Concept Cat: A two-armed cluster randomised controlled trial
Statistical Analysis Plan

Evaluator (institution): RAND Europe
Principal investigator(s): Elena Rosa Speciani, Miguel Subosa, Louise Tracey, Erin Dysart

Template last updated: August 2019

PROJECT TITLE[1]	Concept Cat: A two-armed cluster randomised controlled trial
DEVELOPER (INSTITUTION)	Better Communication CIC
EVALUATOR (INSTITUTION)	RAND Europe, University of York
PRINCIPAL INVESTIGATOR(S)	Elena Rosa Speciani, Miguel Subosa, Louise Tracey, Erin Dysart
SAP AUTHOR(S)	Fin Oades, James Merewood, Elena Rosa Speciani, Miguel Subosa, Bhavya Singh.
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at setting level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	3-4, Early Years Foundation Stage 1
NUMBER OF SCHOOLS	89
NUMBER OF PUPILS	Between 4 and 15 children per setting, a mean of 12 children per setting, or 1,040 children in total
PRIMARY OUTCOME MEASURE AND SOURCE	*Basic Concepts* sub-test score from the Clinical Evaluation of Language Fundamentals® Preschool-2 (CELF Preschool-2)
SECONDARY OUTCOME MEASURE AND SOURCE	1. Early Years Toolbox Early Numeracy task 2. *Concepts and Following Directions* sub-test score from the CELF Preschool-2

SAP version history

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

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

Table of contents

Contents

Introduction

Trial overview

This efficacy trial is being conducted to evaluate whether the Concept Cat intervention leads to improvements in early conceptual vocabulary among Early Years (EY) pupils aged 3 to 4. We also intend to conduct sub-group analyses on: (a) pupils who are eligible for the Early Years Pupil Premium (EYPP); (b) pupils identified as speaking English as an additional language (EAL); and (c) pupils identified as having special education needs and disability (SEND). Concept Cat is a whole-class intervention, developed by the founders of Thinking Talking, Stephen Parsons and Anna Branagan.

Based on the results of the pilot study (Hopkins et al., 2022), RAND Europe, in collaboration with the University of York, was provided funding from the EEF to conduct an efficacy trial of the Concept Cat programme; this evaluation forms part of the Department for Education's (DfE's) Early Years Education COVID-19 Recovery Package. Whilst the programme is already widely delivered – with approximately 300 Early Years practitioners trained in the teaching methodology each year – there has yet to be a sufficiently robust evaluation of its efficacy as an Early Years intervention.

Intervention

Concept Cat is a whole-class intervention, targeting children aged between three and four years old, that seeks to facilitate the acquisition of key early verbal concepts. In turn, the acquisition of these early verbal concepts would support the attainment of learning competencies laid out in the Key Stage 1 core science and mathematics curricula. The Concept Cat methodology offers an alternative to the generally unstructured and less explicit way these core concepts are taught in standard practice. Moreover, the Concept Cat approach, which is embedded in daily practice, offers a combination of explicit and implicit teaching of concept words. The programme is delivered over a full academic year (i.e., approximately 30 weeks), with a new key word introduced each week. The words or 'concepts' are selected from a list devised by the authors and based on the work of Ann Locke (Locke, 1985).

The overall structure of the intervention is based on the STAR (Select, Teach, Activate, Review) methodology, where a concept is 'selected', 'taught', then implicitly 'activated' through play and home-based activities, and subsequently 'reviewed' to encourage a deeper understanding of its meaning (Blachowicz & Fisher, 2015). The programme incorporates four core components in its implementation. These are a whole-class introduction to the word, meaningful play sessions, parent-child tasks, and a whole-class review.

The 'whole-class introduction' serves as the foundational step in establishing the key verbal concept each week, constituting the 'explicit teaching' component of the programme delivered through a multi-sensory approach. The lead practitioner introduces a specific word, complemented by a unique visual symbol and a physical gesture. Phonology and repetition are integrated to reinforce the children's memory of the word's phonic characteristics. Following this, 'meaningful play sessions' offer an opportunity for 'implicit teaching' to further solidify the learning from the whole-class introduction. During these sessions, children encounter the word within planned play situations. Additionally, the word is introduced to the children's families, and 'home-based tasks', which involve families using target words whilst engaging is meaningful activities, are recommended. Towards the end of the week and in subsequent weeks, a 'whole-class review' of words learned in previous sessions is also conducted, incorporating various activities with Concept Cat props like the word bag and picky puppet. This review ensures that children encounter the word not only during its focus week but also later in the academic year, reinforcing the knowledge acquired throughout the entire intervention delivery period.

Children with the most limited language proficiency receive extra support during the delivery of the intervention, involving additional modelling of the target words, Concept Cat story, and structured play opportunities. In the context of this evaluation, these children needing additional support are termed 'focus children.' To qualify as a focus child, the following criteria must be met:

- Aged 3-4 years old
- Able to sit and respond to an adult-led task for a few moments
- Uses fewer words and shorter sentences than other children of the same age
- Not a child who doesn't speak much at nursery who speaks fluently at home
- If the child speaks English as an Additional Language (EAL), then they must also have delayed language development in their home language(s)

All staff in treatment settings deliver the programme, with lead practitioners undergoing a three-hour remote training session and other EY staff undergoing a one-hour remote training session provided by Better Communication. Each setting receives seven in-setting support visits from Concept Cat coaches. During the second visit, setting staff, supervised by coaches, assess six children chosen by practitioners to determine the appropriate level of concept words for the setting. Coaches also model the Teach element. Subsequent sessions include additional modelling of Teach, Activate, and Review, providing further support on Concept Cat's implementation. These sessions ensure practitioners are implementing the program as intended and tailored to the setting. Alongside coach support, lead practitioners are encouraged to attend six group support sessions throughout the academic year to share experiences and enhance delivery.

The delivery of the intervention started in the last week of September 2023, and will run for 30 weeks (with a new word being taught in each of these weeks).

Eligibility of schools

As detailed in the trial protocol, settings were recruited into the study with support from Better Communication CIC, with selection of settings from three EY Stronger Practice Hubs (SPHs):

- HEART (West Midlands)
- Bright Futures (Trafford)
- Liverpool City Region and Beyond (Everton).

These three EY Stronger Practice Hubs span across three regional areas in England: the West Midlands, Trafford, and Everton. However, the stratification process used for randomisation split settings in the West Midlands into north and south. Delivery regions were selected by taking into account both the delivery team's regional distribution of coaches and the Stronger Practice Hubs that expressed interest in promoting the programme.

Settings were eligible to apply to participate in the trial if they met the following criteria:

- Settings expect to have a minimum of 15 children aged 3–4 (in Foundation 1) enrolled to attend for at least 15 hours a week in the academic year 2023/2024.
- Settings complete all baseline measures (surveys and child consent/data) and facilitate assessments within their setting prior to randomisation.
- Settings agree to participate fully in the evaluation, including completing the programme (as outlined above) if selected to be in the intervention group and completing all evaluation requirements (both control and intervention) in the academic year 2023/2024.
- Settings have not implemented Concept Cat within the last 2 years, or settings that participated in wave 1 of the Concept Cat pilot.

- Settings do not have any staff employed who have attended Word Aware Early Years training within the last three years.
- Settings have not accessed Concept Cat resources through Lift Lessons (liftlessons.co)

Design overview

This section provides an overview of the study design for this efficacy trial, which is being carried out by RAND Europe, in partnership with the University of York. As detailed in Table 1 below, the trial is a two-arm, waitlisted, cluster randomised controlled trial which primarily assesses the impact of the Concept Cat teaching methodology on early conceptual vocabulary development among children aged 3 to 4 in Early Years education.

Table 1 provides an overview of the trial design, specifying the key stratification variables used in sampling, and the outcomes being assessed.

