Evaluation summary

Project title[1]	Embedding Executive Challenge into Early Maths
Developer *(Institution)*	University of Oxford
Evaluator *(Institution)*	RAND Europe
Principal investigator(s)	Elena Rosa Brown
Protocol author(s)	Elena Rosa Brown, Merrilyn Groom, Sarah Angell
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 settings *(at design stage)*	150
Number of pupils *(at design stage)*	2250
Primary outcome measure and source	The Early Years Toolbox Early Numeracy (Howard et al., 2022).
Secondary outcome measure and source	Combined executive function measure: Heads Toes Knees Shoulders (HTKS-R) (Gonzales et al., 2021;), Corsi blocks (Richardson, 2007).

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

Protocol version history

Version	Date	Reason for revision
1.2 [*latest*]		
1.1		
1.0 [*original*]		N/A

Contents

Tables

Figures

Appendix tables

Please cross link tables, figures and appendices when mentioned throughout the text using the 'insert cross-reference' function. The captions have already been created as part of this template.

Study rationale and background

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

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

A programme integrating executive challenge into play-based activities (without the maths focus) has been trialled in Australia (Howard et al, 2020). It resulted in improvements in executive functions for the intervention settings, though improvements in attainment did not reach statistical significance. This current project – Orchestrating Numeracy and the Executive (the ONE) – builds on this work by, adapting the Australian programme to the United Kingdom Early Years context and to incorporate activities with well-evidenced maths-specific content (e.g., Moss et al., 2016), given mounting evidence that executive functions are key for early mathematical development. This content has already been co-developed with early years teachers and practitioners in pilot settings, and underwent a feasibility RCT in 16 settings in the 2021/22 academic year (Scerif et al., 2023).

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

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

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

years education recovery support package. The projects are a major part of the EEF's increased focus on generating evidence for the early years sector.

Intervention

Name: The ONE programme: "Orchestrating Numeracy and the Executive"

Why (theory/rationale):

Studies have shown that early mathematics achievement is highly predictive of later mathematics performance. We also know that children who fall behind their peers in mathematics early usually continue to develop at a slower rate than their more advanced peers and are likely to remain behind them. Yet, despite promising evidence of early maths programmes, to date not many have been evaluated using robust methods in the UK. Studies with disadvantaged children have also shown that certain elements of executive functioning (EF) are highly correlated with early maths ability, and that disadvantaged children with higher EF are more likely to do better in mathematics, both in preschool and primary school, suggesting that EF may be a key means of narrowing the attainment gap.

Who (recipients and provider):

The team at the University of Oxford and University of Sheffield (the delivery team) will train and support Early Years practitioners to run play-based maths activities that support maths development by embedding executive functioning skills into maths learning. Early Years practitioners will then deliver this play-based intervention to children who are due to start school in the following academic year (3 – 4 year olds). The ONE is a whole-class intervention, so all pupils in the classroom or playgroup inclusive of those children due to start school in the following year will receive the intervention. Large settings may be asked to nominate a target preschool room, with all staff and children in that room receiving the intervention.

The delivery team will recruit 150 settings for the trial, with up to fifteen children in each setting taking part in baseline and endline testing. Both maintained settings and private, voluntary and independent (PVI) early years settings will be eligible for the trial. Recruitment to SPH trials aim to recruit at least 30% of settings from the maintained sector and 30% of settings from the PVI sector.

The ONE is a waitlisted trial. Half of the settings will be randomised into the treatment group, and receive the intervention in the academic year 2023/2024. The other half will be put on a waitlist, with these settings receiving the intervention in the following academic year. These waitlisted settings will form the control group for the evaluation, with children in these settings receiving the usual early learning and care during the academic year 2023/2024. The children in the control group during the academic year 2023/2024 will not be exposed to the intervention, as they should have moved on to primary school by the start of the academic year 2024/2025, so the waitlist design does not hamper long-term evaluation of the intervention.

What (materials and procedures):

The ONE consists of face-to-face training for educators, a pack of 25 activity cards, and resources to be used with the activities.

All preschool staff taking part will participate in a training programme, consisting of weekly 30-minute face-to-face professional development sessions for the first four weeks of the programme. These sessions are scheduled at times and in formats that best suit practitioners at each setting (e.g., 1-to-1, or in a group). The sessions support educators' understanding of how early maths and executive functions co-develop, and they explain how executive functions can be embedded into a range of early maths learning activities whilst ensuring that children across a range of different ability levels are adequately challenged. The sessions also introduce the activity cards. The aim is to help practitioners develop a better understanding of early maths acquisition and the role of executive functioning in maths, to provide practical activities that incorporate this evidence for delivery with young children, and to increase practitioners' confidence in their own ability to run play-based activities that embed executive functions into maths learning.

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

Practitioners are provided with 25 activity cards which describe play-based maths activities across three key areas of early years mathematics (numbers and counting; ordering and patterns; shapes, and spatial awareness; all informed by the evidence-basis provided by early years mathematics experts within the extended delivery team). Each of the activity cards highlight the key mathematical and executive skills they foster, as well as how to gradually increase executive function demands within all of the activities. Some of these activities will be familiar to educators, with additional maths and executive function elements. Other activities are likely to be less familiar, and extend the breadth of maths skills that educators can support. The activities are designed to make use of commonly available resources, supplemented by a low-cost resource pack. The overall aim is for practitioners to scaffold children's maths learning at the optimal level of executive challenge, to boost early maths development.

Practitioners are asked to implement a minimum of three activities per week within the setting, including one activity from each identified area of mathematics in each week. The activities last five to ten minutes and can be embedded into preschool routines such as small group activities, outdoor play, and free play. Practitioners begin running these activities from the first week of the intervention, concurrently with the four-week professional development programme. Practitioners have the flexibility to choose how to implement the activities (big groups, small groups or a combination), as long as staff taking part in professional development and the children in the year preceding the move into Reception are included in these activities.

Where (location):
The delivery team are recruiting settings from West London, East of England, East Midlands, and Yorkshire and Humber. Both the professional development sessions and activities are carried out within the setting.

How (format):
The activities are delivered in the early years settings by the setting practitioners. Professional development will be delivered in-person by the delivery team, who will also provide each setting with a resource pack.

When and how much (dosage):
The ONE is a 12-week intervention. Settings randomised to the treatment condition will receive the intervention between January and April 2024. Baseline testing will occur in the autumn term (October – November 2023) and endline testing in the summer term (April – June 2024). Settings randomised to the waitlist will receive the intervention in September 2024.

Practitioners are asked to attend a four-session professional development programme delivered over four weeks, and to deliver a minimum of three activities per week for the 12-week duration of the programme. The professional development sessions are conducted on site at the settings during the first four weeks of the intervention. During this period, practitioners will start delivering activities in between visits. The delivery team will meet with practitioners on site to reflect on the implementation of the activities within the classroom from weeks 1 to 4 of the intervention, and will conduct a reflection session with one nominated educator per setting, during the 8th and 12th week of the intervention.

The delivery team will collect training attendance logs for practitioners. The evaluation team and subcontractors will collect child attendance patterns from settings. This will act as a proxy to understand child-level exposure to the intervention. It is important to note that attendance patterns do not capture true dosage at the child-level, as many settings adopt an open-play environment where children can choose whether to participate in the activity. Practitioners will be asked to keep a record of all completed activities using either a printed poster or electronically.

