021). We will follow the guidelines for MI recommended in Jakobsen et al. (2017). Given children with valid primary or secondary outcome data will be tested as endline, we will 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 will use Multiple Imputation using Chained Equations (MICE) method for imputation to allow us to impute both missing baseline and endline.

MI will only alleviate bias if 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 whilst 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 will be undertaken using the *mice*[19] package in R, or if not practical, using the *mi suite*[20] in Stata 17 or higher. The advantage of using the *mice* package in R is that multiple imputation can be carried out on the multi-level models outlined in the primary outcome section (Grund et al., 2018). If *mi suite* in Stata must be used, we will instead use clustering to account for the multilevel nature of the data, but this will mean the imputed model will structurally diverge from the primary outcome model.

Compliance and dosage

As the ITT approach is inherently conservative, capturing the average effect of offering the ONE intervention, we propose additionally estimating treatment effects for complying setting.

[19] For more information on implementing missing data analysis using the MI approach in R with the *mice* command, see here: https://www.jstatsoft.org/article/view/v045i03

[20] For more information on implementing missing data analysis using the MI approach in Stata with the `mi` command, see here https://www.stata.com/meeting/switzerland16/slides/medeiros-switzerland16.pdf

The intention-to-treat (ITT) approach, while reducing bias associated with non-random attrition, may dilute the estimated effect of an intervention due to non-compliance. We will use complier average causal effect (CACE) analysis to measure the effect of the programme among settings that are fully compliant with the intervention. This additionally analysis measures the average effect of fully compliant participation in the ONE on numeracy outcomes.

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 outlined in the protocol, full participation in the professional development arm of the training was seen as the necessary pre-condition to successful implementation, so full compliance with the intervention will be defined as at least one staff-member from each setting participating in each of the professional development sessions. Implementation of the ONE activities at least three times a week over the intervention period will instead form part of our dosage analysis, and is discussed below and additionally outlined in Table 5.

Table 5. Setting level compliance measure

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[21]	Number of hours a week the child usually attends a setting

In a situation of imperfect compliance, whereby not all participating setting are deemed compliant using the criteria outlined above, we will undertake a complier average causal effect (CACE) analysis. We will use 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 (Education Endowment Foundation, 2022). The CACE analysis rests on two main assumptions:

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

[21] Attendance data was collected at baseline prior to randomisation. It will be collected again at endline (Summer term). The endline attendance data patterns will be used in the compliance analysis unless the dataset is of a significantly poor data,

As discussed above, the CACE analysis will take compliance as a binary measure (where 1=compliant; 0=non-compliant as defined above in Table 5) defined at the setting level based on attendance logs for professional development sessions. The results of this model will establish the extent to which compliance with the intervention implementation requirements lead to improved outcomes for pupils. It will be estimated for the primary outcome, EYT numeracy, only.

The first stage of this 2SLS approach will estimate 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 will estimate a compliance rate and will be estimated using the following equation:

$$Y_j = \beta_0 + \tau\, ONE_j + \beta_2 Z_j + u_j$$

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)

u_j = setting level residual

Results for the first stage, and the associated F stat, will be reported.

The second stage of the IV estimation predicts the outcome as a function of all covariates included equation 1 (see Analysis - Primary Outcomes), but substitutes the treatment indicator ($\mathrm{ONE}_j\tau$ in equation 1) with the compliance rate estimated in the first regression (Angrist & Krueger, 1991; Angrist, 2006). Due to ease of estimation, we will use an OLS IV approach to analysis, clustering the errors at the school level. This does not mimic exactly the multi-level 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. This model will be estimated for the primary outcome measure only.

Dosage, reflecting the implementation of intervention activities, will be captured as a continuous variable. Dosage will be assessed at both the setting and child level. As outlined in the protocol and above in Table 5, the dosage at the setting level will be measured by the frequency of intervention activities provided to the children over a 12-week period. Since the intervention requires settings to three activities per week during the intervention period, the dosage will be top-coded at 36, creating a continuous dosage measure with a potential range from 0 to 36.

However, as outlined in protocol, while dosage measures at the setting level might be an appropriate approximation for individual-level dosage in certain settings, the varying attendance patterns and the free-flow nature of a classroom environment suggest that setting-level dosage might not accurately represent child-level dosage. Firstly, collecting attendance in early years settings is more difficult than in schools: unlike schools, not all settings have electronic attendance records and reports. Even if attendance perfectly matched the days intervention activities were offered, there is no guarantee that all children would choose to participate, given the free-flow nature of most early years settings. Requiring settings to provide attendance of children at each activity was deemed too burdensome for settings. Given the limitations of attendance collection in early years settings, we propose measuring child-level dosage by the registered attendance patterns of children at baseline as an estimate for pupil-level dosage Attendance will be recorded in hours, with most children expected to attend settings for at least 15 hours a week, as this is offered free to all 3- and 4-year-olds in England. In the protocol, we additionally suggested attending actual electronic attendance where possible, but this added an extra burden on settings and so we did not proceed with this third dosage measure.

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

This dosage analysis does not differentiate between the kind of activity that the student attends (numbers and counting; ordering and patterns; shapes, and spatial awareness). The implication of not differentiating between the types of activities is that the estimated treatment effect will represent the average impact of attending any intervention activity, rather than the specific impact of attending a particular type of activity. This means that the analysis will not capture the potentially different effects that various activities might have on the primary outcome measure (EYT Score). However, given the trial's design, any attendance patterns are likely to be randomly distributed. This, along with the fact that 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 will be undertaken in R or Stata, version 17 or higher.