Table 1: Trial Design Overview

Trial design, including number of arms		**Two-arm, waitlisted, cluster randomised controlled trial**
Unit of randomisation		Early Years settings
Stratification variables (if applicable)		1) Setting type (Private, voluntary, or independent [PVI] vs. school-based settings [SBS]); 2) Region (Northern West Midlands, Southern West Midlands, Trafford, Everton)
Primary outcome	variable	Early Conceptual Vocabulary
	measure (instrument, scale, source)	'Basic Concepts' subtest from Clinical Evaluation of Language Fundamentals Preschool-2 UK (CELF-Preschool 2 UK). These will be used as raw scores (min = 0; max = 18).
Secondary outcome(s)	variable(s)	1) Early Conceptual Vocabulary 2) Early Numeracy
	measure(s)	1) 'Concepts and Following Directions' subtest from Clinical Evaluation of Language Fundamentals Preschool-2 UK (CELF-Preschool 2 UK). These will be used as raw scores (min = 0; max = 22). 2) Early Years Toolbox (EYT) Early Numeracy task. These will be used as raw aggregate scores (min = 0; max = 85).
Baseline for primary and secondary outcome	variable	Early Conceptual Vocabulary
	measure	'Basic Concepts' subtest from Clinical Evaluation of Language Fundamentals Preschool-2 UK (CELF-Preschool 2 UK).

Given that the intervention has a whole-class focus, randomisation occurred at the setting level, with each setting being allocated to either a group that receives the Concept Cat intervention (the treatment group), or a group that receives business as usual (the control group). Randomisation was stratified according to region so that each region had settings delivering Concept Cat whilst also ensuring that the delivery team had an appropriate number of settings per trainer in each region. As presented in

table 1, the four regional areas used for stratification were: Northern West Midlands, Southern West Midlands, Trafford, and Everton. All settings situated in the West Midlands are part of the HEART stronger practice hub, all settings situated in Trafford are part of the Bright Futures hub, and all settings situated in Everton are part of the Liverpool City Region and Beyond hub. Stratification was also based on setting type (i.e., PVI or SBS)[2] to ensure a similar representation of each type of setting in each region. Having a balance of both setting types in the treatment and control group ensures that findings from the trial were applicable to all setting types.

The primary outcome being measured is early conceptual vocabulary, as operationalised by the 'Basic Concepts' subtest from the Clinical Evaluation of Language Fundamentals Preschool-2 UK (CELF-Preschool 2 UK). This trial also seeks to evaluate the impact of receiving the intervention on two secondary outcomes: an alternative measure of early conceptual vocabulary, operationalised by 'Concepts and Following Directions' subtest from CELF-Preschool 2 UK; and early numeracy, operationalised by the Early Years Toolbox (EYT) Early Numeracy assessment (ENA). For all outcomes, scores from the 'Basic Concepts' subtest from CELF-Preschool 2 UK will be used as the baseline assessment. For more information on the outcomes used, please see the Outcome Measures section.

Settings allocated to the treatment group received Concept Cat training and are expected to deliver the Concept Cat programme during the academic year 2023/2024, while those allocated to control are expected to carry on with business as usual until the following academic year (2024/2025) when they will receive training and support to deliver Concept Cat. All settings (i.e., regardless of assignment to the treatment or control group) are provided incentives in two tranches: £200 on completion of baseline assessments and a further £200 on completion of all endline assessments. These funds are to be used at the discretion of the setting and could be used to buy an intervention programme of their choice once the trial ended.

Randomisation was conducted by the RAND evaluation team on the 20th of September 2023. While EEF guidance suggested collecting baseline measures before randomisation (EEF 2022), because of the tight timelines, settings that had booked testing were randomised before they had all completed testing (but after they had booked their testing). The randomisation was blinded and conducted using STATA software, with the code used for this randomisation provided in the Appendix. The results of the randomisation were shared with the delivery team so they could organise staff allocation , but settings were not informed of their allocation until they completed testing. To mitigate against potential attrition, only settings that had booked a date for baseline testing, had shared pupil data, and had signed and returned a DSA were eligible for randomisation.

Sample size calculations overview

The initial power calculations documented in the evaluation Protocol (Oades et al., 2023) were based on both information provided in the EEF's Invitation to Tender, as well as on subsequent meetings with EEF. Power calculations and minimum detectable effect size calculations were performed using the *PowerUp!* tool (Dong & Maynard, 2013).

The MDESs at the protocol stage were calculated using a two-level random assignment, based on the assumption of equal allocation of settings to intervention and control groups. These initial MDES calculations also assumed that the average number of children per setting was 15, and that the average number of EYPP-eligible children per setting was 3. However, upon collection of pupil baseline data, these MDES calculations have been updated to reflect the actual trial sample, with an average total setting size of 12 children (minimum of 4; maximum of 15), and an average of 2 EYPP-eligible children per setting.

The assumed desired power of 0.8, alpha of 0.05, and normally distributed primary outcome variable were maintained from the protocol stage. We have assumed a high pre-test/post-test correlation of

[2] PVI (Private, voluntary or independent setting); SBS (School-based setting)

0.75. The CELF-P2 has a published test-retest correlation of 0.95 for receptive vocabulary (Eadie et al., 2014), but we have made our estimate more conservative for two reasons. Firstly, the retest in this particular instance was administered between two and 24 days after the initial assessment. Furthermore, in the NELI effectiveness trial, we found pre-test/post-test correlations of 0.75 using another version of the CELF (Dimova et al., 2020). Table 2 displays the results of these power calculations on our analytical sample at randomisation, generating an MDES of 0.253 for the overall sample, and an MDES of 0.346 for the EYPP-eligible sub-sample.

Table 2: MDES calculations overview, at Protocol stage and at Randomisation

		Protocol		Randomisation	
		OVERALL	EYPP	OVERALL	EYPP
Minimum Detectable Effect Size (MDES)		0.249	0.312	0.253	0.346
Pre-test/ post-test correlations	level 1 (pupil)	0.75	0.75	0.75	0.75
	level 2 (setting)	0.15	0.15	0.15	0.15
Intracluster correlations (ICCs)	level 2 (setting)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		15[3]	3	12	2
Number of schools	intervention	45	45	45	45
	control	45	45	44	44
	total	90	90	89	89
Number of pupils	intervention	675	68	527	78
	control	675	68	513	78
	total	1350	136	1040	156

It is important to note that we are still in the process of collecting baseline data from settings. As detailed in Table 4 in the 'Imbalance at Baseline' section, we are still missing EYPP data from a number of settings, and the MDES calculations presented in the table above are subject to change following the completion of this data collection. Whilst the power calculations presented in Table 4 retain assumptions about the ICC from protocol stage, the impact team also recalculated the MDES using the scores collected at baseline. From these baseline scores, we found the intra-cluster correlation (proportion of variance in outcome explained by setting-level differences) to be 0.11, generating an MDES of 0.225.

Outcome measures

Baseline measures

Baseline assessment for pupils in EY settings participating in the trial will consist of one measure for both primary and secondary outcomes:

[3] Based on the number of children that can be visited over the course of two days of testing.

1) **Early Conceptual Vocabulary** operationalized using the 'Basic Concepts' subtest from Clinical Evaluation of Language Fundamentals Preschool-2 UK (CELF-Preschool 2 UK).

Baseline testing was conducted using the same measure as endline, maximising the precision of the impact estimate for the primary outcome and accounting for imbalance at baseline. The 'Basic Concepts' subtest is one of seven subtests within the CELF Preschool-2, a standardised, individually administered assessment of expressive and receptive linguistic ability specifically designed for children aged three to six. It is widely used in EY outcome assessments (for more detail, see *Primary outcome measures* section). We note that whilst the use of the 'Basic Concepts' subtest as baseline for Early Numeracy (see secondary outcomes) will likely reduce the pre- and post-test correlations for this outcome, we expect there to be sufficient overlap between the latent outcomes being measured by each test for it to be a suitable baseline when modelling all outcome measures. The decision to not to collect baseline scores for the Early Years Toolbox Early Numeracy Assessment (ENA) was taken to both avoid over-burdening settings and ensure that baseline testing could be completed within 5 weeks to facilitate full delivery of the intervention.