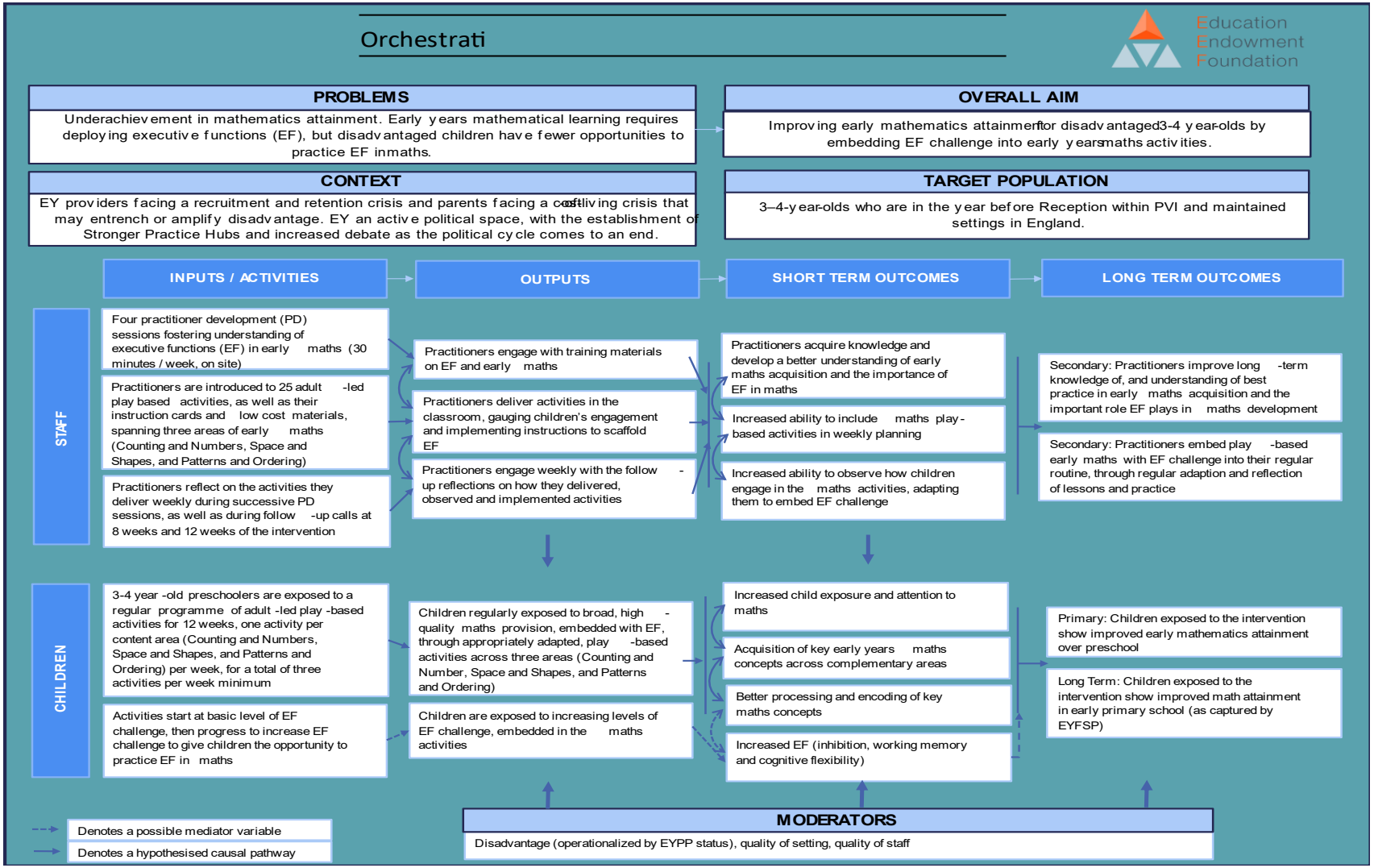
Tailoring (adaptation):
A final refinement phase has informed implementation of the intervention in this trial since completion of the feasibility study (Scerif et al., 2023). This refinement phase sought feedback particularly from educators in settings serving low-income communities. Activity cards were refined to highlight ways of differentiating activities for children who may start off from a lower knowledge basis in mathematics, or children with special educational needs (SEND) or children with English as an Additional Language (EAL). Conceptual clarifications to individual activity cards were also implemented, to help educators understand what key elements of the activities they should retain, and how these activities could be differentiated. For example, “Number Robot” was refined by providing examples of logical rules that started from matching to simple addition and subtraction, facilitating differentiation, while also retaining key mathematics and executive demands. Based on feedback from the refinement phase, no further adaptation is planned for this intervention. In the pilot, some executive

functions and maths adaptations were used by practitioners in activities, and any adaptations that occur in this trial will be recorded since they are important for fidelity to the intervention.

Control condition:

The control condition will be business-as-usual. Control settings are those settings initially randomised to the waitlist.

Figure 1: Logic model



Impact evaluation design

Research questions

1. What is the difference in maths attainment, measured by the Early Years Toolbox Numeracy, of children in the year prior to entering Reception in early years settings receiving the ONE intervention in comparison to those in control settings receiving business-as-usual?
2. 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?

Design

Table 1: Trial design

Trial design, including number of arms		Two-armed randomised waitlisted controlled trial
Unit of randomisation		Setting
Stratification variables (if applicable)		Region (West London, East of England, East Midlands, and Yorkshire and Humber); setting type (Private, Voluntary, Independent (PVI) or Maintained)
Primary outcome	**Variable**	Maths attainment related to acquisition of maths concepts
	Measure (instrument, scale, source)	Early Years Toolbox (EYT) numeracy measure, 0 – 120, Howard et al., 2022[2]
Secondary outcome(s)	**Variable(s)**	Executive functioning (composite measure) Executive functioning (visual-spatial)
	Measure(s) (instrument, scale, source)	- HTKS-R (composite measure), 0 – 118, Gonzales et al., 2021 - Corsi blocks (visual-spatial measure), 0 – 15, as used in Blakey et al., 2020. As one measure is composite and the other is domain-specific, they

[2] 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.

		will not be combined to form a single measure of EF.
Baseline for primary outcome	**Variable**	Maths attainment
	Measure (instrument, scale, source)	Early Years Toolbox (EYT) Numeracy measure, 0 – 120, Howard et al., 2022
Baseline for secondary outcome	**Variable**	Executive functioning (composite measure) Executive functioning (visual-spatial)
	Measure (instrument, scale, source)	- HTKS-R (composite measure), 0 – 118, Gonzales et al., 2021 - Corsi blocks (visual-spatial measure), 0 – 15, as used in Blakey et al., 2020. As one measure is composite and the other is domain-specific, they will not be combined to form a single measure of EF.

A two-group, stratified cluster-randomised controlled trial will be used. Settings will be the unit of randomisation, and children aged three to four will be the unit of analysis. We will stratify on region (London, East of England, East Midlands, and Yorkshire and Humber), which will achieve an equal number of settings within each region assigned to the intervention and control group respectively. We will also stratify on setting type (i.e., Private, Voluntary, and Independent (PVI) or Maintained), which will achieve an equal number of types of settings in treatment and control.

Participant selection

The study participants will be children in the year prior to entry into Reception, aged three to four, in private, voluntary and independent (PVI) settings and in maintained settings. Parents or carers are provided with a parent information sheet and given the opportunity to opt out of the trial prior to baseline collection, or withdraw their child from the evaluation at any time. Whilst classrooms may include children who will not be attending school the following year, younger children are not in scope for this evaluation. Settings will be in London, East of England, East Midlands, and Yorkshire and Humber. Settings with fewer than ten children within the relevant age range enrolled in September will initially be waitlisted and included on a case-by-case basis if necessary to reach the recruitment target of 150 settings. Should settings have more than 15 children eligible for inclusion in this evaluation, baseline assessors will assess children in the order they appear on a randomised class list. This ensures random selection into the evaluation.

There are no additional exclusion criteria. The number of hours a child attends will not be an exclusion criterion at the setting level, as the ONE intervention is a whole-class intervention that will be delivered to all children regardless of attendance pattern. However, given the attendance patterns of children vary, not all children who attend the setting will be in attendance during baseline assessment. Those who are not in attendance and are not assessed on baseline assessment days will not be included in endline assessment. Baseline assessment will be conducted on at least two different days 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.

There are no exclusion criteria on the basis of SEND or EAL status. Children are assessed at baseline regardless of SEND or EAL background, given that neither are accurately or consistently recorded in early years settings. However, if children refuse to or are evidently unable to engage with any of the assessments at baseline, they will not be included in the trial.

Each setting can only take part in one Stronger Practice Hub programme and cannot be involved in another trial that includes the same children and the same outcomes of interest (i.e., maths and EF). Settings also cannot take part in both the evaluation of the ONE and the DfE Early Years Professional Development Programme in the same year.

Planned treatment units and how they will be recruited

There will be approximately 75 treatment units (75 settings), recruited by the delivery team. Recruitment will use multiple complementary strategies:

1. direct emailing of all settings in the named Local Authorities (LA) whose contact details are available publicly, soliciting expressions of interest (EoIs) in the project to be followed up by individual calls to answer questions about the project, and offering open webinars about the project;
2. onboarding of LA Early Years specialists in each of the named LAs and cascaded recruitment e-mails / EoIs and calls / webinar offers to their contacts;
3. targeted additional "cold-calling" of settings serving low-income neighbourhoods (as defined by IMD < 5) to enrich EYPP-eligibility;
4. onboarding of Stronger Practice Hubs and cascaded recruitment as above, to their contacts.

