Independent evaluation of Maths Champions in nursery settings to develop children's early numeracy: A two-armed cluster randomised controlled trial (Maths Champions II)

Statistical Analysis Plan

Evaluator: University of York and Durham University
Principal investigators: Dr Lyn Robinson-Smith and Hannah Ainsworth

PROJECT TITLE	Independent evaluation of Maths Champions in nursery settings to develop children's early numeracy: A two-armed cluster randomised controlled trial (Maths Champions II)
DEVELOPER (INSTITUTION)	National Day Nurseries Association (NDNA)
EVALUATOR (INSTITUTION)	University of York and Durham University
PRINCIPAL INVESTIGATORS	Dr Lyn Robinson-Smith and Hannah Ainsworth, University of York
PROTOCOL AUTHORS	Lyn Robinson-Smith, Katie Whiteside, Carole Torgerson, Xiaofei Qi, Caroline Fairhurst, Louise Elliott, Catherine Hewitt, Kalpita Baird (née Joshi), Victoria Menzies, David Torgerson, Hannah Ainsworth
SAP AUTHORS	Caroline Fairhurst, Kalpita Baird, Danielle Podmore
TRIAL DESIGN	Two-armed cluster randomised controlled trial with random allocation at the nursery level
TRIAL TYPE	Effectiveness
CHILD AGE RANGE AND KEY STAGE	3-4 years at baseline (starting reception in September 2022)
NUMBER OF SETTINGS	Planned: 138 nursery settings; Actual: 134 nursery settings
NUMBER OF PUPILS	Planned: 1380 children; Actual: (estimated) 1302 children
PRIMARY OUTCOME MEASURE AND SOURCE	Maths attainment score (Assessment Profile on Entry for Children and Toddlers [ASPECTS])
SECONDARY OUTCOME MEASURE AND SOURCE	• Practitioner confidence (short survey adapted from Chen et al., 2014) • Language (ASPECTS reading and phonological awareness score) • Child development at 2 years old (Ages and Stages Questionnaire) and correlation to ASPECTS at 3 and 4 years old (for investigation as a measure of prior attainment, not a secondary attainment outcome) • Longitudinal: Child attainment at the end of Reception year at school: Early Years Foundation Stage Profile (EYFSP) data, completed at the end of Reception, collected from National Pupil Database (NPD).

Table of Contents

This analysis plan was written post-randomisation and prior to receipt of any outcome data and deals only with the statistical analysis of effectiveness for the main trial and the longitudinal analysis. This document has been written based on information contained in the study Evaluation Protocol (amended) (2) (uploaded 9 November 2021) published on the EEF website, in which full details of the background and design of the trial are presented.

SAP version history

Any changes made to the protocol which impact on the SAP, and any changes made to the SAP after its initial publication, will be specified here. There are no such changes to note to date.

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	07/07/2022	Creation of original document

Changes from protocol included in this SAP

The version of the protocol at the time of writing is Evaluation Protocol (amended) (2) (uploaded 9 November 2021) published at the EEF website. The following are changes between details provided in this version of the protocol and this SAP:

- The protocol indicates that 'If a research assistant visits a nursery to complete the baseline testing with children, baseline scores will be adjusted for within the primary outcome statistical model, which will account for any differences hypothetically caused by type of assessor at baseline.' This SAP specifies this analysis as a sensitivity analysis rather than the primary analysis model.
- The protocol indicates that all three on the subscales for the practitioner confidence survey will be analysed; however, this is an error as we are only using one subscale so only this will be analysed, as detailed in this SAP.
- Edits to the compliance criteria have been made in discussion with NDNA. For this programme it was agreed that the Maths Champion could hold Level 3 qualifications rather than graduate level. Feedback from the previous MC study suggested some practitioners would benefit from some coaching guidance. The coaching course was added as an optional element (rather than compulsory) for those who felt they needed additional support in this area. Any setting who is able to embed a minimum of 8 mandatory resources is considered well engaged with the activities. As webinars are optional, any setting that makes the time to attend at least 2 webinars is considered well engaged.
- The protocol indicates that 'Subgroup analyses looking at gender, the average number of hours the child attends the nursery setting, eligibility for EYPP [Early Years Pupil Premium], whether a child was eligible for FEEE [Free Early Education Entitlement] at 2 years old and whether the child was pre-identified to be tracked and monitored as part of the programme will be considered and detailed in the SAP.' However, data relating to which children would be tracked as part of the Maths Champions programme was not collected pre-randomisation and so this particular subgroup analysis (relating to the pre-identified children tracked and monitored as part of the programme) will not be conducted.
- Within the longitudinal analysis, the protocol indicates that attainment in Mathematics as part of the Early Years Foundation Stage Profile (EYFSP) will be assessed by summing the scores from the early learning goals (ELGs) that make up this domain, and that this will be analysed as a continuous outcome using a linear mixed model. However, recent changes to the EYFSP mean that ELGs are only scored as 'emerging' and 'expected' (the 'exceeding' option has been removed). Two ELGs make up the Mathematics domain, scoring these as 1 and 2, the sum would only range from 2-4; therefore, it is inappropriate to analyse this as a continuous measure. Instead, we shall convert this to a dichotomous variable, in terms of whether or not the participant achieved 'expected' across both ELGs. A similar approach will be taken for the Literacy domain, which consists of three ELGs. Analyses will be via mixed-effect logistic regression models.
- Within the longitudinal analysis, the protocol indicates that the analysis models will be adjusted for baseline Core Mathematics Standard Score and the minimisation factor of number of children with parent/carer agreement to participate within the setting; however, these are typos and the models will be adjusted for baseline ASPECTS numeracy/language score, and setting-level minimisation factors (as per the primary analysis model).