Settings were asked to provide the evaluation team with a list of all eligible pupils. Elklan were provided with these pupil lists and a protocol for random ordering.[4] On the day of testing, the testers used the randomly ordered number lists and the pupil list to test children in a specific order. This ensured that there was no selection bias in who was tested on the day. Testers used this approach until a maximum of 15 children were selected for testing. Selection for testing was informed by child-level eligibility criteria, namely that the child would be expected to be in attendance for the full 30-week delivery period of the intervention.

Baseline assessment testing was carried out by Elklan in September 2023, prior to randomisation. This testing was performed, on behalf of Elklan, by independent test administrators – all of whom were qualified speech and language therapists. Test administrators were trained in the use and administration of CELF Preschool-2, including how to conduct practice sessions. Data was initially collected using paper-based sheets as per standard delivery of CELF Preschool-2 before being uploaded by the test administrators to a secure portal to facilitate continual quality assurance by both Elklan and the evaluation team.

Primary outcome measures

Because this is an efficacy trial, we will only explore the impact of the intervention on one primary outcome, as requested by the EEF (EEF, 2022). The primary outcome of interest, in the assessment of the efficacy of the Concept Cat intervention, is the same as the baseline measure:

1) **Early Conceptual Vocabulary** operationalized using the 'Basic Concepts' subtest from Clinical Evaluation of Language Fundamentals Preschool-2 UK (CELF-Preschool 2 UK).

Post-test assessment of participating pupils' early conceptual vocabulary will be conducted after the full delivery of the programme in June 2024. The primary outcome being assessed following implementation is the same subtest from the CELF-Preschool 2 UK. This is the same primary outcome used in the pilot study (Hopkins et al., 2022). The CELF Preschool-2 is a standardised, individually administered assessment of expressive and receptive linguistic ability specifically designed for children aged three to six. It is widely used in EY outcome assessments. The CELF Preschool-2 consists of seven subtests, including Basic Concepts. 2 The CELF Preschool-2 was judged fit for purpose since: (a) it is designed to be brief (taking around five to seven minutes to administer per subtest); (b) it directly measures receptive vocabulary development; (c) it is UK norm-referenced; (d) and it has strong psychometric properties, with test-retest reliability ranging from 0.77 to 0.96 (for ages 3 to 3,11) and 0.74−0.95 (for ages 4 to 4:11; EEF, 2023).

[4] This was a serries of numbered lists based on setting size, with numbers being presented in random order for the given number of eligible pupils in a setting.

The 'Basic Concepts' subtest measures a child's knowledge of four fundamental concepts: dimension and size; direction, location, and position; number and quantity; and quantitative equality. The testing process requires the child to respond to a description of a concept provided by selecting a picture from a set of options that they believe best corresponds with or exemplifies said concept. The score the child receives at the end of the test is equal to the number of correctly identified concepts, and the test is terminated after five consecutive incorrect responses. The primary analysis will utilise the raw scores of this subtest (scored from 0 to 18).

We note that the pilot found slight ceiling effects in the Basic Concepts subtest at baseline and understand that this may be attributable to design factors, such as the sampling strategy used (e.g., a slightly older pupil group than the efficacy trial is targeting, lower-than-average number of pupils with English as an Additional Language) (Hopkins et al., 2022). The distribution of the CELF-P2 Basic Concepts scores at baseline among pupils in our trial is displayed in Figure 1 in the Appendix, and display a significant left skew in the distribution of these scores. It is plausible that this skew may become more pronounced at endline, producing ceiling effects in the outcome and potentially leading to underestimation of the treatment effect.

We will employ three approaches to assess whether the outcome demonstrates a ceiling effect. Firstly, we will visually inspect the distribution of endline scores to identify any clustering at the upper end of the scale. This examination will be complemented by Pearson's coefficient of skewness, providing insight into the extent of deviation from a normal distribution. Additionally, we will analyse the proportion of children achieving scores within 1 standard deviation of the maximum score. Uttl (2005) found that ceiling effects may significantly increase the likelihood of Type II errors when 25% of outcome scores fall within 1 standard deviation of the maximum value for tests consisting of 9 to 15 items. We will therefore employ this cut-off when exploring the possibility of a ceiling effect in our endline data. By integrating these three approaches, we aim to gauge the likelihood of a ceiling effect in the primary outcome, guiding our approach to the main analysis. Our proposed analysis in the presence of ceiling/floor effects is detailed in the Additional Analysis section.

Secondary outcome measures

The impact of the Concept Cat intervention on two secondary outcomes is also being captured in this trial:

1) **Early conceptual vocabulary,** alternatively operationalized by the 'Concepts and Following Directions' subtest from CELF-Preschool 2 UK.
2) **Early numeracy,** as measured using the Early Years Toolbox (EYT) Early Numeracy Assessment (ENA).

These are outlined further below:

1) Early conceptual vocabulary

The first secondary outcome being evaluated in this efficacy trial offers an alternative measure for early conceptual vocabulary, as operationalised by the 'Concepts and Following Directions' subtest from CELF-Preschool 2 UK. This subtest measures conceptual vocabulary by evaluating a child's ability to: (a) understand spoken directions containing concepts that require logical operations; (b) remember names, orders and characteristics of items mentioned; and (c) identify the target from among several choices. A key distinction between this subtest and the 'Basic Concepts' subtest is that it is more explicitly linked to vocabulary, rather than solely measuring a child's understanding of the concepts themselves. Given Concept Cat's specific focus on vocabulary, the latent constructs being measured by this sub-test were deemed adequately proximal to those targeted by the intervention. The secondary analysis will utilise the raw scores of this subtest (scored from 0 to 22).The frequency distribution for CELF-P2 'Concepts and Following Directions' subtest is displayed in Figure 2 in the Appendix. This distribution points towards a significant right skew in this variable, with bunching of child scores at the lower end of the scale. As with the primary outcome in relation to ceiling effects, we will explore the possibility that this secondary outcome is exhibiting a floor effect,

and may re-estimate the secondary outcome model using a tobit regression. This is further detailed in the Additional Analysis section.

2) Early numeracy

As outlined in the program's theory of change, Concept Cat specifically concentrates on instructing fundamental verbal concepts crucial to the mathematics and science curriculum. In doing so, it seeks to improve future attainment in Key Stage 1 mathematics and science. Recognizing the central role of mathematics and science in the intervention's desired outcomes, early numeracy will be incorporated as a secondary measure to assess Concept Cat's efficacy. This will be measured using the EYT-ENA, a set of eight short game-like tasks covering five distinct skill domains using a maximum of 85 items:

- **number sense**, which pertains to early numerical concepts and language (12 items) and rapid quantitative comparison (6 items);
- **cardinality and counting**, which refers to counting a subset of items (6 items), identifying digits and quantities (6 items), matching digits and quantities (6 items), completing number sequences (6 items), discerning the relative position of digits based on their quantity (6 items), and identifying the ordinal position of an object with respect to other objects in a line (6 items);
- **numerical operations**, which measures a child's ability to derive information from a basic, verbal mathematical problem (6 items) and solving basic numerical equations (6 items);
- **spatial and measurement constructs**, which assesses a child's ability to understand spatial and measurement concepts, such as length, size, and geospatial relations (13 items); and
- **patterning**, which refers to children's ability to discern and complete increasingly complex patterns (6 items).

These five skill domains seek to measure fundamental numeric abilities found to have an important role in shaping social, emotional, cognitive, and life outcomes (Dawson et al., 2020). The tests are administered using iPad-based assessment tools suitable for use with children in Early Years settings by Early Years practitioners, with all eight tasks taking approximately 5 minutes to administer. The ENA has been found to demonstrate good psychometric properties, yielding a test-retest reliability of 0.89 (Howard et al., 2022).

The overall ENA score is the sum of a child's correct responses to each item, giving a maximum score of 85. Whilst the ENA allows for discrete categorisation of scores across different skill domains, this evaluation will use the overall score in aggregate.