Intra-cluster correlations (ICCs)

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.

The ICC applied in the power calculations detailed in the section about Sample Size and Power Calculations Overview, as well as at the protocol stage (refer to the protocol for further details), is set at 0.18. This value is derived from the ICC observed in previous EEF trials.

In the final report, we will present ICCs as they were at the protocol stage, at the time of randomization, and at the analysis stage. The ICC at the analysis stage will focus on the primary outcome measure. Its calculation will involve two approaches: (i) using the model corresponding to equation 1, and (ii) employing a model akin to that in equation 1 but without any covariates. This second model accounts for the clustering of pupils in schools and is referred to as the 'empty model'.

ICCs will be estimated using Stata's *estat icc* command, or an R equivalent.

Effect size calculation

As outlined in Analysis section, unless otherwise stated, we will calculate 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

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 is used in the denominator, as we assume the standard deviations of both the treatment and control groups are drawn from the same underlying population distribution. If there is cause to question this assumption at the analysis stage, we will use the unconditional standard deviation of the control group, in line with EEF guidance.

Effect sizes will be computed for each of the estimated models, and reported alongside their 95% confidence interval (CI).The effect sizes will then be converted into months of progress (for attainment measures) to facilitate interpretability.

All ES will be estimated using the *eefanalytics* Stata package.[22] If there is evidence of non-normality in residuals ofr any models, we will report non-parametric bootstrapped confidence intervals instead, as per the *eefanalytics* package.

[22] See here for more information https://ideas.repec.org/c/boc/bocode/s458904.html

References

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology*, 2(1), 23-44.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, 106(4), 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).

Blakey, E., Matthews, D., Cragg, L., Buck, J., Cameron, D., Higgins, B., Pepper, L., Ridley, E., Sullivan, E., & Carroll, D.J. (2020). The Role of Executive Functions in Socioeconomic Attainment Gaps: Results from a Randomized Controlled Trial. *Child Development, 91*, 1594-1614.

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., & 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., 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*.

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

Education Endowment Foundation. (2022). *Statistical analysis guidance for EEF evaluations*. Education Endowment Foundation.

Glennerster, R. & 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., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111-149.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32, 4:. 341 - 370 https://doi.org/10.3102/1076998606298043.

Howard, S.J., Vasseleu, E., Batterham, M., 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., Neilsen-Hewett, C., de Rosnay, M., Melhuish, E. C., Buckley-Walker, K. (2022). Validity, reliability and viability of pre-school educators' use of early years toolbox early numeracy. *Australasian Journal of Early Childhood*, 47(2), 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., & Styles, B. (2010). *A guide to running randomised controlled trials for educational researchers*. Slough: NFER.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & 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), 1-10.

Lachowitz, M; Preacher, K; Kelley, K. (2018). A Novel Measure of Effect Size for Mediation Analysis. *Psychological Methods, 22(2),* 244.

Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., & Wilson, S. J. (2017). Learning-related cognitive self-regulation measures for prekindergarten children: a comparative evaluation of the educational relevance of selected measures. *Journal of Educational Psychology*, *109*(8), 1084.

MacKinnon, D. (2012). *Introduction to Statistical Mediation Analysis* (1st ed.). Routledge.

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in psychology*, *5*, 599.

McClelland, M.M., Gonzales, C.R., Cameron, C.E., Geldhof, G.J., Bowles, R.P., Nancarrow, A.F., Merculief, 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.

Paull, G, and Popov, D. *The role and contribution of maintained nursery schools in the early years sector in England*. London: Department for Education, 2019.

Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research*, 44(3), 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, & computers*, *36*, pp.717-731.

Purpurpa, D.J. and Lonigan, C.J,, 2015. Early Numeracy Assessment: The Development of the Preschool Early Numeracy Scales. *Early Education and Development, 26:* 286–313.

Scerif, G., Gattas, S., Hawes, Z., Howard, S., Merkley, R., & O'Connor, R. (2023). Orchestrating Numeracy and The Executive: The One Programme. https://doi.org/10.31234/osf.io/2gxzv

Senn, S. (1994). *Testing for baseline balance in clinical trials*, Statistics in Medicine, 13: 1715-1726.

StataCorp. 2023. Stata 18 Structural Equation Modeling Reference Manual. College Station, TX: Stata Press

Torgerson, C., & Torgerson, D. (2013). *Randomised Controlled Trials in Education: An Introductory Handbook*.

Vallis, Dimitris, Singh, Akansha, Uwimpuhwe, Germaine, Higgins, Steve, Xiao, ZhiMin, De Troyer, Ewoud and Kasim, Adetayo, (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., Hirsh-Pasek, K., 2014. Contributions of executive function and spatial skills to preschool mathematics achievement. *Journal of Experimental Child Psychology, 126*, pp.37-51.

Watson, S. (2024). Package ‘crctStepdown’: Univariate Analysis of Cluster Trials with Multiple Outcomes. https://cran.r-project.org/web/packages/crctStepdown/crctStepdown.pdf

Yu, Qingzhao, and Bin Li. 2022. *Statistical Methods for Mediation, Confounding and Moderation Analysis Using r and SAS*. Chapman & Hall/CRC. https://www.routledge.com/Statistical-Methods-for-Mediation-Confounding-and-Moderation-Analysis/Yu-Li/p/book/9780367365479.