- The protocol indicates that, in line with the effectiveness trial analyses, subgroup analyses as part of the longitudinal analysis will consider children that were eligible for the EYPP, FEEE at 2 years old and gender. However, in the effectiveness trial analysis, a subgroup analysis considering the average number of hours that the child attends nursery will also be conducted, so this will be conducted for the longitudinal analysis also, for consistency.

Introduction

Maths Champions is a programme developed by the National Day Nursery Association (NDNA) with the aim of improving the knowledge, skills, and confidence of nursery practitioners in order to improve the quality of maths provision within their setting. This two-armed cluster randomised controlled trial (RCT) with random allocation at the nursery level will evaluate the effectiveness of the Maths Champions programme on the mathematical development and skills of children aged three and four years. Although all children within the treatment nursery settings will receive the intervention, the primary outcome of the evaluation will focus on the mathematical attainment of children who are aged three and four years at the start of the intervention, due to attend primary school in September 2022 and attend nursery for a minimum of 15 hours per week. The research questions include:

RQ 1. What is the impact of the Maths Champions programme, in comparison to usual early years setting provision, on the maths skills of pre-school children aged 3-4? *[Primary outcome]*

RQ 2. How effective is the Maths Champions programme at improving nursery practitioners' confidence in supporting children's maths development in comparison to usual early years setting provision? *[Secondary outcome 1]*

RQ 3. What is the impact of the Maths Champions programme, in comparison to usual early years setting provision, on the development of language (reading and phonological awareness) of pre-school children aged 3-4? *[Secondary outcome 2]*

RQ 4. What is the feasibility of accessing ASQ-3 data completed when children were 2 years old from NHS digital and how does this data correlate to maths and language development at 3 and 4 years old (measured using ASPECTS)? *[Secondary outcome 3]*

These research questions will be answered by analyses due to be conducted in Autumn 2022, and written up in a report to be submitted to the EEF in late 2022.

Longitudinal analysis research questions include:

LRQ 1. What is the impact of the Maths Champions programme, in comparison to usual early years setting provision, on the mathematical development of children at the end of reception, as measured by the two mathematical early learning goals of the EYFSP?

LRQ 2. What is the impact of the Maths Champions programme, in comparison to usual early years setting provision, on the literacy of children at the end of Reception, as measured by the three literacy early learning goals of the EYFSP?

LRQ 3. What is the impact of the Maths Champions programme, in comparison to usual early years setting provision, on children's overall development and school readiness, as measured by whether the child achieved a good level of development in the EYFSP?

We shall request National Pupil Database (NPD) data for randomised children only, provided their parents/carers gave consent for their child's data to be accessed. Data for the longitudinal analysis will be available in late 2023, and will be analysed and written up in an addendum report due to be submitted to the EEF in Spring 2024.

Design overview

Table 1: Study design overview

Trial design, including number of arms		Two-armed cluster randomised controlled trial
Unit of randomisation		Nursery setting
Minimisation factors		Nursery type (2 levels: PVI (private, voluntary, independent); SN (school-based nursery) and maintained settings); Nursery size (2 levels: < 30, which was the median number of children leaving for primary school in 2022 at participating settings; ≥ 30); Number of staff at the nursery holding a degree qualification in early years (2 levels: 0 graduates; ≥1 graduate)
Primary outcome	variable	Child maths attainment after 7 months intervention exposure
	measure (instrument, scale, source)	ASPECTS maths attainment score, 0-29, Centre for Evaluation and Monitoring (CEM) at Cambridge Assessment
Secondary outcomes	variables	Practitioner confidence (in teaching children maths) after 7 months intervention exposure Child language attainment after 7 months intervention exposure Child development at 2 years old and its correlation to child development at 3 and 4 years old Longitudinal: Child attainment at the end of Reception year at school
	measures (instrument, scale, source)	Practitioner confidence: Maths. Adapted 'Early Math Beliefs and Confidence Survey' by Chen et al. (2014). Only the adapted subscale 'Confidence in helping nursery aged children learn maths' ASPECTS language (reading and phonological awareness) score, 0-53, CEM at Cambridge Assessment Ages and Stages Questionnaire (ASQ-3) at 2 years old, data gathered via NHS digital and its correlation to ASPECTS (for investigation as a measure of prior attainment, not a secondary attainment outcome) Longitudinal: EYFSP data (teacher-assessed, completed at the end of Reception) collected from NPD
Baseline for primary outcome	**variable**	Child maths attainment
	measure (instrument, scale, source)	ASPECTS maths attainment score, 0-29, CEM at Cambridge Assessment
Baseline for secondary outcome	**variable**	Child language attainment
	measure (instrument, scale, source)	ASPECTS Language (reading and phonological awareness) score, 0-53, CEM at Cambridge Assessment

There are two assessment points in this trial – baseline (conducted October-December 2021), and post-intervention (planned Jun/Jul 2022). Participating children at the nurseries were assessed using the ASPECTS at baseline and will be followed up post-intervention. Nursery staff will also complete a practitioner confidence survey post-intervention. We will request for the survey to be completed by all practitioners in each setting who work with children aged 3 years or older, including the nominated Maths Champion (MC) and Deputy Maths Champion (DMC) in intervention settings and comparable staff in control settings. Nursery-level assessment using the Early Childhood Environmental Rating Scales 3 (ECERS-3) and the Early Childhood Environmental Rating Scale extension (ECERS-E) was planned to take place within a sample of four intervention settings at baseline and at outcome testing; however, due to the impact of COVID (particularly the emergence of the Omicron variant at the time of the ECERS baseline observations followed by high staff/pupil absence and tightening of visitor policies within settings) it was only possible to complete ECERS at baseline in three settings. ECERS data will be analysed as part of the implementation and process evaluation and so not detailed further in this SAP.