Analysis

Primary outcome analysis

The following section is informed by the EEF's analysis guidance for efficacy trials (EEF, 2022).

As detailed in the protocol, this efficacy trial has one primary research question:

RQ1. What is the difference in early conceptual vocabulary development, measured by the Basic Concepts subtest of the Clinical Evaluation of Language Fundamentals Preschool-2 UK (CELF-Preschool-2), of pupils in settings receiving Concept Cat in comparison to those pupils in control settings receiving business as usual?

To address RQ1, the primary outcome for this efficacy trial will be pupils' early conceptual language development, as assessed by the Basic Concepts subtest of CELF Preschool-2 scores. As detailed in the outcomes section, the CELF Preschool-2 is a standardised, individually administered assessment of expressive and receptive linguistic ability specifically designed for children aged three to four.

To address the primary research question, an intention-to-treat (ITT) analysis will be undertaken using multi-level modelling (MLM) with fixed effects and random intercept. This will assume a consistent average treatment effect across schools, in line with the whole-class delivery of the intervention, but will account for the school-based variation in the mean outcome pre- and post-intervention. As the analysis of Concept Cat's efficacy will be based on an ITT principle, data is analysed according to the group randomised, regardless of whether the treatment was received as intended, and irrespective of withdrawal from the intervention post-randomisation, or deviations in programme implementation (Torgerson & Torgerson, 2008). This principle is key in ensuring an unbiased analysis of 'real world' intervention effects and is in line with the EEF's guidance (2022).

The EEF (2022) recommend clustering using the unit of randomisation. To account for the nested nature of the data, a hierarchical linear model with two levels (setting, pupil) will be fitted controlling for pre-intervention scores at the pupil level. This will allow for the potential school-level heterogeneity in the expected impact of the intervention. Following the stratified randomisation, the model will also include terms identifying the regions and PVI status (strata).

Impact will be estimated by fitting the model in equation (1). Equation (1) is known as a 'random intercepts' model because $\beta_{0j}=\beta_0+u_j$ is interpreted as the school-specific intercept for school j and $\beta_{0j} \sim_{i.i.d} N(\beta_0, \sigma_u 2)$ is random (it is a number that can take any value):

$$(1) \qquad Y_{ij} = \beta_0 + \beta_1 \mathrm{CONCEPTCAT}_j + Z_j\beta_2 + X_{ij}\beta_3 + u_j + e_{ij}$$

Where:

Y_{ij}= 'Basic Concepts' subtest scores from CELF-Preschool 2 UK for child $_i$ in setting $_j$ at endline;

β_0 = the cluster-level coefficient for the slope of a predictor on Early Conceptual Vocabulary;

CONCEPTCAT$_j$ = a binary indicator of the setting assignment to intervention [1] or control [0];

β_1 = the individual-level coefficient denoting the estimated effect of assignment to treatment on the primary outcome.

Z_j = school-level characteristics i.e. the stratifying variables of geographical location and PVI status (as used for randomisation);

X_{ij} = pupil level characteristics for pupil $_i$ in school $_j$, including the 'Basic Concepts' subtest from CELF-Preschool 2 UK score at baseline to reduce bias in estimates (EEF, 2022);

u_j = setting-level residuals and

e_{ij}= individual-level residuals.

The coefficient β_1 in Equation 1 above will represent the outcome of the trial, with respect to the primary outcome measure. Equation 1 will also be replicated for each of the two secondary outcome measures (see secondary outcome analysis section). If we expect early conceptual vocabulary scores to be age-correlated, this may introduce bias into the measurement of β_1. As an additional sensitivity analysis, we will run the regressions with age (in months) included as a covariate in X_{ij}.

The inclusion of baseline scores for the CELF-Preschool 2 'Basic Concepts' subtest as a control variable allows us to control for prior attainment, providing a more conservative and comparable method that will be usable across potential future EEF trials (EEF, 2022).

The effect size (Hedges' *g*) will be calculated for β_1 (see Effect Size Calculation section). The effect size will be standardised using unconditional variance in the denominator and confidence intervals will be reported to communicate statistical uncertainty as 95% confidence intervals, in line with EEF guidance

(EEF 2022). This will tell us the average effect of the intervention on pupil outcomes in treatment schools compared to those in control schools.

In addition to the primary analysis model specified above, we will also provide means and standard deviations of pre-and post-test scores for the 'Basic Concepts' subtest. The distributions of these scores will be summarised using histograms.

All analyses will be run in either Stata (versions 18 onwards) or R.

Secondary outcome analysis

As detailed in the protocol, this efficacy trial has two secondary research questions:

RQ2. What is the difference in early conceptual vocabulary development, measured by the Concepts and Following Directions subtest of the CELF-Preschool 2, of pupils in settings receiving Concept Cat in comparison to those pupils in control settings receiving business-as-usual?

RQ3. What is the difference in early numeracy development measured by the Early Numeracy Assessment (ENA) of the Early Years Toolbox (EYT) of pupils in settings receiving Concept Cat in comparison to those pupils in control settings receiving business-as-usual?

As detailed in the Outcomes section, we aim to answer these research questions through exploring the impact of the Concept Cat intervention on two secondary outcome measures: early conceptual vocabulary (measured by the 'Following Directions' subtest of CELF-Preschool 2) and early numeracy (measured by the EYT-ENA).

For both secondary outcomes, the secondary analysis will follow the same procedures from the primary analysis, using Equation (1) to estimate each respective secondary outcome model, substituting the primary outcome variable in turn with each of the secondary outcome variables.

For the secondary outcome analysis for early conceptual vocabulary and early numeracy, the X_{ij} vector for pupil $_i$ in school $_j$ will again be represented by the baseline scores for the *Concepts and Following Directions* subtest of CELF-Preschool-2. Whilst neither early conceptual vocabulary nor early numeracy are age-standardised, in the main analysis we will not include age in the pupil-level characteristics, X_{ij}, as the trial is within one year group, somewhat minimising concerns over age effects. Again we will run the regressions with age (in months) included as a covariate in X_{ij} as an additional sensitivity analysis, accounting for the potential bias introduced by ENA composite scores being correlated with age.

To reduce the likelihood of Type I errors arising from the use of multiple secondary outcomes, we will conduct a testing correction for our secondary analysis. We will use the Romano-Wolf correction, given that it accounts for potentially correlated multiple outcomes, and the dependence structure of test statistics, through bootstrap (or permutation) resampling from the original data (Clarke et al., 2019). Whilst this method is more analytically and computationally intensive than the Bonferroni correction, it has been found to provide stronger control against the family-wise error rate, especially when the multiple outcomes are correlated (IBID).

This testing correction process will either be performed using the rwolf package in STATA, or the crctStepdown[5] package in R.

[5] crctStepdown is a bespoke package developed by Watson, Akinyemi, and Hemming (2023): Permutation-based multiple testing corrections for P$$ P $$-values and confidence intervals for cluster randomized trials - Watson - 2023 - Statistics in Medicine - Wiley Online Library

Sub-group analyses

As detailed in the trial protocol, this efficacy study has three further research questions pertaining to Concept Cat's impact on particular sub-groups:

RQ1a. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of Early Years Pupil Premium/Free Early Education Entitlement (EYPP/FEEE)-eligible pupils, compared to non-EYPP/FEEE-eligible pupils?

RQ1b. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of pupils with English as an Additional Language (EAL), compared to non-EAL pupils?

RQ1c. What is the impact of the Concept Cat teaching methodology on the early conceptual vocabulary development of pupils with Special Educational Needs or Disability (SEND), compared to non-SEND pupils?

We will employ two main approaches to sub-group analysis.

The first approach will incorporate re-running the model used in the primary analysis, but on a restricted sample of children identified as being a member of the relevant subgroup. In the case of EYPP-eligible children (RQ1a), a key target group for the EEF, the study team will conduct sub-group analysis for pupils eligible for these funding provisions, using the same model used for the primary analysis (Equation 1), but where the analytical sample only contains pupils from this sub-group. Information on pupil eligibility was obtained from settings themselves in the baseline data collection phase, and analysis will be undertaken using a binary EYPP variable, where EYPP-eligible = 1; non-EYPP/FEEE-eligible = 0.