The design is a waitlisted design, which means settings that do not receive The ONE in the 2023/2024 academic year will receive it in the 2024/2025 academic year. This is line with the funding linked to SPH where delivery is expected to be rolled out for two years, but it is also helpful for recruitment as all settings will receive the intervention.

The delivery team will reimburse settings for the research burden. In recognition of the costs associated with undertaking and implementing professional development, nurseries will be reimbursed for 50% of their staff's time/cover cost to attend training. In addition, to thank settings for participating in the research and evaluation activities, settings will receive a payment of £150 after baseline testing and £150 after endline testing.

Outcome measures

Baseline measures

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

Primary outcome

The primary outcome of the study will be maths attainment related to acquisition of maths concepts. The Early Years Toolbox (EYT) will be used to measure maths attainment. The EYT is a freely available battery of iPad-based numeracy, EF, language, self-regulation, and social development measures. Each measure was designed to be brief (i.e. approximately five minutes, including instruction and practice) and engaging. We believe the numeracy subtest is suitable for the following reasons: a) the mathematics measure has parallel versions with good validity, good test-retest reliability and have highly comparable results whether administered by a researcher or an educator; b) is easy to administer using iPads; c) has shown developmental sensitivity over the target age-range of this trial (Howard, et al. 2022).

The EYT early numeracy assessment consists of 120 interspersed items (with scores ranging from 0 to 120), which cover the following domains: number sense, cardinality and counting, numerical operations, spatial and measurement concepts, and patterning. The assessment implements automated stop and start rules: the starting item is determined by the child's age and the test is stopped after five consecutive incorrect responses. As a result, children rarely encounter all items. The average duration of the assessment under these start/stop rules was 7 minutes (range: 3 – 15 minutes).

The assessment is administered via an iPad app, in the format of a child-friendly game. It features a cartoon robot who encounters a variety of numerical problems which the child helps to solve by tapping the screen or providing a verbal response. An example screenshot is provided in Figure 2.

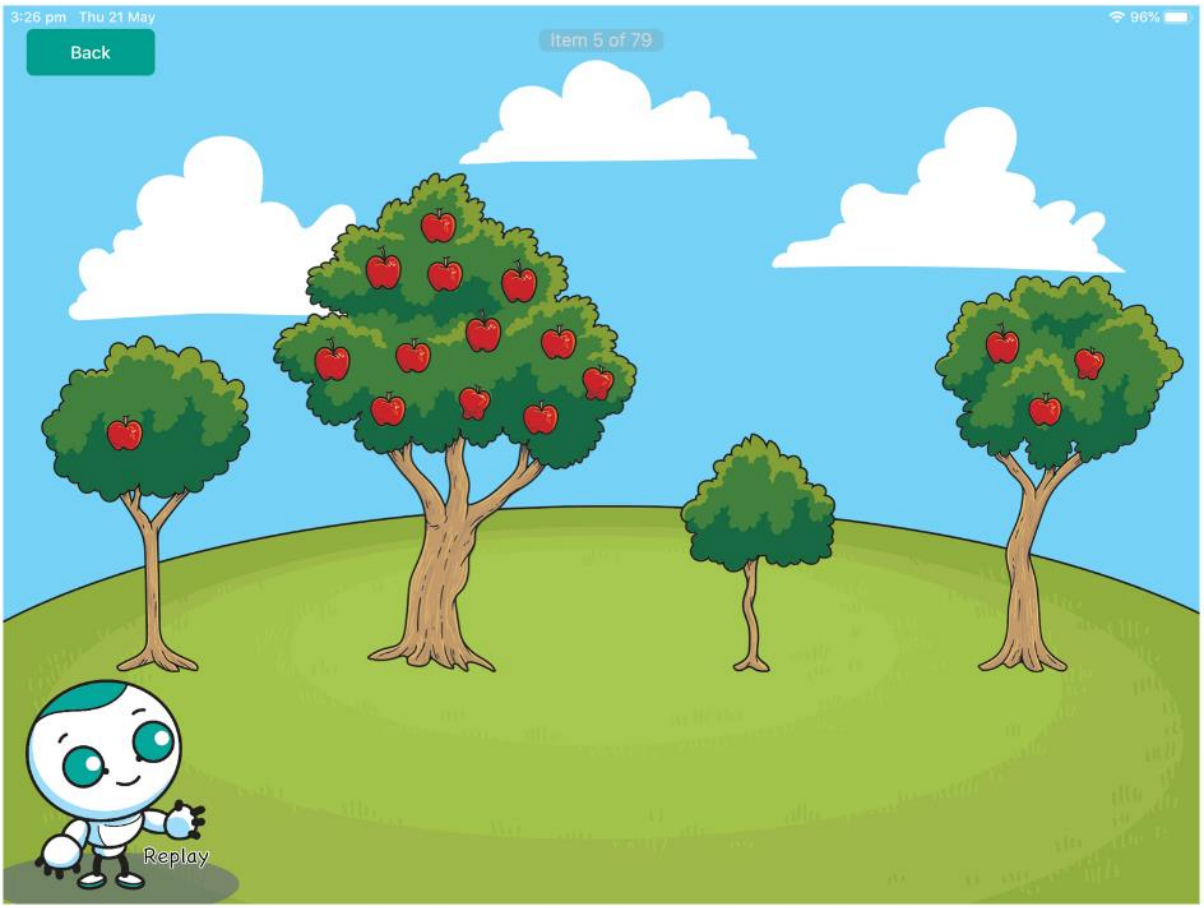


Figure 2: Example EYT Early Numeracy Item. Taken from Howard et al. (2022). The child is instructed to: "Point to the tree with only a few apples". They respond by tapping one of the trees.

.

The EYT early numeracy assessment has been validated on a sample of 246 3-5 year olds from two metropolitan areas in New South Wales, Australia (Howard et al., 2022). In this sample of children, the EYT showed good construct and concurrent validity, with significant positive correlations between EYT early numeracy two other validated early numeracy scales (DAS-II and PENS) and high test-retest reliability (r=0.89). EYT early numeracy scores also showed evidence of developmental sensitivity, with age a significant predictor of EYT early numeracy scores. Finally, the assessment showed excellent consistency between scores collected by researchers and scores collected by educations, with scores both highly correlated (r = 0.89) and not significantly different on average.

There are no validated UK norms available for the EYT early numeracy subscale. Therefore, it is impossible to establish benchmarks for children in the UK relative to their age. There are preliminary Australian norms available on the developer's website, derived from 575 Australian children between 3-5 years old attending pre-school (Howard et al, 2023). However, given baseline and endline maths attainment will be collected for all children in this trial, access to age-standardised scores is not critical for analysis. In addition, raw scores will be used for analysis.

Baseline and endline outcome testing will be administered by Qa Research with trained test administrators who are blind-to-allocation. Tests will be administered on a one-to-one basis with test administrators attending settings in person. All instructions and the stopping rule are embedded within the tablet-based task. The data collected will be immediately uploaded

to the GDPR-compliant cloud. The administrator is present to only resolve technical difficulties, and do not provide scripted prompts.

Secondary outcome

Given evidence of a link between executive functioning and early numeracy skills, and the prominent role it plays in the intervention and theory of change, we will measure executive functioning as a secondary outcome. We will use Heads-Knees-Toes-Shoulders Revised (HKTS-R) as a composite measure of executive function, and Corsi blocks as a domain-specific test of visual-spatial memory.

HKTS-R is a composite measure of executive function and behavioural self-regulation, providing a game-like measure of multiple EF components (working memory, inhibitory control, and cognitive flexibility) appropriate for use with 4-year-olds. Given the growing evidence that EF is a unidimensional concept in young children (Karr et al, 2018; McClelland et al., 2021; Morra et al., 2018), a composite measure of EF was considered appropriate for this age. It also better reflects the broader conceptualisation of EF in the intervention and theory-of-change than a domain-specific measure.