At baseline, where possible, a practitioner/teacher from within each nursery, who was familiar with the children, completed ASPECTS with participating children. To support this, a research assistant (RA) was provided to complete baseline assessments in nurseries that were unable to complete assessments themselves within the agreed timeframe; in the end, an RA conducted baseline assessments in eight nursery settings, and this will be investigated in a sensitivity analysis (see Analysis section).

In the protocol, we proposed to pre-test up to ten children per nursery. However, if there were less than ten children per setting, all eligible children were tested where possible. If there were more than ten eligible children then the list was randomly ordered and the first ten children in the list, who were present in the setting on the day of testing, were tested. Where possible, we want to include at least one child with Early Years Pupil Premium (EYPP) status per setting to have adequate power to conduct analyses in the EYPP subgroup. Therefore, up to three eligible children with EYPP status were randomly selected to appear at the top of the list (this was all the eligible children with EYPP status if they numbered three or fewer), then the remaining unselected children (EYPP and non-EYPP) were randomly ordered below these.

At the time of outcome testing, ASPECTS will be administered in all settings by independent RAs blinded to (unaware of) trial arm, wherever possible. However, there may be scenarios where it is not possible for a blinded RA to collect the data (e.g., the RA is sick and there isn't the opportunity to send another blinded RA to the setting). In such cases, the post-test data may be collected by an unblinded, independent assessor (e.g., a member of the YTU research team), or a practitioner/teacher within the setting. This will reduce attrition, but may impact on the internal validity of the trial as the possibility of bias is introduced (e.g., unblinded practitioners might be inclined to help or administer the post-test differently in a way that benefits children in the treatment/control group). While we expect this to affect very few cases, the impact of the ASPECTS being administered by someone other than a blinded RA will be investigated in a sensitivity analysis (see Analysis section).

It is possible that some participating children may leave the setting by the time of post-testing. A decision on whether to pursue children who have left the setting will depend on what this proportion looks like; if the numbers of those that have left are low, it may prove too resource-intensive and we would be better focussing our efforts on ensuring a high return of data from children who remain at participating nurseries. If a child leaves the nursery before outcome testing, attempts may be made to locate the child and arrange post-testing with them. The

child’s original nursery will be asked to provide the new nursery destination if they know the child has moved settings; but, if they do not know, then the child’s parents may be contacted. If they are at a new nursery, we might ask if the new setting agrees for someone to visit the nursery to complete ASPECTS testing with the child. Or, where possible, a practitioner at the new setting, who will be blind to the child's trial allocation, will be asked to complete the outcome test with the child. Alternatively, parents may be asked if a visit could be arranged at home or a local place like a library, or we might ask the parent to bring the child to their old nursery.

Randomisation

Nurseries were randomly allocated 1:1 to either receive the Maths Champions intervention; or to continue with usual nursery provision (control). A statistician at York Trials Unit (YTU), who is not involved in nursery recruitment, randomised nursery settings using minimisation to ensure balance across the trial arms on nursery type, nursery size and the number of graduate staff (see Table 1 for the levels of each minimisation factor). A dedicated computer program, MinimPy (Saghaei and Saghaei, 2011), was used for randomisation. The trial statistician will not be blind to group allocation at analysis. Settings were randomised in 17 batches between October and December 2021 after child recruitment and baseline data collection had been completed in the setting.

Naïve minimisation with base probability 1.0, following a random start, was conducted (i.e., deterministic minimisation). Naïve minimisation was deemed to be sufficient as the allocations were conducted in batches, rather than prospectively, meaning predictability was not a concern and hence a random element was not required (Altman and Bland, 2005).

The median number of children leaving for primary school in 2022 was 30 and was calculated based on expected numbers from 138 settings that expressed interest in the trial. The final number of nurseries randomised into the trial was 134 (Intervention 66; Control 68).

Sample size calculations overview

Table 2: Sample size estimations

		Protocol		Randomisation*	
		OVERALL	EYPP**	OVERALL	EYPP**
Minimum Detectable Effect Size (MDES)* **		0.20	0.38 / 0.30	0.20	0.30 / 0.39
Pre-test/ post-test correlations	level 1 (child)	0.59	0.59	0.59	0.59
	level 2 (nursery)	N/A	N/A	N/A	N/A
Intracluster correlations (ICCs)	level 2 (nursery)	0.17	N/A / 0.17	0.17	N/A / 0.17
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two
Average cluster size		10	1 / 2	9.7	1 / 2
Number of nurseries	Intervention	69	69	66	66
	Control	69	69	68	68
	Total	138	138	134	134
Number of pupils	Intervention	690	69 / 138	638	66 / 132
	Control	690	69 / 138	666	68 / 136
	Total	1,380	138 / 276	1,304	134 / 268

** based on estimated number of children for whom ASPECTS was completed at baseline, to be confirmed in final report*

*** figures either side of the / represent the two scenarios: i) aggregating data to setting-level; and ii) conducting analysis at pupil-level*

**** all estimates assume 15% pupil-level attrition at post-test*

From protocol

The following assumptions were made: a setting-level intra-cluster correlation coefficient (ICC) of 0.17 derived from the Maths Champions I trial (Robinson-Smith, 2018); a baseline and outcome testing correlation of 0.59 (from Maths Champions I); 10 children per setting at baseline; and 1:1 allocation at nursery setting level. Based on 138 nurseries (i.e., 1,380 children), we would have 80% power to show an effect size of 0.20 of a standard deviation between the control and the intervention groups, allowing for 15% attrition at the child level.