We will also conduct further analysis by introducing an interaction term to the primary outcome model between the binary EYPP-eligibility indicator and treatment assignment. This will allow us to explore the potential differential effects of Concept Cat on EYPP/FEEE-eligible pupils. This will allow us to account for the EPYY-eligible subgroup and treatment allocation while retaining the whole analytical sample in the model.

The mediating impact of EYPP-eligibility will be estimated by fitting the model in equation (2). Equation (2) is known as a 'random intercepts' model because $\beta_{0j}=\beta_0+u_j$ is interpreted as the school-specific intercept for school j and $\beta_{0j} \sim_{i.i.d} N(\beta_0, \sigma_u 2)$ is random (it is a number that can take any value):

$$(2) \qquad Y_{ij} = \beta_0 + \beta_1 \mathrm{CONCEPTCAT}_j + \beta_2 \mathrm{EYPP}_{ij} + \beta_3 (\mathrm{CONCEPTCAT}_j * \mathrm{EYPP}_{ij}) + \beta_4 Z_j + \beta_5 X_{ij} + u_j + e_{ij}$$

Where:

Y_{ij} = 'Basic Concepts' subtest scores from CELF-Preschool 2 UK for child $_i$ in setting $_j$ at endline;

β_0 = the cluster-level coefficient for the slope of a predictor on Early Conceptual Vocabulary;

CONCEPTCAT$_j$ = a binary indicator of the setting assignment to intervention [1] or control [0];

β_1 = the individual-level coefficient denoting the estimated effect of assignment to treatment on the primary outcome.

CONCEPTCAT$_j$ *EYPP$_{ij}$ = the interaction term between assignment to treatment and EYPP-eligibility

β_2 = the individual-level coefficient denoting the mediating effect of EYPP-eligibility on the impact of assignment to treatment on the primary outcome.

Z_j = school-level characteristics i.e. the stratifying variables of geographical location and PVI status (as used for randomisation);

X_{ij} = pupil level characteristics for pupil $_i$ in school $_j$, including the 'Basic Concepts' subtest from CELF-Preschool 2 UK score at baseline to reduce bias in estimates (EEF, 2022);

u_j = setting-level residuals and

e_{ij}= individual-level residuals.

Identical analyses will also be conducted with sub-groups of EAL (RQ1b) and SEND (RQ1c) pupils, to ascertain the degree to which Concept Cat may impact both pupils who speak English as an additional language and pupils with special educational needs or disability. For all subgroup analysis, effect sizes and statistical uncertainty will be calculated and communicated as per the primary analysis (also see section on **Error! Reference source not found.**).

Additional analyses

There is significant variability in setting sizes, with the smallest setting included in the trial only having 4 children tested, and the largest having 15 children tested, with an overall mean of 12 children tested per setting. As set out in the protocol, the intended number of children for each setting included in the trial was 15. After baseline testing was complete, we have 28 settings (15 in treatment, 13 in control) out of 89 which have met this threshold. We will therefore carry out additional subgroup analyses on both these settings which met this threshold, and the settings which did not.

A number of settings included in this efficacy trial are part of the same Early Years Federations. In addition to the analysis proposed above, such we will also attempt to explore the possibility of interaction and collaboration between these settings delivering Concept Cat, depending on if the data collected as part of the IPE suggests that interactions have taken place.

As detailed in the Outcome measures section, significant left and right skews have been observed in our baseline scores for our primary and secondary outcomes respectively. This may be particularly problematic for our primary outcome, as it is plausible that scores within the sample will increase from baseline to endline, potentially resulting in a ceiling effect produced by the censoring at the top end of the scale. Ceiling effects can increase bias in model estimates due to insufficient variability at the upper limit of a scale, leading to significant uncertainty around estimated treatment effects. In this case, it would likely lead to an underestimation of the true treatment effect. We will inspect the distribution of endline scores for the 'Basic Concepts' subtest of the CELF-P2 assessment, and depending on the nature of this distributions, we may re-estimate the impact of the Concept Cat intervention on early vocabulary development through the use of a tobit model. Tobit models are designed to re-estimate linear models whilst accounting for censoring in the outcome, allowing for more accurate measurement of associations between an independent variable and a latent outcome ((McBee, 2010)). They combine the use of a probabilistic density function to estimate the relationship between an independent variable and the outcome below the threshold, with a cumulative density function to model the relationship above the threshold (McBee, 2010). The structure of and other covariates used in this model will otherwise be identical to the primary analysis model.

Whilst the right-skewed distribution for the 'Following Directions' subtest is arguably less problematic, given that this skew will plausibly become less pronounced at endline, we will nonetheless similarly inspect the distribution of endline scores, both visually, and with a statistical test of the degree of skewness in the distribution using Pearson's coefficient of skewness. Should a floor effect be observed, we will run a tobit model as an additional sensitivity analysis to re-estimate the treatment effect on this outcome. As with the primary analysis, the structure of and other covariates used in this model will otherwise be identical to the secondary analysis model.

Imbalance at baseline

Whilst a robust randomisation process should generate treatment and control groups with equivalent characteristics at baseline, beyond imbalance occurring by chance (Glennerster & Takavarasha, 2013), balance at baseline for the analytical sample will be tested for at both the setting and pupil level.

As displayed in table 3, we have checked the distribution of the following setting level variables across treatment and control groups through cross-tabulations.

- Proportion of PVI settings;
- Proportion of settings by region
- Proportion of children eligible for EYPP/FEEE;
- Proportion of children identified as EAL;
- Proportion of children identified as SEND

At the pupil, level we will assess balance at baseline for the following characteristics:

- Gender;
- EYPP eligibility;
- SEND status;
- EAL status
- Prior attainment using the CELF Preschool-2 'Basic Concepts' subtest

Table 3: Baseline characteristics of groups as randomised

School-level (Continuous)	**Control group**		**Intervention (Concept CAT) group**	
	n/N (missing)	**Count (%)**	**n/N (missing)**	**Count (%)**
Setting Type				
1. School-based	32/44 (0)	73%	32/45 (0)	71%
2. PVI	12/44 (0)	27%	13/45 (0)	29%
Region				
1. Trafford (Bright Futures SPH)	11/44 (0)	25%	11/45 (0)	24%
2. Everton	11/44 (0)	25%	12/45 (0)	27%
3. West Midlands (HEART North)	12/44 (0)	27%	11/45 (0)	24%
4. West Midlands (HEART South)	10/44 (0)	23%	11/45 (0)	24%
School-level (Continuous)	**n/N (missing)**	**Mean**	**n/N (missing)**	**Mean**

Proportion of female children	44/44 (0)	53.4%	45/45 (0)	53.3%	
Proportion of EYPP/FEEE-eligible children	40/44 (4)	16.7%	40/45 (5)	16.3%	
Proportion of children identified as EAL	44/44 (0)	26.7%	42/45 (3)	28.2%	
Proportion of children identified as SEND	42/44 (2)	6.4%	44/45 (1)	7.3%	
Child-level (categorical)	**n/N (missing)**	**Count (%)**	**n/N (missing)**	**Count (%)**	
Gender					
1. Female	276/513	54%	284/527 (4)	54%	
2. Male	237/513	46%	243/527 (4)	46%	
EYPP					
1. EYPP-eligible	78/513 (41)	17%	78/527 (62)	17%	
2. Not EYPP-eligible	394/513 (41)	83%	387/527 (62)	83%	
SEND					
1. SEND	30/513 (38)	6%	25/527 (21)	5%	
2. Not SEND	445/513 (38)	94%	481/527 (21)	95%	
EAL					
1. EAL	117/513 (18)	24%	129/527 (34)	26%	
2. Not EAL	378/513 (18)	76%	364/527 (34)	74%	
Child-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**	**Effect size [95% CI]**
Age in months	513/513	43.41 (3.53)	527/527	43.54 (3.51)	
CELF Preschool-2 'Basic Concepts' subtest	513/513	11.12 (4.39)	527/527	11.60 (4.19)	0.021 (-0.002 – 0.043)

At the time of writing this SAP, we have complete data on three of the four setting level characteristics (PVI status, EAL status, and SEND status); and two of the three pupil-level characteristics (gender and

age). Table 3 presents the balance between treatment and control groups at randomisation for these available baseline characteristics.