HTKS-R is divided into four parts. First, there is a warm-up round in which children are asked to provide the opposite verbal response to what the assessor says: when the assessor says “Heads”, the child responds “Toes” and vice versa. All children continue onto the next section regardless of their score on this round. Part I requires children to do the opposite action: when the assessor says “touch your head”, the child should touch their toes, and vice versa. For part II and part III, they are required to hold in their minds a second set of rules. In part II, the shoulders/knees rule is introduced: when the assessor says “touch your shoulders”, the child should touch their knees, and vice versa. This round alternates between all four instructions (touch your heads/toes/knees/shoulders). The final round switches the rules around: when the assessor says “touch your head”, the child now needs to touch their knees (and vice versa), and when the assessor says “touch your shoulders”, the child now needs to touch their toes (and vice versa). Children receive two points for each correct response, one point each time they self-correct and zero points for each incorrect response. Before each part, children are given a number of unscored practice rounds in which the assessor may prompt the child if he or she makes a mistake. For parts I, II and III, the child is required to reach a score of at least 4 points to continue onto the following part. Scores can range between 0 and 118 for HKTS-R.

HTKS-R has been shown to be: a) short, taking just 5-7 minutes, and easy to administer (Gonzales et al., 2021; McClelland et al., 2021), b) strongly correlated with other measures of EF (Gonzales et al., 2021), c) predictive of young children’s academic achievement (McClelland et al, 2021), and d) displays construct and predictive validity (McClelland et al., 2021). HKTS-R is preferred over HKTS for this trial as it has been shown to exhibit fewer floor effects for young children from lower socio-economic backgrounds (Gonzales et al, 2021; McClelland et al, 2021), exhibiting floor effects in just 3% of 4-year-olds (McClelland et al., 2021). At baseline, however, many of the children in the sample will be younger than 4, and it is not clear to what degree floor effects are prevalent with HKTS-R in 3-year-olds.

Given the possibility of floor effects in the youngest children at baseline under HKTS-R, we have included an alternative, domain-specific measure of executive function: Corsi Blocks.

The Corsi Block task primarily measures visuo-spatial memory (Corsi, 1972; Arce and McMullen, 2021). Corsi Blocks has been well-validated in young children, and used in evaluations of EF on the same age-range without any evidence of floor effects, even in children from disadvantaged backgrounds (Blakey et al., 2020).

In the Corsi Block task, nine black wooden blocks are attached to a board, as shown in Figure 3. The blocks are numbered on one side and positioned such that only the person administering the test can see the numbers. The examiner taps on a sequence of blocks, and the child then tries to tap out the same sequence they just observed. The sequence is coded as correct if the child taps the correct blocks, regardless of the order in which the blocks are tapped. The test starts with three sequences of just two blocks, and the span length (length of the sequence) increases until the child fails to recall two of the three sequences of that span, up until a span of six.

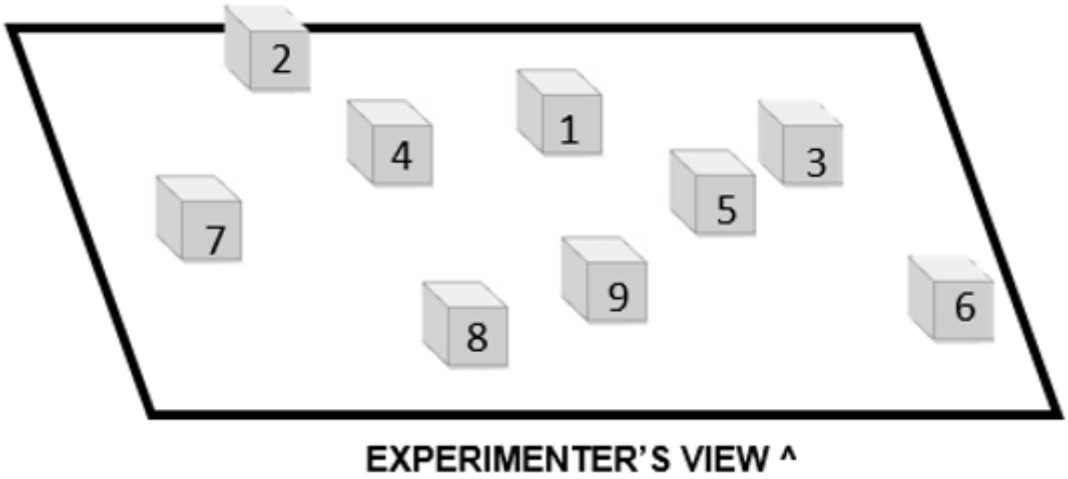


Figure 3: Corsi Blocks layout. Taken from the training manual for The ONE assessors.

Although a domain-specific measure, Corsi Blocks has been shown to be: a) short and easy to administer, b) correlated with other measures of EF and with mathematical ability in 3- and 4-year-olds (Blakey et al., 2020), and c) continues to be predictive of mathematical performance into the primary years (Alloway and Passolunghi, 2011). Given the concerns over possible floor effects of HKTS-R at baseline, Corsi Blocks was chosen as suitable alternative measure of executive function, despite being domain-specific.

Facilitating longer-term follow-up

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 directly 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.

All data protection documentation, from privacy notices and project information sheets, to DSAs, clearly state that data collected will be linked to UPNs and follow-up analysis conducted.

Sample size

Table 2: Sample size calculations

		Overall	EYPP/FEEE
Minimum Detectable Effect Size (MDES)		0.204	0.250
Pre-test/ post-test correlations	level 1 (child)	0.8	
	level 2 (setting)	0.2	
Intracluster correlations (ICCs)	level 2 (setting)	0.18	
Alpha		0.05	
Power		0.8	
One-sided or two-sided?		Two-sided	
Average cluster size		12.5[3]	2.4
Number of settings	Intervention	75	75
	Control	75	75
	Total	150	150
Number of pupils	Intervention	1125	180
	Control	1125	180
	Total	2250	360

The minimum detectable effect size (MDES) for this study has been 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 have made a number of assumptions: randomisation at the setting level with 50:50 allocation, alpha of 0.05 and power at 0.8, 15 children per setting, and pre-test/post-test correlations of 0.8[4]. In line with other early years trials, we have assumed an intra-cluster correlation of 0.18. As is standard practice, we assume an alpha of 0.05 and a power of 0.8. All MDES calculations were made using PowerUp!.

As is standard in EEF trials, we will 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

[3] Taking the average of 10 to 15 children per setting.
[4] 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.

EYPP is lower for 3- and 4-year-olds than take up of FEEE amongst 2-year-olds[5], using EYPP as the basis for power calculations provides a more conservative estimate of MDES. We estimate that the average number of 3- and 4-year-olds registered for EYPP in each setting across England is 2.4[6]; assuming the intervention settings are representative of settings across England, we thus estimate that 360 pupils in the sample will be in receipt of EYPP within the intervention.

The above calculations do not take attrition into account. If we assume attrition at the setting level of 23%[7] and pupil level attrition at 20% we have a range of potential MDES from 0.207 to 0.235, as can be seen in the table below.

	N settings	N children	MDES
At randomisation	150	2250	0.204
Setting attrition 23%	116	1035	0.233
Pupil level attrition 20%	150	1500	0.207
Setting level attrition 23% and pupil level attrition 20%	116	1160	0.235

Randomisation

Randomisation will be stratified by region and type of setting (i.e., PVI or maintained), with settings the unit of randomisation and children the unit of analysis. Given recruitment is organised regionally, with key covariates, such as EYPP/FEEE eligibility likely to vary with region, stratifying by region in randomisation helps ensure balance across treatment and control. The nature of the intervention, which involves professional development and group-run activities in a free-flow playroom environment, means it is impossible to avoid contamination between groups within settings and makes individual-level randomisation infeasible. We will also stratify on setting type, which will achieve an equal number of types of settings in treatment and control. This is because evidence suggests that maintained settings have fundamental differences compared to PVI, such as higher qualified staff, greater availability of additional and specialist services available, and a higher proportion of EYPP-eligible and SEND children (Paull and Popov, 2019; Bonetti 2020). Stratification on region and setting type will help control for some of the variation across settings, improving the precision of the estimate of treatment effect and ensure findings are applicable to all setting types. Within each stratum, settings will have an equal probability of being assigned to treatment or control.

[5] 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

[6] 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

[7] This is based on the findings of a synthesis of EEF's early years trials. Source: https://d2tic4wvo1iusb.cloudfront.net/production/documents/Early-Years-Lessons-learnt-from-EEF-trials.pdf?v=1690972141

Randomisation will occur with a 50:50 allocation to treatment and control. Settings allocated to treatment will receive ONE training and will be expected to deliver the ONE intervention in in the 2023/2024 academic year. Settings allocated to control will be expected to carry on with business as usual in the 2023/2024 academic year and will receive ONE training in the 2024/2025 academic year.