This is calculated via the following. Assuming 1,380 children are randomised, and there is 15% attrition at post-test, there will be 1,173 children included in the analysis. These 1,173 children are spread across 138 settings, which in a cluster trial equates to an approximate effective sample size in an individually randomised trial of 516. This is obtained by dividing the sample size by the design effect of $1 + (m - 1) \times \rho$ where m is the average cluster size at analysis, and ρ is the ICC. Altogether, this equates to $1{,}173/[1 + (8.5 - 1) \times 0.17]$ (Rutterford, 2015). We assume the correlation between the ASPECTS outcome

measured at pre- and post-test is 0.59; therefore, an analysis adjusting for pre-test score, as we plan here, will have the same power with 516 pupils as a t-test comparing two equal size groups with a total of $516/(1-0.59^2)$ pupils (Borm, 2007). Stata v17 was used to estimate the MDES based on a t-test comparing two groups with a total of 792 pupils, using the command *power twomeans 1, sd(1) power(0.8) n(792)*, which gives 0.20.

We will conduct a subgroup analysis for the primary outcome in EYPP pupils. Owing to the proposed sampling strategy of eligible children to participate in the trial, we hope to have at least one EYPP pupils from each setting included in this analysis. If most nurseries only have one EYPP pupil who contributes to this analysis, then the analysis for this will be conducted at the setting level, aggregating child outcomes by taking the mean for eligible EYPP children in that setting. Assuming a baseline and outcome testing correlation of 0.59 (no design effect assumed since at setting-level), with 138 nurseries we would have 80% power to show an effect size of 0.38 of a standard deviation between the control and the intervention groups in the EYPP subgroup.

If, however, more than half the settings have two or more eligible EYPP pupils who contribute to the analysis and the average number per setting is ≥2, we will conduct this analysis at the pupil level, and account for the clustering by setting. Assuming an ICC of 0.17; an average of 2 children per setting at analysis; a baseline and outcome testing correlation of 0.59; and 1:1 allocation at setting level, we would have 80% power to show an effect size of approximately 0.30 of a standard deviation between the control and intervention groups in the EYPP subgroup.

At randomisation

In total, 1,304 children were assessed using ASPECTS at baseline across 134 settings (average cluster size of 9.7). Assuming an ICC of 0.17, a pre-post test correlation of 0.59 and 15% pupil-level attrition, we will have 80% power to detect an effect size of 0.20 between the two arms.

For the EYPP analysis, at setting level, the MDES would be 0.39, and at pupil level (assuming 2 pupils per setting) the MDES would be 0.30.

It is important to note that the figures cited in the ‘At randomisation’ section are based on the best estimates of pupils who have completed baseline ASPECTS at the time of writing this SAP. Thus, these figures are subject to change.

Analysis

Analysis will follow the EEF's (2018) most recent guidance[1]. All analyses will be conducted in STATA v17 (StataCorp, 4905 Lakeway Drive, College Station, Texas 77845 USA), or later (to be confirmed in final report). All analyses will be conducted on an intention-to-treat (ITT) basis, where data are available, using all settings and children in the groups to which they were randomised irrespective of whether or not they actually received the intervention.

Statistical significance will be assessed using two-sided tests at the 5% significance level. Estimates of effect with 95% confidence intervals (CIs) and p-values will be provided.

The number of children identified as eligible for the evaluation, the number for whom parental consent was received, the number selected to take part in the evaluation, and the numbers actually tested for ASPECTS at baseline and outcome assessments will be reported with reasons for non-participation given where available. The number of children who leave the nursery before outcome testing will be reported, along with the number of these it was possible to obtain post-test ASPECTS for (if a decision is made to pursue outcome testing for children who leave their setting).

A CONSORT diagram will be produced to show the flow of settings and children through the trial.

The pairwise correlation between baseline and outcome measurements for ASPECTS scores will be presented. Histograms of pre- and post-test scores will be produced. The observed ICC for ASPECTS scores associated with setting (both baseline and outcome testing) will be presented with a 95% CI. All outcome data will be summarised descriptively by trial arm. Effect sizes based on the difference between the groups at the outcome testing will be presented as Hedges' g with 95% CI, and converted to an estimate for the number of months' progress.

Imbalance at baseline

Nursery, practitioner and child-level characteristics and baseline data will be summarised descriptively by randomised group, both as randomised (to check the randomisation achieved balance) and as analysed in the primary analysis (to check whether attrition has introduced selection bias into the complete-case sample). This will include considering the proportion of children who have a 'positive screen' on the ASQ-3 domain scores, defined as scoring less than two standard deviations below the mean area score. No formal statistical comparisons will be undertaken, except to report the differences in pre-test scores (maths and language scores from ASPECTS and ASQ-3 domains) as a Hedges' g effect size and 95% CI. Continuous measures will be reported as a mean, standard deviation (SD), median, minimum and maximum, while categorical data will be reported as a count and percentage.

Primary outcome measure and analysis

The early maths subscales of the ASPECTS will be used as baseline and outcome testing. The maths score ranges from 0 to 29, and a higher score indicates greater attainment.