Setting-level characteristics

The balance of both continuous and categorical setting-level characteristics demonstrates effective randomisation. Given that setting type was a stratification variable for randomisation, it inevitably displays balance across the two samples, with 27% and 29% of schools being classed as PVI in control and treatment respectively. It should be noted that absolute balance is not possible due to one of the PVI schools dropping out prior to baseline testing, leading to an odd number of PVI schools in the overall sample. Given that, like PVI status, region was a stratification variable used in in randomisation, the proportional makeup of settings from each of the four regional areas are well balanced between treatment and control groups. The average gender make-up of settings is similarly well balanced between both samples, with the average percentage of female children across settings being 53.4% and 53.3% in control and treatment groups respectively. The average proportion of EYYP-eligible children presents a similar balance at 16.7% and 16.3%; as does the proportion of EAL children at 26.7% and 28.2%. The average proportion of SEND children in settings is indicative of a good level of balance across control and treatment groups at 6.4% and 7.3% respectively.

Child-level characteristics

The distribution of child-level variables appears balanced across treatment and control groups. Gender is evenly distributed between the two groups, with 54% of both the control and intervention groups being male, and 46% being female. Similarly, the proportion of EYPP-eligible children is balanced, with 16.7% and 16.3% being eligible in the control and intervention groups, respectively. The percentage of children classified as SEND is well balanced between the two groups, constituting 5% of the control group and 6% intervention of the groups. Lastly, the proportion of pupils speaking English as a foreign language, whilst marginally less even, is nonetheless similar across both arms, accounting for 24% and 26% of the control and intervention groups, respectively.

The age of children is very well balanced between treatment and control, at 43.54 and 43.41 respectively. Similarly the balance of CELF-P2 Basic Concepts baseline scores is sufficiently even across both treatment and control groups, with means of 11.60 and 11.12 respectively. The standard deviations are similarly consistent across both trial arms, at 4.19 and 4.39 for treatment and control.

We can therefore conclude that randomisation has correctly resulted in sufficient balance in observable characteristics between control and intervention groups.

Missing data

Missing data can arise from a variety of sources and at multiple stages of a trial, including as a result of participant attrition at the setting or pupil level, from errors by test administrators, and from errors in data collection. Whilst the intention-to-treat basis of the investigation requires inclusion of all available data from settings as randomised, the robustness of estimated effect sizes from this analysis may be impeded if we do not have complete endline data for all randomised settings. Should the missingness encountered at endline fall below 5%, the mechanism of missingness will be unlikely to bias estimates and the data will be considered as ‘missing completely at random’ (MCAR). In this scenario, a complete case analysis will be employed. Should missingness encountered for the primary outcome be beyond this 5% threshold, the study team will further investigate the nature of this missingness in accordance with the EEF analysis guidance (EEF, 2022).

As an initial step in the missing data analysis, we will examine attrition across trial arms to evaluate bias. Cross-tabulations will be presented for the proportions of missing values on characteristics available at baseline (see table 3 in the previous section), at both pupil and school levels, and on the primary and secondary outcome measures. This will provide an introductory insight into the patterns in

missingness across treatment and control groups. To establish whether systematic differences exist between those who are 'missing' and those who are not, we will model missingness at follow-up (defined as pupils with missing primary outcome data at endline) as a function of the baseline covariates included in the cross tabulations. This will take the form of a logistic regression model, using a binary outcome variable denoting missingness at endline (where 1=missing; 0=complete), and will mirror the multi-level structure of the models used in the main analysis, with pupils nested within settings.

Should the above analysis reveal systematic patterns to missing primary outcome data, as explained by other covariates, we will compare the results of a complete case primary analysis model to the results of the same model with these additional covariates included. If the results of these two models are similar, the complete case analysis is unlikely to be biased, but the interpretation of these results is conditional on the inclusion of these additional covariates. However, should there be significant differences observed between the results of these two models, it will be likely that the data is MNAR. In this case, the estimated effect sizes produced by our analysis will need to be interpreted with significant caution and further sensitivity analysis will be required (see below). The above approach to data that is MAR employs a complete case analysis, and we acknowledge that should we be presented with significant missingness among any of the additional covariates, the sample size may be compromised, and the statistical power of our estimates could be significantly reduced, resulting in wide standard errors. In this instance, depending on the nature and extent of missingness encountered, we may additionally utilise a multiple imputation (MI) approach, generating replacement values for observations with missing primary outcome data based on characteristics available at baseline. Such imputation will use a fully conditional specification (FCS) method, and averages of at least 100 imputations will be used to calculate revised estimates to reduce standard errors and improve statistical power (Graham et al., 2007).

In accordance with the EEF analysis guidance (2022), multiple imputation is required where covariates included in the primary model (including additional covariates included in the primary outcome model following inspection of MAR data patterns), are found to be missing conditional on other covariates, and thus bias estimates generated using a complete case analysis. Given that we already have complete information for all the covariates included in our primary analysis model (CELF-P2 'Basic Concepts' baseline scores, region, and PVI status), we would only anticipate to perform MI in instances where additional covariates added to our primary outcome model are themselves MAR. Should this be the case, MI will be used in the same way as for suggested for the primary outcome above.

It should be noted, however, that in instances where missingness exceeds the 5% threshold but does not appear to be completely explained by the variables in the dataset, we cannot rule out that this data is 'missing not at random' (MNAR). This is where there is a systematic difference between observations that are 'missing' and observations that are not, but where this cannot be explained by observable characteristics captured in the data. Should it be suspected that our data is MNAR, we propose to carry out additional sensitivity analysis using the approach laid out by Carpenter et al. (2007). This would involve initial generation of parameter estimates through multiple imputation on the primary outcome under the false assumption that the data is MAR. Importance sampling will then be used to weight each imputation estimate based on the assumed deviance from MAR before averaging them, generating a revised estimate which is adjusted to the possibility of data being MNAR.

Any multiple imputation will be conducted using either *mi suite* in STATA 17 or higher, or using the *mice* package in R.

Analysis in the presence of non-compliance

As alluded to earlier, whilst the intention to treat (ITT) approach used in this trial will provide a more conservative estimate of intervention efficacy, capturing impact in a 'real-world' setting, it may also underestimate the benefit of Concept Cat specifically among those who receive the intervention. Given that Concept Cat is a continuous whole-class intervention taking place across an extended 30-week period, it is inevitable that incomplete receipt of the intervention will occur (e.g., through pupil

absence). Whilst the primary analysis outlined above assesses the impact of *offering* the intervention, we also propose to estimate the average treatment effect on the treated (ATT) through exploring its estimated impact in the presence of non-compliance, capturing the effect of the intervention specifically on those who received it. For this analysis, we will utilise the EEF's definition of compliance as 'the extent to which the critical ingredients of the intervention are delivered to and/or received by the target participants' (EEF, 2022).

After discussions with the delivery team, both Concept Cat session attendance and the number of words learned during these sessions were chosen to inform a measure of compliance. Information on words learned is collected by the delivery team at the setting level through recording the weekly focus word, providing a measure of week-on-week fidelity to the Concept Cat intervention.

Measuring child attendance is more complicated. Not all settings systematically collect attendance data, making it difficult to have accurate records of whether children were in settings on certain weeks. As a proxy for child attendance the evaluation team will use the attendance patterns from the start of the intervention in Autumn 2023. Attendance patterns record the number of hours, days, or sessions a child is expected to attend a setting. It does not give granular data on whether the child actually attended on those days (i.e., they may have been absent due to illness or a holiday), but in absence of actual attendance data this was considered a useful proxy. Data on attendance patterns at the start of the intervention (Autumn 2023) will be collected at endline. This will lead to a single metric of attendance - attendance patterns at the start of the intervention.