Randomisation will take place after baseline data collection. Randomisation will be conducted in Stata by a member of the evaluation team, who will be blind to treatment allocation. The code used to randomise settings as well as all relevant variables will be recorded and communicated to the implementation team in a PDF file to prevent editing.

Statistical analysis

Primary analysis

The primary analysis will be on an intention-to-treat (ITT) basis, under an analysed-as-randomised approach, whereby the analysis will include all randomised settings and baselined children in the groups to which they were randomly assigned, regardless of the treatment actually received, or deviations in programme implementation. The ITT approach is inherently conservative as it captures the averaged effect of offering the intervention, regardless of whether the participants complied with assignment. This principle is key in ensuring an unbiased analysis of intervention effects and is in line with the EEF's guidance (see EEF 2022).

The primary outcome will use raw scores on the EYT numeracy subscale, with prior attainment accounted for by raw scores on the same EYT subscale at baseline (see the 'Outcome measures' section for more detail). To estimate the impact on the primary outcome we will use a two-level multilevel model (children in settings) to account for clustering of data.

The main analysis consists of the model for outcomes of pupils nested in settings, which is:

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

where Y_{ij} is the EYT numeracy subscale score for child $_i$ in school $_j$; β_0 is the cluster-level coefficient for the slope of a predictor on number skills; ONE_j is a binary indicator of the school assignment to intervention [1] or control [0]; Z_j are school-level characteristics, here the stratifying variable of geographical location (as used for randomisation); X_{ij} represents characteristics at child level (child$_i$ in setting$_j$), specifically the pre-intervention EYT numeracy subscale score; $_{uj}$ are setting-level residuals and e_{ij} are individual-level residuals.

Equation (1) is known as a 'random intercepts' model because $\beta_{0j}=\beta_0+u_j$, the setting-specific intercept for school j, is random (it is a number that can take any value) in nature, with an assumed distribution $\beta_{0j} \sim_{i.i.d} N(\beta_0, \sigma_u 2)$. Our target parameter (i.e. the focal result of the trial) is τ, a binary treatment/control indicator variable. The effect size (Hedge's *g*) will be standardised using unconditional variance in the denominator and confidence intervals will be reported to communicate statistical uncertainty in line with EEF guidance (see EEF 2022). This will tell us the average effect of the intervention on children's numeracy outcomes in treatment settings compared to those in control settings.

Secondary analysis

The following secondary outcome analyses are planned: (1) an analysis of composite EF through HTKS-R, and (2) an analysis of visual-spatial memory using Corsi Blocks. This analysis uses the measures described in the 'Outcome measures' section.

Secondary outcomes will be assessed following the same specification to Equation (1) listed under 'Primary outcome analysis' above, but we will substitute either (1) the raw score on HTKS-R or (2) raw scores on Corsi Blocks for the secondary outcomes (i.e. *Y* in Equation 1). The vector of pupil-level characteristics in Equation (1), Xij, will include the relevant baseline scores, either (1) baseline score on HTKS-R or (2) baseline score on Corsi Blocks. Further details will be discussed in the Statistical Analysis Plan (SAP).

Analysis will be carried out on the raw scores on both sub-measures, rather than making use of factor analysis to generate an underlying latent measure of executive functioning. It is evident from the literature that EF is a unidimensional construct in this age range (Blakey et al., 2020; McClelland et al., 2021), which appears to be well-captured by HTKS-R (Gonzales et al., 2021; McClelland et al., 2021). Whilst confirmatory factor analysis is a common method to generate a unidimensional latent EF factor, in most examples in the literature, it is used when researchers have access to domain-specific measures of all three components of EF (working memory, inhibitory control and cognitive flexibility). The purpose of collecting Corsi Blocks is to safe-guard against possible floor effects at baseline in HTKS-R, rather than to augment HTKS-R. There is limited theoretical evidence to suggest HTKS-R needs to be augmented by domain-specific measures in order to generate an underlying EF factor, given it is a composite measure of EF already.

Sub-group analyses

Although the trial has not been powered to detect an effect on children from disadvantaged backgrounds, this sub-group is important particularly given the EEF' s focus. We propose to collect data from settings on pupils' eligibility for Early Years Pupil Premium eligibility (EYPP) and Free Early Education Entitlement (FEEE) at the age of two. Given the target population is 3- and 4-year-olds, the most relevant measure of disadvantage is receipt of the Early Years Pupil Premium (EYPP). However, take up of EYPP is known to be lower than FEEE, so as a secondary measure of disadvantage, we propose also capturing FEEE eligibility at the age of 2 for those children.

Analysis will, in the first instance, be undertaken with the binary EYPP variable (EYPP - eligible=l; non-EYPP-eligible=0), as it is the most relevant measure of disadvantage to the target population. Analysis of the subgroup will use two approaches, as suggested in EEF analysis guidance (EEF, 2022). The first will run the primary model given in Equation (1) on the EYPP subgroup only. Effect sizes and statistical uncertainty will be calculated on the EYPP subgroup following the procedure outline in the above section on *Primary Outcome*.

In order to make use of the entire sample, the treatment effect on the EYPP subgroup will also be estimated using an interaction model:

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

This is the same model specification as in equation (1), with the addition of the EYPP_j indicator of disadvantage and an interaction term combining EYPP eligibility and treatment allocation

($EYPP_j$ * ONE_j). The primary coefficient of interest in the interaction model is β_2, which can be interpreted as the additional treatment effect experienced by children from disadvantage background: a positive β_2 is indicative of a treatment acting as a 'gap-closer' and a negative β_2 indicative of treatment acting as a 'gap-widener'. The treatment effect size will be calculated by hand using the coefficients in the interaction models and the unconditional standard deviation of the EYPP sub-sample[8], according to EEF guidance (2022), and compared with that calculated from the model on the EYPP sub-sample.

Given the lower take-up of EYPP, as a robustness check the subgroup effect size will be recalculated using FEEE in place of EYPP. In accordance with EEF guidelines, and with the procedure outlined above for EYPP, we will again generate two estimates of effect size of the treatment on the FEEE sub-sample: i) by running the primary model given in equation (1) on the FEEE sub-sample, and ii) by running the analysis on the entire sample using the interaction model specified in equation (2) and calculating the treatment effect size by hand.

Analysis in the presence of non-compliance

As the ITT approach is inherently conservative, capturing the averaged effect of offering the intervention, we also propose to look at treatment effects in the presence of compliance. This additional 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. It is not clear whether full compliance necessitates settings being both fully compliant in the professional development activities and the implementation activities, or whether partial participation is sufficient. Given the competing pressures facing settings, mandating that settings satisfy both conditions would likely be too strict a definition, with relatively few settings achieving this. Discussions with the delivery team during the set-up meetings highlighted that full participation in the professional development arm of the training was a necessary pre-condition to successful implementation. As such, compliance with the intervention was defined as at least one staff-member from each setting participating in each of the professional development sessions. Implementation of the intervention activities instead forms one possible definition of 'dosage' and is discussed below.

In a situation of imperfect compliance, whereby not all intervention settings are deemed compliant according to the above criteria, we will undertake a complier average causal effect (CACE) analysis, using two-stage least squares (2SLS) estimation to recover the local average treatment effect (LATE) of attending a compliant setting on numeracy outcomes. Further details of how compliance will be analysed will be presented in the SAP.

Additional analyses and robustness checks

Dosage

[8] This is calculated according to the following formula: $\frac{ONE_j\tau+(ONE_j*EYPP_i)\beta_2}{sd}$

Measuring ‘dosage’ in Early Years settings is uniquely complicated by the fact that attendance and engagement varies across children, making 'dosage' measures inconsistent between children and settings. Whilst it is often easier to measure dosage at the setting level (based in this intervention on the number of times intervention activities are on offer to children), this is not necessarily an appropriate measure of child-level dosage. Inconsistencies in dosage across the treatment pool can arise either from the setting level (i.e., absences, setting-specific constraints, and turnover in intervention-trained staff leading to differences in the number of activities offered to children) or from the children themselves, through varying and inconsistent setting attendance or through varying engagement with intervention-related activities in a free-flow or free-choice playroom environment. A setting offering the intervention activities three times a week during the twelve week period will not necessarily translate into all baselined children in those settings participating in the intervention activities three times a week, due to differences in attendance patterns and engagement.