Numeracy attainment for children in the intervention group and those in the control group will be compared using a linear mixed model at the child-level. Group allocation, baseline ASPECTS numeracy score, and setting-level minimisation factors (nursery type [PVI or SN/maintained settings]; nursery size; and number of staff at the nursery holding a degree qualification in early years) will be included as fixed effects in the model. The continuous

[1] Please see the Statistical Analysis Guidance.

variables that were dichotomised to use as factors in the minimisation procedure (nursery size and number of graduate nursery staff) will be included in their continuous form in the model.

Pupil-level fixed effects:

- Baseline ASPECTS numeracy score (continuous)

Setting-level fixed effects:

- Number of staff at the nursery holding a degree qualification in early years (continuous)
- Nursery size (continuous)
- Nursery type (PVI or SN/maintained settings; binary)

Adjustment will be made for clustering at the setting level by including setting as a random effect, and robust standard errors will be specified to account for any potential heteroscedasticity.

Model equation:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 w_i + \beta_3 y_i + \beta_4 I_{Ti} + \beta_5 I_{Ai} + u_i + y_{ij}$$

Y_{ij} = response (post-test ASPECTS numeracy score) of the j-th of n_i members of the i-th cluster (nursery), *i=1,…,m, j=1,…,n_i*
m = number of clusters (nursery)
n_i = size of cluster (nursery) i
x_{ij} = baseline ASPECTS numeracy score for j-th member of i-th cluster (nursery)
w_i = number of staff holding a degree qualification in early years in i-th nursery
y_i = size of i-th nursery
I_{Ti} = indicator variable for type of i-th nursery (0=PVI, 1= SN/maintained)
I_{Ai} = indicator variable for group allocation of i-th cluster (nursery) (0=Control, 1=Intervention)
β_0, β_1, β_2, β_3, β_4, β_5 = fixed effect parameters
$u_i \sim N(0, \varphi_b^2)$ = setting-specific random effect and $\gamma_{ii} \sim N(0, \varphi_w^2)$ = individual-specific random effect

The normality of the standardised residuals will be checked using a visual inspection of the QQ plot. If the model assumptions are in doubt, a sensitivity analysis will be conducted in which transformations of the outcome data will be tried to improve the model fit; these will include considering whether using the log or square root of the post-test ASPECTS score as the dependent variable in the analysis model improves the normality of the standardised residuals.

The primary analysis model shall include all post-test results regardless of the method of data collection (e.g., via a blinded RA or unblinded practitioner/teacher at the setting) or location (original or new setting). This will be explored further in sensitivity analysis, described below.

Secondary outcome measures analysis

Language

The language score from ASPECTS will be assessed at baseline and outcome time-points. This is scored from 0 to 53, where a higher score indicates greater attainment. The language score from ASPECTS will be analysed in the same way at the primary outcome, by comparing the score between the intervention and control groups, with a linear mixed model at the child-

level. Group allocation, baseline ASPECTS language score, and setting-level minimisation factors will be included as fixed effects in the model and setting as a random effect.

Practitioner Confidence

Practitioner confidence (in teaching children maths) will be assessed at outcome using a short online survey adapted from Chen et al. (2014). We will request for the survey to come completed by all practitioners in each setting who work with children aged 3 years or older, including the nominated MC and DMC in intervention settings and comparable staff in control settings. The survey will be completed at post-intervention only.

The original survey consists of three subscales: Belief about Nursery Aged Children and Maths (8 items); Confidence in Helping Nursery Aged Children Learn Maths (11 items); and Confidence in Own Maths Abilities (9 items). The three subscales produce separate scores and cannot be combined. All three subscales were collected and analysed in the Maths Champions I trial; however, in the first MC trial, the intervention was designed to improve practitioners' own maths abilities (which could justify the inclusion of subscale 3), while in this second trial improving practitioners' own maths abilities is no longer a focus of the intervention. Moreover, the Evaluation and Delivery teams agreed that there were limitations in using subscale 1, since there could be debate about what constitutes a correct/better belief, which would make interpretation of a difference in scores between the randomised groups challenging. Also, some of the questions ask about beliefs about characteristics of incoming children, which the intervention would not be expected to change. Therefore, only the second subscale: Confidence in Helping Nursery Aged Children Learn Maths (11 items) will be used.

Practitioners will be asked to rate their agreement with each item on a Likert scale, from 1 = strongly disagree to 5 = strongly agree. Scores for items will be summed to produce a summary score ranging from 11 to 55, and a higher score indicates greater confidence. The developers offer no guidance on how to handle missing item-level data for this instrument and so the scale will only be scored if a valid response is provided across all 11 items.

Responses to items in the practitioner confidence survey will be summarised descriptively by trial arm. These will be presented for all respondents and disaggregated by MC and DMC of each nursery (where these persons can be identified).

The subscale score will be compared between the two arms using a linear mixed model, adjusting for the setting-level minimisation factors (number of graduate staff, nursery type and nursery size) and highest qualification in mathematics of the respondent as fixed effects, and setting as a random effect.

Ages and Stages Questionnaire III (ASQ-3)

The Ages and Stages Questionnaire (ASQ-3) (Squires & Bricker, 2009) is used to capture the skills and development of children at 2 years old. This measure is being investigated as a measure of prior attainment, rather than as an outcome measure. The domains of the ASQ-3 include: communication, gross motor, fine motor, problem-solving and adaptive skills. A score is assigned to each development domain. Within any screened domain, less than two standard deviations below the mean area score is considered a positive screen.