These two metrics were selected, after consultation with the delivery team, because they each capture specific aspects of compliance. The proposed attendance measure will inform us how often children were expected to be present to receive the intervention. However, relying solely on attendance might incorrectly classify settings where pupils were present, but the intervention was not delivered by staff, as 'compliant'. Likewise, exclusively focusing on 'words learned' might mistakenly categorize absent pupils as 'compliant'.

In order to generate a single measure of compliance, these two metrics need to be combined in a way that enables meaningful and straightforward interpretation. We therefore propose to generate a single binary measure of compliance, based on specific pre-set cut-offs. This binary variable will take on the value of 1 (indicating 'compliance') if both the following conditions are satisfied:

- The child meets the eligibility criteria of having attended at least 15 hours every week over the 30 week delivery period.
- The setting has taught 30 words over the 30 week delivery period.

Details of this compliance measure are presented in Table 4 below. While the ultimate compliance indicator for the CACE analysis will be a binary measure with specific thresholds for each aspect of compliance, we will also offer a descriptive overview of both pupil attendance and setting delivery. This additional information aims to provide insight into the extent to which these factors contribute to determining children's actual receipt of the intervention.

Table 4: Pupil level compliance measures

Compliance criterion	Data source	Compliance indicator
Setting-level compliance	Log of number of words taught by settings recorded by delivery team	Binary compliance indicator which takes on a value of 1 for each child who met the eligibility criteria for attendance, and who's setting taught 30 words over the full delivery period.
Child-level compliance	Data from settings indicating whether a child attended 15 hours or more per week during Autumn 2023	

Given the binary nature of this proposed compliance indicator, however, we will not be able to use it to understand how 'dosage' (i.e., the amount of the intervention received by pupils) impacts the primary outcome.

For each this aspect of the analysis, we will calculate the Complier Average Causal Effect (CACE) through a two-stage least squares (2SLS) instrumental variable (IV) approach. The first stage of this approach will involve regressing the binary compliance indicator on allocation to treatment, providing an estimate of how assignment of pupils and settings to receive Concept Cat encourages uptake of the intervention. This will effectively provide an overall 'compliance rate'. Results for the first stage will report the correlation between the instrument and the endogenous variable; and an F test. The second stage of this approach will model the primary outcome (as presented in Equation (1), but will include this predicted compliance in place of treatment assignment, and then will instrument this predicted compliance with treatment assignment. The Hedges' g derived from this model will similarly be calculated to provide an estimate of the CACE. The results of each of these two models will allow for us to discern the degree to which compliance with the Concept Cat intervention improves outcomes for pupils. This analysis will be conducted for the primary outcome only.

Depending on the granularity of attendance data, we will also provide histograms and frequency tables for these two separate elements of compliance to demonstrate the extent of compliance across both children and settings in the treated group.

Intra-cluster correlations (ICCs)

Given that a key analytical tenet of this efficacy trial is acknowledgement of the clustered nature of intervention receipt, with pupils nested within settings, the ICC is an essential metric for disentangling the degree of variance in the outcomes that is explained by between- and within-cluster variation.

The ICC of 0.15 used for the power calculations reported in the section on Sample Size and Power Calculations and at protocol stage (see the Concept Cat evaluation protocol for more information) is based on the ICC used in the power calculations for the trial proposal. This figure takes into account the generally higher ICC rates reported in EY settings compared to in school settings.

In the final report, we will report ICCs as they were at three stages: during the protocol stage; at randomisation; and during analysis. The ICC at analysis stage will be based on the primary outcome measure at both baseline and endline; and will be calculated using: (i) the same model as Equation (1) and; (ii) a model similar to that documented in Equation (1) but with no covariates, accounting for the clustering of pupils in schools (the so-called empty model).

Effect size calculation

We will use the effect sizes (hereafter ES) for cluster-randomised trials given in the EEF evaluator guidance (EEF, 2022), as adapted from (Hedges, 2007):[6]

[6] Where,
$(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the adjusted difference in means in the outcome between treatment and control group, accounting for the multi-level structure of the data.
$sd_1{}^2$ is the variance of the treatment group; and equally defined for the control group.
n_1 is the number of individuals in the treatment group, equally defined for the control group.

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\frac{(n_1 - 1)sd_1{}^2 + (n_2 - 1)sd_2{}^2}{n_1 + n_2 - 2}}}$$

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between the intervention and control group adjusted for baseline characteristics and $\sqrt{\frac{(n_1-1)sd_1{}^2+(n_2-1)sd_2{}^2}{n_1+n_2-2}}$ is an estimate of the pooled unconditional population standard deviation.

From the primary outcome model, we will take each group's adjusted mean and variance to calculate the effect size. This variance will be the total variance (across both pupil and school levels, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchinson & Styles, 2010). A 95% CI for the ES, which takes into account the clustering of pupils in schools, will also be reported. Effect sizes will be calculated for each of the models estimated.

References

Blachowicz, C., & Fisher, P. (2015). *Teaching Vocabulary in All Classrooms*. Pearson.

Carpenter, J., Kenward, M., & White, I. (2007). Sensitivity analysis after multiple imputation undermissing at random: a weighting approach. *Statistical Methods in Medical Research*, *16*, 259-275.

Clarke, D., Romano, J., & Wolf, M. (2019). The Romano-Wolf Multiple Hypothesis

Correction in Stata. *IZA Institute of Labor Economics*, Article 12845. https://docs.iza.org/dp12845.pdf

Dawson, A., Stokes, L., Huxley, C., Runge, J., Takala, H., Manzoni, C., Hudson-Sharp, C., & Williams, C. (2020). *Early Years Toolbox: Pilot Report*. https://educationendowmentfoundation.org.uk/public/files/Early_Years_Toolbox_Report_(final).pdf

Dimova, S., Illie, S., Rosa Brown, E., Broeks, M., Culora, A., & Sutherland, A. (2020). *The Nuffield Early Language Intervention. Evaluation Report*. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Nuffield_Early_Language_Intervention.pdf

Dong, N., & Maynard, R. (2013). PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies. *Journal of Research on Educational Effectiveness*, *6*, 24-67. https://doi.org/10.1080/19345747.2012.673143

Eadie, P., Cattram, N., Carlin, J., Bavin, E., Bretherton, L., & Reilly, S. (2014). Stability of language performance at 4 and 5 years: measurement and participant variability. *International Journal of Language & Communication Disorders*, *49*(2), 215-217.

EEF. (2022). Statistical analysis guidance for EEF evaluations.

Glennerster, R., & Takavarasha, K. (2013). *Running Randomized Evaluations: A Practical Guide*. Princetown University Press. https://doi.org/https://doi.org/10.2307/j.ctt4cgd52

Graham, J., Olchowski, A., & Gilreath, T. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, *8*, 206-213.

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

Hopkins, T., Harrison, E., Coyne-Umfreville, E., & Packer, M. (2022). A pilot study exploring the effectiveness of a whole-school intervention targeting receptive vocabulary in the early years: Findings from a mixed method study involving students as part of a practice-based research placement. *Child Language and Teaching Therapy*, *38*(2), 212-229.

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*, 92-106. https://doi.org/https://doi.org/10.1177/18369391211061188

Locke, A. (1985). *Living Language*. NFER-Nelson.

McBee, M. (2010). Modeling Outcomes With Floor or Ceiling Effects: An Introduction to the Tobit Model. *Gifted Child Quarterly*, *54*, 314-320. https://doi.org/10.1177/0016986210379095

Oades, F., Subosa, M., Speciani, E. R., Dysart, E., & Tracey, L. (2023). *Trial Protocol for Concept Cat: A two-armed cluster randomised controlled trial*.