We propose measuring dosage both at the setting level and at the child level, comparing the two estimates. Dosage at the setting level can be captured by the number of times intervention activities were offered to the children over a 12-week period. Given the intervention requirement is three activities offered in each week of the intervention period, dosage will be top-coded at 36, to generate a continuous measure of dosage with a possible range of 0 to 36. The number of activities offered during the intervention period will be collected directly from settings. Analysis of dosage will follow that of compliance described above, using a 2SLS approach to estimate the average treatment effect of *being offered* one additional maths-based activity with embedded executive challenge on child-level numeracy outcomes.

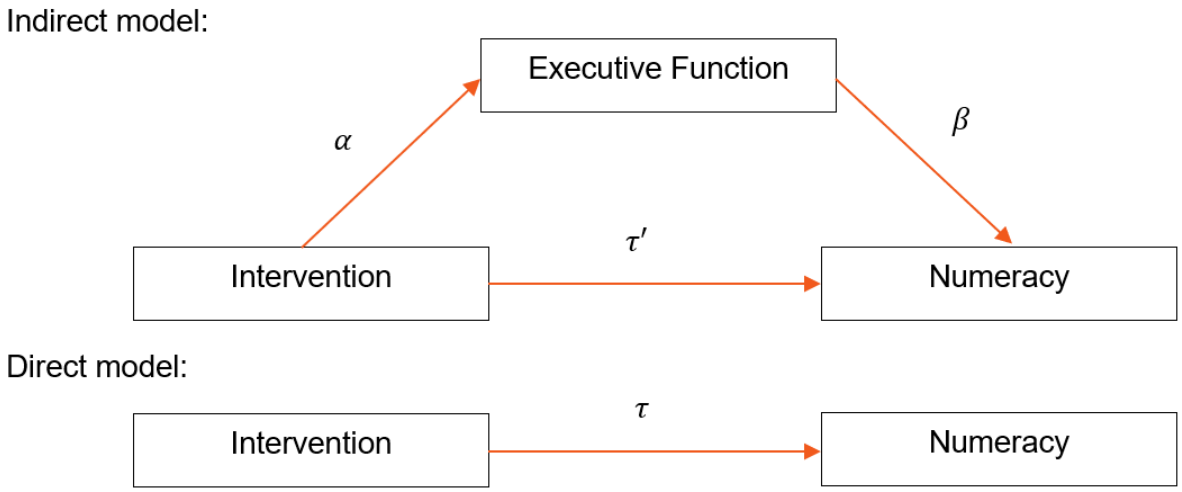
Whilst setting-level measures of dosage may serve as a suitable proxy for individual-level dosage in settings, varying attendance patterns and the nature of a free-flow classroom environment means that setting-level dosage may be a poor proxy for child-level dosage in Early Years settings. Even if attendance were perfectly correlated with the number of intervention activities children engaged in, the constraints of early years settings make it difficult to measure; unlike in schools, not all early years settings electronically record and report attendance. We propose capturing child-level dosage in two ways: i) for all settings, we will collect enrolled attendance patterns of children at baseline as a proxy for child-level dosage; and ii), as a robustness check, for those settings who can easily share attendance electronically we will collect actual attendance over the intervention period. In both instances, attendance will be measured in hours, with most children expected to attend settings for at least 15 hours a week, given this is offered free-of-charge to all 3- and 4-year-olds in England. Analysis of child-level dosage will follow that of staff-level dosage described above, using a 2SLS approach to estimate the average treatment effect of *an additional hour spent* in an intervention setting on child-level numeracy outcomes.

Mediation Model As part of this efficacy trial, we will undertake a mediation analysis to better understand the hypothesised role executive functioning plays in the effect of the intervention on numeracy outcomes. The Theory of Change explicitly models EF as a mediating variable in mathematical development. Given the strength of the literature that supports the mediating role of EF in mathematical performance (Blakey et al, 2020), there is considerable theoretical support for this mediation model.

We propose testing a simple mediation model, that measures whether EF mediates, or accounts for, the effect of the intervention on numeracy and maths attainment. More formally, EF could be considered a mediator if: a) the intervention significantly effects maths attainment,

b) the intervention significantly predicts EF, and c) EF significantly predicts maths attainment, when controlling for treatment status (Preacher and Hayes, 2004). In other words, EF is a mediator of maths attainment if there is evidence of that EF provides a causal link between the intervention and numeracy and maths attainment, as depicted in Figure 4: Mediation Model Diagram.

Figure 4: Mediation Model Diagram



The mediation model will be estimated using either Stata 17 or R. We will use bootstrapped standard errors and confidence intervals, according the procedure recommended by Preacher and Hayes (2004), as it has been shown to improve power when compared to other alternatives. Further details will be provided in the statistical analysis plan (SAP).

Imbalance at baseline

A well-conducted randomisation should yield groups that are equivalent at baseline, with any imbalance at baseline occurring by chance (Glennerster & Takavarasha 2013). However, to check for, and monitor, imbalance at baseline in the realised randomisation, baseline equivalence testing will be conducted at the setting and child level. At the setting level, we will check the balance by means of cross-tabulations and histograms over Ofsted ratings and proportion of pupils eligible for EYPP. At the child level, balance will be assessed over gender, EYPP/FEEE status, and baseline attainment.

Missing data analysis

Missing data can arise from item non-response or attrition of participants at setting and child levels. Even though it is important to include all data, it can be problematic to apply the ITT principle if we are not able to complete follow-up testing for all randomised settings or children. To better understand the pattern of missing data and its impact on the analysis, we will explore the extent of missingness, and whether there is a pattern in missingness.

Attrition across both trial arms will be explored as a basic step to assess bias. For less than 5% missingness overall, we propose to only carry out a complete-case analysis, regardless of the missingness mechanism. To gauge systematic differences in missingness, we propose to model missingness at follow-up as a function of baseline covariates, including treatment status. This will allow us to further investigate the pattern in missingness. Depending on the pattern of missingness, multiple imputation may be implemented according to EEF guidance

(2022). However, should data be missing not at random, multiple imputation will not be sufficient to generate unbiased estimates of the treatment effect, and sensitivity analysis will be carried out and reported alongside the headline impact estimates.

Implementation and process evaluation (IPE) design

Research questions

Fidelity and Adaptation

RQ.1: To what extent, how and why was the ONE delivered as planned (including training of practitioners and the implementation of the intervention by teachers)?

a) Which components of the intervention were delivered with the highest fidelity and which were implemented with the lowest fidelity (and why)?
b) To what extent does fidelity moderate outcomes of the ONE?
c) What, if any, adaptations are made to the ONE during implementation? Why are these made (are they logistical or philosophical, pro-active or reactive, in keeping with intervention logic or deviating from it)? What impact did they have on pupil responsiveness and outcomes?

Dosage

RQ.2: To what extent does variation in attendance of children in settings and engagement of children in activities affect the perceived impact of the intervention?

a) Do patterns of child attendance at their setting affect the impact of the intervention?
b) To what extent do children vary in their engagement with the activities, given the free-play environment of most early years' settings, and is there a perceived difference in impact of intervention due to engagement with the activities?

Programme differentiation and monitor of control/comparison groups

RQ.3: What are the expected activities, outputs, outcomes and impacts of the ONE?

a) To what extent does the ONE result in changes to teachers' knowledge and understanding of executive functions in math and EY skills?
b) To what extent does the ONE result in changes to teachers' ability to incorporate maths learning into routine, play-based activities?
c) To what extent does the ONE result in changes to teachers' ability to adapt activities to appropriate levels of challenge based on observed pupil engagement?
d) When can outcomes and impacts reasonably be expected to materialise, and what would make them sustainable in the longer term?

RQ.4: How do the activities, outputs, outcomes and impacts of the ONE differ from business-as-usual?

a) What characterises business-as-usual in settings? How often do they engage in maths and EF activities with the children?
b) To what extent do settings engage in other structured pedagogical activities with the children (e.g., reading and language-based activities, science activities, etc)?

c) How often do they receive professional development? Has recent professional development been targeted at numeracy and executive function?

Unintended Consequences

RQ.5: To what extent does the ONE result in positive or negative unintended consequences for children, practitioners and settings?

a) Does engagement with the ONE alter staff retention? Does it place increased pressure on staff?
b) Does engagement with the ONE crowd out other professional development?
c) Does compliance with the ONE reduce the use of other activities (e.g., activities designed to target language development and early literacy)?