The ASQ-3 is used routinely by health visitors who request that parents complete the questionnaire as part of a health check when their child is 2 years old. The data from the questionnaire is stored, and accessed, via NHS digital. Consent was sought from parents/carers to access ASQ-3 scores from NHS digital for participating children to assess

the feasibility and coverage of ASQ-3 data held within NHS digital and to determine if a correlation exists between ASQ-3 scores at 2 years old and ASPECTS scores at 3 and 4 years old.

The Evaluation Team have been communicating with NHS Digital regarding the current coverage of the ASQ-3; this process, and its findings, will be reported. However, it was ultimately decided that coverage was too low to warrant actually requesting the data, and hence no formal analysis will be performed on the ASQ data, including calculating its correlation with ASPECTS.

Subgroup analyses

Subgroup analyses looking at gender, the average number of hours the child attends the nursery setting (dichotomised at the median number of hours), eligibility for EYPP, and whether a child was eligible for FEEE at 2 years old will be undertaken for the primary outcome. These subgroup analyses will be conducted by including the factor and an interaction term between the factor and allocation in the primary analysis model.

We shall also repeat the primary analysis restricting to the subset of participants eligible for EYPP. As stated in the sample size section, if >50% of the settings only have 1 EYPP pupil included in the model, then this analysis will be conducted at the setting-level, whereby the pre- and post-test ASPECTS numeracy scores will be averaged across pupils for any setting with more than one pupil with EYPP status. Otherwise, if >50% of the settings have two or more eligible EYPP pupils who contribute to the analysis, then the analysis will be conducted at pupil-level as described for the primary analysis. The effect size estimate associated with the intervention for the EYPP subgroup will be calculated from this model.

Sensitivity analyses

An RA conducted baseline assessments in eight nursery settings; in a sensitivity analysis we will adjust for this in the primary outcome statistical model, by including a pupil-level indicator for whether the child was tested at baseline by an RA or a practitioner/teacher in their setting as a fixed-effect covariate, plus an interaction of this factor with trial arm, to account for any hypothetical differences caused by type of assessor.

A further sensitivity analysis will be conducted in which the primary analysis includes an indicator variable for whether or not the post-test ASPECTS was conducted by a blinded RA, plus an interaction term with this factor and trial arm.

A similar sensitivity analysis will be undertaken to account for the location of the post-test (original/new) should the decision be made to collect post-test ASPECTS from children who leave their original setting.

Missing data

The amount of missing primary baseline and outcome data will be summarised, and reasons for missing data explored and provided in the report, where available. If greater than 5% of children with baseline ASPECTS data are missing from the primary analysis model due to missing outcome and/or other covariate data, then multi-level logistic regression will be used to model presence or absence of the primary outcome including all available pupil and nursery-level baseline data as fixed effects, and nursery as a random effect. Significant predictors and possible mechanisms for the missing data will be discussed in the report.

The impact of missing data (if >5%) on the primary analysis will additionally be assessed using multilevel imputation via the REALCOM-impute macro, which is compatible with Stata (http://www.bristol.ac.uk/cmm/software/realcom/imputation.html). Pre- and post-intervention ASPECTS mathematics score data will be predicted by a linear regression model that includes all available pupil and nursery-level baseline variables. This imputation procedure can account for the two-level (pupil and nursery) nature of the data.

A 'burn-in' of 10 will be used (meaning that the first 10 iterations will be discarded to allow the iterations to converge to the stationary distribution before the imputation) and 30 imputed datasets will be created. (The values of 10 and 30 are subject to the convergence of the model and other values may be used during analysis). The primary analysis will then be rerun within the imputed datasets and Rubin's rules (Rubin, 1987) will be used to combine the multiply imputed estimates.

Compliance

Compliance and fidelity will be measured at the nursery setting level. Each setting in the intervention arm will be assessed for their implementation fidelity and compliance. This will be measured by NDNA who will rate each setting on compulsory and optional aspects of the MC programme.

NDNA will rate each setting on aspects of the programme on a scale of 2 = very engaged ('green'), 1 = partially engaged ('amber'), and 0 = not engaged ('red'). This will result in possible scores of 0-16 for core components, with an additional 12 points for optional components.

For the purposes of this rating scale, in this particular trial, we are not differentiating between compliance and fidelity, but seeking to capture information on both compliance and fidelity within one rating scale.

Table 3: Compulsory/Optional Components Compliance and Fidelity Rating

Criteria	Core/ Optional	Description	RAG rating
Identification of suitable Maths Champion (MC; graduate or Level 3 practitioner)	Core	MC with Level 3 or graduate qualifications	Green = 2
		MC identified with <Level 3 qualifications	Amber = 1
		MC with no level 3 qualifications or no MC identified	Red = 0
Identification of suitable Deputy Maths Champion (DMC; qualified to at least Level 3)	Core	DMC with Level 3 qualifications or higher	Green = 2
		DMC with no level 3 qualifications	Amber = 1
		No DMC identified	Red = 0
MC and DMC complete induction	Core	MC and DMC complete induction	Green = 2
		Only MC or DMC complete induction	Amber = 1