Uttl, B. (2005). Measurement of Individual Differences: Lessons From Memory Assessment in Research and Clinical Practice. *Psychological Science*, *16*(6), 460-467. https://doi.org/https://doi.org/10.1111/j.0956-7976.2005.01557.x

Appendix

Histograms of CELF-P2 baseline scores

Figure 1: Histogram of CELF-P2 Basic Concepts subtest at baseline

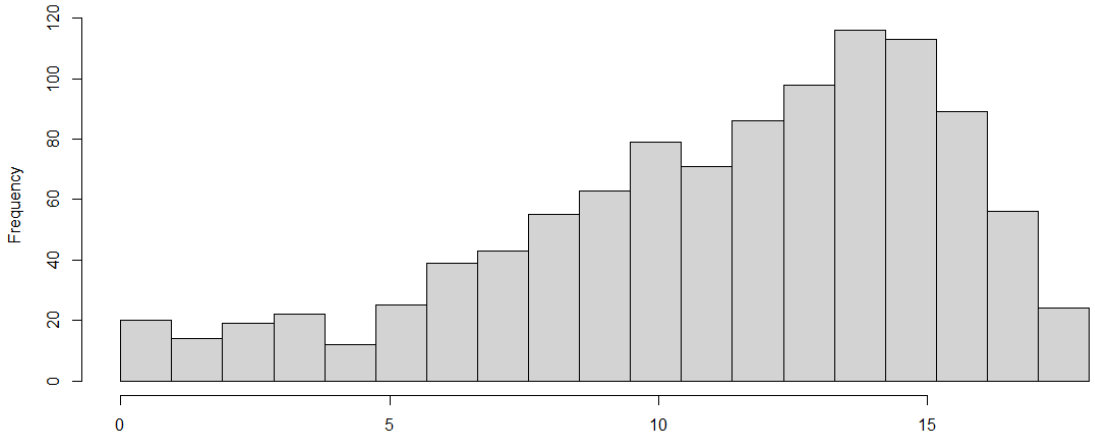
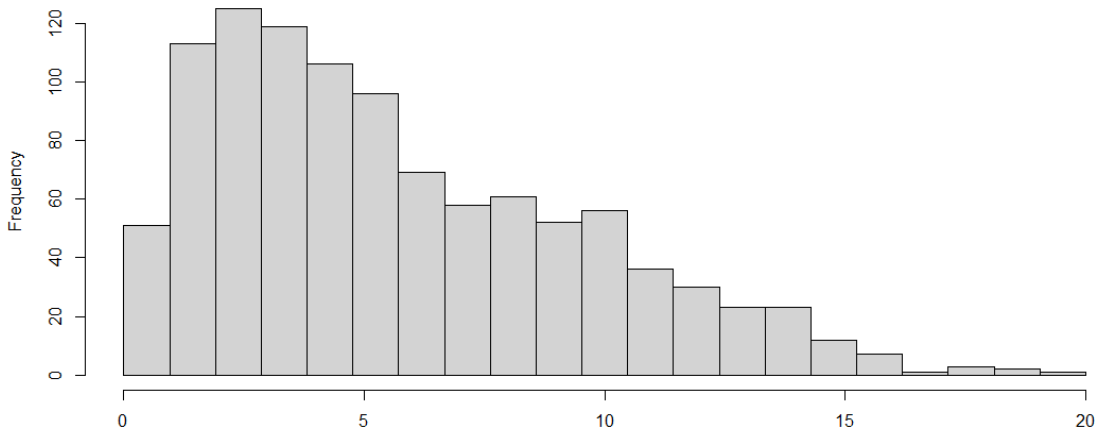


Figure 2: Histogram of CELF-P2 Following Directions subtest at baseline



Randomisation code

**# Concept Cat randomisation code

/*

The aim of this script is to produce treatment and control assignments to schools for the Concept Cat project.

First the script sets up the variables which will be used as Strata during the randomisation.

Next the script generates two random numbers for each observation, and adds them together, to generate a randomly and uniquely assigned identifier for each school - this is done by another researcher to ensure that randomsation is blind.

Then, after checking for duplicates in the unique identifier, the script saves the dataset, drops the information about the school and saves the dataset again - this creates two datasets (1) containing the school information and the unique identifier and (2) containing no information about the school but retaining variables required for randomisation and the unique identifier.

Randomisation is conducted on dataset (2) and the script checks for balance between conditions.

Finally, the script merges the school information back into dataset (2) from dataset (1) to complete the blind randomisation process and end with a list of schools and their assignment.

*/

**# Load the data

*we load in a list of schools with information required for randomisation

use "\\Nevis\RE\Projects\In_Progress\022807.015_P2899_EEF_Concept_Cat_RCT\WIP\04. Data collection\Data Not_Mirrored\Randomisation\list_of_schools_v0.1.dta"

**# Preparing strata variables

*check how many schools have PVI status and how many are in each region

*we should be looking for balance in these since they are the strata in the randomisation

tab settingtypepviormaintained

tab region

*settingtypepviormaintained and region are string variables and need to be numeric for the randomsation code to work properly

encode settingtypepviormaintained, gen(pviormaintained)

encode region, gen(regionnum)

**# Generating random unique identifier

*now we generate a random number for each observation - since stata generates random numbers with replacement, we generate two random numbers and add them together to significantly reduce the change that the same number is assigned to two observations

*set the seed

set seed 991

gen double rand1 = runiform()

gen double rand2 = runiform()

gen uniqueobs = rand1+rand2

*we check that there are no duplicates (the statistical probability of getting duplicates is below 0.00001 since we generated two random numbers independently for each observation)

duplicates list uniqueobs

**# Save the dataset

*now we save the data with the uniqueobs variable including

save "\\Nevis\RE\Projects\In_Progress\022807.015_P2899_EEF_Concept_Cat_RCT\WIP\04. Data collection\Data Not_Mirrored\Randomisation\list_of_schools_with_unique_identifier_v0.1.dta", replace

**# Prepare the dataset for randomisation

*now we drop the school name and address, so that we can randomise based on the unique identifier

drop schoolname schooladdress mainschoolcontactname mainschoolcontactemail hassentpupildemographics hasstartedtestingwithelklan hascompletedtestingwithelklan testingdatebookedwithelklan numberofpupilsegress numberofpupilstestedegress numberofpupilstestedelklan dsaforwardedtokerin dsacompletedsignedbysettingandke notes hassentdsa

*and we save the data again - this is the data we will do the randomisation on

```
save "\\Nevis\RE\Projects\In_Progress\022807.015_P2899_EEF_Concept_Cat_RCT\WIP\04. Data collection\Data Not_Mirrored\Randomisation\list_of_schools_with_unique_identifier_for_randomisation_v0.1.dta", replace


**# Run the randomisation


*we use the randtreat package to run the randomisation

ssc install randtreat, replace

*we use randtreat to generate a new variable "treatment" which is stratified by eia status (equal (non-)eia in treatment and control), creating two different groups (multiple(2)) and using a seed of 999 to ensure that this code generates the same assignment each time

*since we have an unequal number of schools in some regions, we need to assign a misfit strategy to deal with the schools left out of the equal group assignment - here we assign the strata method which effectively randomly assigns it to one group in this context

randtreat, generate(treatment) setseed(990) strata(pviormaintained regionnum) multiple(2) misfits(strata)


**# Check the balance of the randomisation


*now we check that we have balance in the two strata by tabulating their counts in both conditions

tab treatment pviormaintained

tab treatment regionnum


**# Merge school information back into treatment allocation


*we remerge to bring the school information back into the dataset

merge 1:1 uniqueobs using "\\Nevis\RE\Projects\In_Progress\022807.015_P2899_EEF_Concept_Cat_RCT\WIP\04. Data collection\Data Not_Mirrored\Randomisation\list_of_schools_with_identifier_v0.1.dta"


**# Save the final dataset


*finally we save the resulting dataset which contains all of the required information

save "\\Nevis\RE\Projects\In_Progress\022807.015_P2899_EEF_Concept_Cat_RCT\WIP\04. Data collection\Data Not_Mirrored\Randomisation\treatment_allocation_v0.1.dta"
```