Context/moderators

RQ.6: What are the barriers and facilitators to successful implementation?

a) To what extent, if at all, does the ONE particularly benefit disadvantaged pupils, compared to business as usual? What are the barriers and facilitators to the ONE benefiting disadvantage pupils?
b) To what extent, if at all, does the ONE particularly benefit EAL pupils, compared to business as usual? What are the barriers and facilitators to the ONE benefiting EAL pupils? Where does this fit in the intervention logic?
c) What are the barriers and facilitators to the ONE improving teaching practice and teachers' knowledge?

Mediators

RQ 7: 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)?

Cost

RQ.8: What is the cost of delivering the ONE and how does this compare to business-as-usual?

Research methods

We have developed a mixed-methods Implementation and Process Evaluation (IPE) data collection plan which includes document reviews, observations, surveys and semi-structured interviews, as outlined in Table 3: IPE methods overview . The tools and approaches are based on the theory of change which was co-constructed with the delivery team and the EEF. This ensures that the approach is theory-based and intervention-led.

Table 3: IPE methods overview

IPE dimension	RQ addressed	Research methods	Data collection methods	Sample size and sampling criteria	Data analysis methods

Fidelity	1,3	Observations	Training Observations (RAND Europe)	Four training sessions (one training sessions per region),	Qualitative analysis of observation results
Fidelity Programme implementation and differentiation	1,3	Interviews	Semi-structured interviews with trainers (RAND Europe)	Four interviews (an interview with one trainer per region)	An inductive approach and thematic analysis
Fidelity Adaptation Programme differentiation Usual practice Unintended consequences Moderators Cost	1,2,3,4, 5,6,8	Interviews	Semi-structured interviews with practitioners and managers (RAND Europe)	Managers and practitioners in 12 settings, chosen through purposive sampling.	An inductive approach and thematic analysis.
Fidelity Adaptation Programme differentiation Usual practice Unintended consequences	1,2,3,4	Document and management information review		Intervention training and activity materials, training attendance logs, intervention activity logs	Descriptive statistics, compliance and dosage analysis using 2SLS, thematic analysis
Fidelity Adaptation Programme differentiation Unintended consequences Moderators Cost	1, 2, 3, 4, 5, 6, 8	Surveys	Online questionnaires (RAND Europe)	All practitioners at baseline and endline, and managers at endline	Thematic analysis, descriptive statistics

Further details of each method are provided below.

1. *Document Review*

Several documents are being collected by the delivery team (i.e., Oxford University and University of Sheffield) that will be used by RAND Europe to understand implementation. These documents will be used as they reduce data collection burden on settings.

- **Intervention training and activity materials**: these will be provided to RAND Europe by the delivery team and will be used to understand fidelity and programme differentiation.
- **PD attendance logs**: these will be collected by trainers at the training and will record practitioner attendance at training. These will be collected and used by RAND Europe to understand fidelity.
- **Intervention activity log**: a record of the number of intervention activities offered to children in the settings over the course of the intervention period. These will be collected by the delivery team and used by RAND Europe to understand fidelity and dosage. Settings will give feedback on the number of activities they completed per week and what those activities were. They are given two choices for submitting this feedback: either via a feedback poster that the delivery team collect at the end of the programme or via QR code taking them to a Qualtrics survey asking for this information.

2. *Observations*

Observations of the training will be carried out by RAND Europe to understand the way in which training imparts the core components of the intervention. Observers from RAND will attend one training session per region, all of which are to be delivered by the delivery team. Training sessions will be selected randomly.

Observing training can offer insights into: the suitability of training content and the delivery mode, how the training was received, the level of knowledge among practitioners participating in training, and practitioners' hopes and concerns.

3. *Interviews*

Interviews with trainers: RAND Europe will conduct semi-structured interviews with four trainers, one per region. Interviews with trainers will allow us to understand the extent to which trainers feel prepared to deliver training, their level of knowledge of the aims of the programme, and any concerns they have.

Interviews with managers/practitioners: RAND Europe will conduct semi-structured interviews with practitioners and managers from 12 different settings that have been allocated to treatment. These interviews will complement the breadth of information gathered by the surveys and help us understand practitioners' experiences of delivery barriers/facilitators, overall level of fidelity, modifications made to the intervention, and unintended consequences along with an overall picture of settings. We will use purposive sampling to create a sample that will allow us to answer our research questions.

Interviews will be designed to last between 30 to 45 minutes and be conducted via Teams and/or phone at a time that is convenient to the interviewee. Interviews will take place towards the end of the delivery period (i.e., Summer term). No incentives will be offered.

4. *Baseline and Endline Surveys*

An online baseline practitioner survey, sent in the Autumn term, will allow us to capture a broad picture, in all settings, of business as usual, turnover/retention, and underlying knowledge. Surveys will be conducted after randomisation as the timeline prior to randomisation will not

allow sufficient time for the survey to be conducted alongside all other pre-randomisation requirements.

Online endline practitioner and manager surveys will capture changes, as well as information about practitioners' experiences of delivery barriers/facilitators, and self-reported impacts on skills and confidence. The endline survey will also be used to collect data on actual costs incurred as well as time, and any other resource implications.

Control practitioners and managers will also be requested to complete an online endline survey to help understand business as usual in the control group.

Analysis

Analysis of IPE data will be carried out through a combination of descriptive statistics and qualitative analysis using data from the research methods described above. An inductive approach will be used to code qualitative data from interviews, structuring analysis around the Theory of Change (ToC) to minimise bias. Credibility will be enhanced by triangulation across data collection approaches (i.e., interviews, surveys, observations) and researcher triangulation, with more than one researcher involved in data collection and analysis. Thematic analysis will be undertaken to analyse data from the inductive coding using themes that will support us to understand and answer the research questions.

Cost evaluation design

Costs will be evaluated using data gathered through the interviews and surveys administered to managers from across all setting. We will establish the counterfactual by evaluating the cost of business as usual in control settings, which may include the direct and indirect costs of running programmes comparable to the ONE.

The cost of the ONE implementation will be calculated per child over the course of the pre-school year. Whilst standard EEF guidance (EEF, 2022) is to calculate cost per pupil-school-years, over the course of three years, this is inappropriate in Early Years settings where many children do not attend settings for that length to time. This will be refined further in discussion with the delivery team, the EEF and other Stronger Practice Hubs evaluation teams, in an effort to ensure calculated costs are comparable across early years interventions.

The aspects of costs incurred by settings that will be gathered through our data collection tools include: direct costs of running the programme, practitioner time used for training (as staff cover is normal required to allow practitioners to undertake training), preparation and delivery of the programme, supplemental material cost incurred to deliver the programme, additional staff time used to support delivery of the programme (e.g. the need to employ other cover staff to allow practitioners time to prepare the activities). Costs such as time spent, stationery and other supplies will be monetised using market estimates. We will use sensitivity analysis to account for heterogeneity of costs between settings. The evaluation will measure pre-requisite costs, start-up costs (e.g. training), and recurring costs (e.g. costs of materials, staff time required for support).

Ethics and registration

The trial will be registered on the International Standard Randomised Controlled Trial Number (ISRCTN) registry, which is used to describe randomised controlled trials (RCTs) and efficacy

trials at inception. Once registered, this protocol will be updated with the assigned registration number.

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

Prior to children's data being sent to the delivery team, parents will be sent information sheets and withdrawal forms by the setting and will have the opportunity to return these. Parents can withdraw their children at any time from the data collection activities. Parents will be given two weeks between when information sheets are sent out and when child data is collected from settings, to allow parents to withdraw children from the evaluation before any data is collected. If parents choose to withdraw their children from data collection later on, their data will not be collected or will be deleted, as appropriate (see the privacy notice in Appendix 2).

RAND Europe will collect consent forms for all practitioners, managers and trainers that participate in an interview. The front page for each online survey will contain a privacy notice informing respondents that participation in the survey is entirely voluntary. The consent form in the survey will be built into the data collection tool so that those moving past a certain page (following the privacy notice and information on the research) will have given consent for the data to be used in the research.

None of the evaluation team has any conflicts of interest and all members of the study team have approved this protocol prior to publication.