		Neither MC or DMC complete induction	Red = 0
Completion by the MC of 2 courses: Developing Mathematical Confidence in the Early Years: the big ideas of number sense; Developing Mathematical thinking in the Early Years: shape space, measures and pattern – including Characteristics of Effective Learning and sustained, shared thinking.	Core	Both completed	Green = 2
		One completed	Amber = 1
		Neither completed	Red = 0
Use of audit tool	Core	Audit Tool used and audit completed	Green = 2
		Audit Tool used but audit not completed	Amber = 1
		Audit Tool not used	Red = 0
Completion and continued use of an action plan	Core	Action plan done and used as working document throughout	Green = 2
		Action plan done, started to be used but then not implemented	Amber = 1
		Action plan not done/not used	Red = 0
Use of up to 10 mandatory resources provided through online platform: 3-4 year olds: Build a maze, Number hunt, Delivering the post, Mud kitchen, Cars down a ramp, Patterns, Construction, Tidy up time, Snack time, Outdoor games	Core	Use of at least 8 mandatory resources	Green = 2
		Use of 5-7 mandatory resources	Amber = 1
		Use of 4 or less mandatory resources	Red = 0
Engagement with one-to-one support provided by NDNA	Core	Setting always receptive to support from NDNA	Green = 2
		Setting sometimes receptive to support from NDNA	Amber = 1
		Setting never receptive to support from NDNA	Red = 0
Possible Total Score Core Components			**16**
Track and Monitor development of 6 children on termly basis.	Optional	All done and evidence uploaded	Green = 2
		Some done but needed support	Amber = 1
		None done	Red = 0
Monthly webinars	Optional	Attend two or more	Green = 2
		Attend one	Amber = 1
		Attend none	Red = 0
Completion by the DMC 2 courses:	Optional	Both completed	Green = 2

Developing Mathematical Confidence in the Early Years: the big ideas of number sense; Developing Mathematical thinking in the Early Years: shape space, measures and pattern – including Characteristics of Effective Learning and sustained, shared thinking.		One completed	Amber = 1
		Neither completed	Red = 0
Completion by MC/DMC of Coaching as an Educational Lead course	Optional	Both MC and DMC complete	Green = 2
		Only MC completes	Amber = 1
		Neither MC nor DMC complete, or DMC completes but MC does not	Red = 0
Reflection and completion of case study based on outcomes of action plan	Optional	Case study submitted demonstrating impact of change as a result of the programme	Green = 2
		Case study started or planned	Amber = 1
		Case study not started or planned	Red = 0
Compliance review via online platform – note: this is the portfolio review.	Optional	Case study submitted demonstrating impact of change as a result of the programme	Green = 2
		Case study started or planned	Amber = 1
		Case study not started or planned	Red = 0
Possible Total Score Optional Components			**12**
Possible Total Score Core and Optional Components			**28**

Dosage is defined as the length of time (in weeks) a nursery setting is delivering the MC programme. In this effectiveness trial, the intended duration of programme delivery is 7 to 8 months. This will start on the day NDNA make contact with the setting to begin the MC programme and end when post-testing occurs, or when the setting expresses a desire to no longer implement the Maths Champions programme or when NDNA withdraw their support, whichever is sooner. The dosage will be summarised.

The compliance scores (total scores for core components, optional components, and both combined) will be summarised. Complier Average Causal Effect (CACE) analyses will be considered to account for compliance/engagement of the nurseries with the programme. An instrumental variable, two-stage least squares (2SLS) approach will be used, with random group allocation as the instrumental variable (Dunn, 2005) with cluster standard errors to account for clustering at the nursery level. Three CACE analyses for the primary analysis, will be conducted; one will use the continuous compliance score considering the total score out of 16 for all the core components, and two will define compliance at the nursery level as a dichotomous variable as described below:

- Settings engaging at least **minimally** with the programme (defined as the nursery being rated an amber score of 1 or a green score of 2, in at least one of the core

aspects of the programme, total core component score of at least 1 out of 16), vs setting received no intervention at all (control nurseries plus all intervention nurseries for whom all core components of the programme were rated red, score of 0); and

- Settings who deliver the programme with **good fidelity** (defined as the nursery being rated an amber score of 1 or a green score of 2 in all of the core aspects of the programme (minimum score of 8 and all components scoring at least 1)) vs settings who deliver **no intervention or deliver with poor fidelity** (control nurseries plus all intervention nurseries for whom at least one core component of the programme is rated red score of 0).

Results for the first stage (of the 2SLS process) will be reported alongside i) the correlation between the instrument and the endogenous variable; and, ii) a F test.

Intra-cluster correlations (ICCs)

The ICC associated with nursery for the primary outcome (both pre and post-test) will be presented alongside a 95% CI. The ICC at post-test will be computed for the primary analysis model, and also for an empty model (i.e., one without covariates). The ICC at pre-test will be calculated for a linear model with pre-test as the outcome and setting as a random effect.

Effect size calculation

Effect sizes will be calculated by dividing the adjusted mean difference between the intervention and control group (accounting for prior attainment and the minimisation factors) by the pooled unconditional standard deviation obtained from the model run without these covariates. A 95% CI for the effect size will be calculated by dividing the 95% confidence limits for the adjusted mean difference by this same denominator. All parameters used in these calculations will be provided in the final report.

$$\mathrm{ES} = \frac{(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}}{sd_{pooled}}$$

where, $(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}$ denotes the difference in means between trial groups adjusting for pre-test score and the minimisation factors, from the multilevel analysis model; and sd_{pooled} denotes the pooled, unconditional standard deviation of the two groups (square root of the sum of the within- and between-cluster variances).

Longitudinal follow-up analyses

The longitudinal analysis will involve accessing participating children's EYFSP data via the NPD, to determine if the Maths Champions programme, administered to nursery children (ages 3-4 years old) had any longer-term effects at the end of Reception (4-5 years old).

The analysis will follow the EEF's (2019b)[2] most recent published guidance on longitudinal analysis of EEF trials. The analysis will consider mathematics, literacy and readiness for school.