Data protection

Our team has extensive experience handling personal data, and our researchers are accredited by the Office for National Statistics to use, for instance, data from the National Pupil Database. RAND would obtain personal data from settings as a data controller. The lawful basis for RAND Europe's use of that data under the General Data Protection Regulation (GDPR) is 'legitimate interest'.[9] Legitimate interest is an appropriate basis because the data collected as part of this evaluation will be used in ways that people would reasonably expect (i.e. for the benefit of improving support for executive functioning and early mathematical development in children) and that have minimal privacy impact. Legitimate interests apply where processing is necessary for the purpose of the legitimate interest pursued by the controller (see GDPR Article 6 (1) (f)) and for statistical and research purposes (See GDPR Article 89). The University of Oxford and their collaborator, the University of Sheffield (the delivery team) rely on public interest as the legal basis for use of the data, because research.

RAND will also obtain outcome data from its testing subcontractor, who will act as a *processor* pursuant to appropriate data sharing terms in the subcontract. Data obtained by the testing subcontractor is expected to be on the basis of legitimate interests and parents

[9] For more information about legitimate interest, please see:

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/what-is-the-legitimate-interests-basis/.

shall be provided with age-appropriate fair processing privacy notices that explain the use, storage and secure handling of the data. RAND will conduct a privacy impact assessment prior to commencing and data collection in accordance with ICO guidance on processing of child data.

The evaluation team will use basic identifiers (name and date of birth) to associate information with individuals to create datasets for research purposes. The rights and freedoms of the subjects will not be affected as information will only be identifiable during processing to the evaluation and delivery teams and not otherwise. If parents choose to withdraw their children from data collection, their children's data will not be collected, or will be deleted if already collected. Research data (not sensitive/personal data) will be kept securely by the evaluation and delivery teams for the duration of the study and deleted one year thereafter (RAND Europe) or in 2028, to allow for completion of DPhil theses associated with research.

The evaluation team have put appropriate security measures in place to keep personal data secure and to prevent any unauthorised access to or use of it. The evaluation team will collect and store all evaluation data in accordance with the Data Protection Act (2018) and GDPR requirements. Evaluation data will be stored on secure servers. Data transferred between the delivery and evaluation teams will be encrypted or use secure file transfer protocols. Data will be shared securely using specialised software (Syncplicity). No data will be saved on servers or shared with processors outside the United Kingdom or the European Economic Area (EEA) or the United Kingdom (if outside the EEA) pursuant to EEA approved terms.

Personnel

Delivery team: University of Oxford and Sheffield University
Project Leaders: Gaia Scerif and Emma Blakey
Research Project Manager: Caroline Korell
Research Officers: Rosemary O'Connor, Toni Loveridge, Carmel Brough, Sophie Smith, Joanna Archibald, Alicia Mortimer, Siobhan Murray, Alexandra Turner, Hannah Palmer, Molly Staley, Holly Amos.
Research Co-investigators: Victoria Simms (Ulster University), Zachary Hawes, Steven Howard, Rebecca Merkley, Fionnuala O'Reilly

Evaluation team: RAND Europe
Principal Investigator and overall project leader: Elena Rosa Brown
Project Manager: Merrilyn Groom
Fieldwork and analysis team: Sarah Angell

Risks

Risks have been identified and outline in Table 4.

Risks	Assessment (Likelihood/Impact)	Mitigation strategy	Impact post-mitigation
Recruitment failure	Likelihood: moderate Impact: high	This can be mitigated by regular dialogue over any recruitment issues and ensuring that the design introduces minimal burden to settings. Timelines will be discussed and agreed to ensure there is adequate time for all activities.	Moderate
Attrition	Likelihood: moderate Impact: high	We propose recruiting more settings to build in a buffer for attrition of child (assumed 23%) and setting level (assumed 15%). Settings will be given clear information about participation before signing up. Test burden will be kept low to maximise participation and reduce drop out. We will explore the possibility of settings collecting parent emails so that pupils that are removed from settings can be followed up.	Moderate
Small number of FEEE children for analysis	Likelihood: moderate Impact: moderate	We propose that the delivery team aim to recruit settings in areas of high deprivation to support analysis of FSM children in both the impact and IP evaluations	Low
Low participation rates for	Likelihood: moderate Impact: moderate	Sufficient data collection window given with real-time monitoring of response rates to allow for reminders to be	Low

IPE surveys and interviews		targeted. Ensure that the survey can be completed on phones to make it easier for practitioners who may not have easy access to a computer.	
Cross-contamination	Likelihood: low Impact: high	Setting-level randomisation proposed. Involvement in other maths-based Stronger Hubs interventions made an exclusion restriction. Information about other comparable programmes will be collected in surveys and (if necessary) factored into analysis. Settings that are part of nursery chains (i.e., multiple settings under one central administer) will be asked if they can participate as individual settings. If so they will be randomised as individual settings and contact details collected from each setting to ensure there is no confusion (i.e., the central administrator making mistakes with allocation).	Low
Quality of reporting	Likelihood: moderate Impact: moderate	Applying RAND QA processes, including expert review. PI with considerable experience of EEF reporting standards	Low
Lack of coordination between RAND Europe, the EEF and the Delivery team	Likelihood: moderate Impact: moderate	All teams have attended initial meetings and have agreed on roles and responsibilities. Regular contact between key team members from each organisation will be maintained throughout	Low
Evaluation team members' absence or turnover	Likelihood: moderate Impact: low	The team can be supplemented by researchers with considerable experience in evaluation from the larger RAND Europe pool.	Low

Timeline

Table 4: Timeline

Dates	Activity	Staff responsible/ leading
Spring term 2023	Recruitment of settings	Delivery team
	MoUs received for settings involved	Delivery team

Autumn term 2023	Baseline testing	QA Research
December 2023	Randomisation	RAND
Spring term 2024	Delivery of programme in settings	Delivery team
Spring term 2024	Training of practitioners during first four weeks of delivery	Delivery team
Spring term 2024	Delivery team holds check-in call with settings in 8th and 12th week of delivery	Delivery team
Summer term 2024	Endline testing & IPE surveys	QA Research/RAND
Summer – Autumn 2024	IPE and impact evaluation, report writing	RAND

References

Alloway, T.P. and Passolunghi, M.C., 2011. The relationship between working memory, IQ, and mathematical skills in children. *Learning and Individual Differences*, *21*(1), pp.133-137.

Arce, T. and McMullen, K., 2021. The Corsi Block-Tapping Test: Evaluating methodological practices with an eye towards modern digital frameworks. *Computers in Human Behavior Reports, 4,* pp. 1-16.

Blakey, E., Matthews, D., Cragg, L., Buck, J., Cameron, D., Higgins, B., Pepper, L., Ridley, E., Sullivan, E. and Carroll, D.J., 2020. The role of executive functions in socioeconomic attainment gaps: Results from a randomized controlled trial. *Child Development*, *91*(5), pp.1594-1614.

Blair, C., and Razza, R. P., 2007. Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten. Child Development, 78(2), pp.647-663.

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)

Bonetti, S., and Blanden, J., 2020. Early years workforce qualifications and children's outcomes: An analysis using administrative data. *Education Policy Institute.*

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.

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.

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.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, Melhuish, E. and Chadwick, S. (2023) Early Years Toolbox website: 'Norms'. http://www.eytoolbox.com.au/toolbox-norms

Karr, J.E., Areshenkoff, C.N., Rast, P., Hofer, S.M., Iverson, G.L. and Garcia-Barrera, M.A., 2018. The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological bulletin*, *144*(11), p.1147.

Moss, J., Bruce, C.D., Caswell, B., Flynn, T. & Hawes, Z. (2016). Taking shape: Classroom activities to improve young children's geometric and spatial thinking. Toronto, ON: Pearson. https://wordpress.oise.utoronto.ca/robertson/

Morra, S., Panesi, S., Traverso, L. and Usai, M.C., 2018. Which tasks measure what? Reflections on executive function development and a commentary on Podjarny, Kamawar, and Andrews (2017). *Journal of Experimental Child Psychology*, *167*, pp.246-258.

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.

Ponitz, C.C., McClelland, M.M., Matthews, J.S., and Morrison, F.J. A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes, 2009. *Developmental psychology, 45(*3), pp. 605-619.

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.

Richardson, J.T.E., 2007. Measures of Short-Term Memory: A Historical Review. *Cortex, 43*(5): 635-650.

Scerif, G., Gattas, S., Hawes, Z., Howard, S., Merkley, R., & O'Connor, R. (2023, March 7). Orchestrating Numeracy and The Executive: The One Programme. https://doi.org/10.31234/osf.io/2gxzv

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.