[2] Please see the longitudinal analysis guidance.

Key Research Questions and Outcome Measures

The EYFSP is an observational measure completed by teachers. Teachers rate each child's learning and development against 17 early learning goals (ELGs) using the following two levels: meeting the level of development expected at the end of the EYFS (expected); or not yet reaching this level (emerging). For any of the ELGs, a score of 'A' may be reported to indicate that a child has not been assessed.

See:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1024319/Early_years_foundation_stage_profile_handbook_2022.pdf

Mathematics is a specific area of learning measured in the EYFSP, using two ELGs: Number; and Numerical Patterns. A binary measure of whether or not the pupil met 'expected' levels in both of these ELGs will be the primary outcome for the longitudinal analysis and relates to LRQ 1.

Literacy is another learning area measured by the EYFSP, using three ELGs: Comprehension; Word Reading; and Writing. A binary measure of whether or not the pupil met 'expected' levels in all three of these ELGs will be a secondary outcome for the longitudinal analysis and relates to LRQ 2.

'Good level of development' (GLD) is a dichotomous variable (Yes/No) pre-calculated and provided as a single variable in the NPD. Children are defined as having reached a GLD at the end of the EYFS if they have achieved at least the expected level for the ELGs in the prime areas of learning (communication and language; physical development; and personal, social and emotional development) and the specific areas of mathematics and literacy. This will be a secondary outcome for the longitudinal analysis and relates to LRQ 3.

Analyses will be conducted on an intention-to-treat basis, using two-sided significance at the 5% level. Nursery and child-level characteristics and baseline data will be summarised descriptively by randomised group for participants for whom EYFSP data is available. Outcome data will be summarised descriptively for the two groups, for each research question.

We will consider the correlations between EYFSP and measures collected during the main trial (ASPECTS and ASQ-3).

The three dichotomous outcomes will be analysed via mixed-effect logistic regression, adjusted for baseline ASPECTS score (numeracy for maths outcome, language for literacy outcome, and both (separately) for the GLD outcome), and setting-level minimisation factors. The treatment effect expressed as an adjusted odds ratio (OR) will be reported with a 95% CI and p-value. We will also present the unadjusted and adjusted (i.e., predicted, using the postestimation command *margins, dydx(allocation)*) percentage point difference between the two trial arms with a 95% CI (Ge et al. 2011), and convert the adjusted OR (and 95% CI limits) to an estimated Hedges' g effect size using the Cox index as follows (What Works Clearinghouse):

$$d_{Cox} = \omega[ln(OR)]/1.65$$

Where $\omega = \left[1 - {}^{3}/_{(4N-9)}\right]$ and *N* is the total sample size.

In line with the effectiveness trial analyses, subgroup analyses as part of the longitudinal analysis will consider children that were eligible for the EYPP, FEEE at 2 years old, average

number of hours that the child attends nursery and gender. This will only be undertaken for the primary outcome for the longitudinal analysis of mathematics. The subgroup analyses will be conducted by including the factor and an interaction term between the factor and allocation in the primary longitudinal analysis model. We shall also repeat the primary longitudinal analysis within the subset of participants eligible for EYPP.

References

Altman D G, Bland J M. Treatment allocation by minimisation BMJ 2005; 330 :843 doi:10.1136/bmj.330.7495.843

Borm GF, Fransen J, Lemmens WA. A simple sample size formula for analysis of covariance in randomized clinical trials. J Clin Epidemiol. 2007 Dec;60(12):1234-8. doi: 10.1016/j.jclinepi.2007.02.006. Epub 2007 Jun 6. PMID: 17998077.

Chen, J.-Q., McCray, J., Adams, M. and Leow, C., 2014. A survey study of early childhood teachers' beliefs and confidence about teaching early math. Early Childhood Education Journal, 42(6), pp.367–377.

Dunn, G., Maracy, M. and Tomenson, B., Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. Statistical Methods in Medical Research, 2005. 14(4): p. 369-395.

Ge M, Durham LK, Meyer RD, Xie W, Thomas N. Covariate-Adjusted Difference in Proportions from Clinical Trials Using Logistic Regression and Weighted Risk Differences. Drug Information Journal. 2011;45(4):481-493. doi:10.1177/009286151104500409

Robinson-Smith, L., Fairhurst, C., Stone, G., Bell, K., Elliott, L., Gascoine, L., Hallett, S., Hewitt, C., Hugill, J., Torgerson, C., Torgerson, D., Menzies, V. and Ainsworth, H., 2018. *Maths Champions: Evaluation report and executive summary*. [online] London: Education Endowment Foundation. Available at: <https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Maths_champions_evaluation_report.pdf> [Accessed 16 Mar. 2022].

Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987.

Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015;44(3):1051-1067. doi:10.1093/ije/dyv113

Saghaei, M. and Saghaei, S. (2011) Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials. Journal of Biomedical Science and Engineering, 4, 734-739. doi: 10.4236/jbise.2011.411090. Available at: http://www.scirp.org/journal/PaperInformation.aspx?PaperID=8518

Squires, J., & Bricker, D. (2009). Ages & Stages Questionnaires®, Third Edition (ASQ®-3): A Parent-Completed Child Monitoring System. Baltimore: Paul H. Brookes Publishing Co., Inc.

What Works Clearinghouse (n.d). Procedures Handbook, Version 4.0, p.13-14: https://ies.ed.gov/ncee/wwc/docs/referenceresources/wwc_procedures_handbook_v4.pdf