Counting Collections
Statistical Analysis Plan

Evaluator (institution): Sheffield Hallam University
Principal investigator(s): Dr Martin Culliney

Education Endowment Foundation

Template last updated: August 2019

PROJECT TITLE[1]	Evaluation of Counting Collections, a two-armed cluster randomised trial
DEVELOPER (INSTITUTION)	University of Nottingham
EVALUATOR (INSTITUTION)	Sheffield Hallam University
PRINCIPAL INVESTIGATOR(S)	Dr Martin Culliney, Dr Karen Daniels
PROTOCOL AUTHOR(S)	Dr Martin Culliney, Dr Karen Daniels, Joanne Robson
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Age 4-5, Early Years (Reception)
NUMBER OF SCHOOLS	180
NUMBER OF PUPILS	3600
PRIMARY OUTCOME MEASURE AND SOURCE	Number attainment, GL Sandwell Early Numeracy Test
SECONDARY OUTCOME MEASURE AND SOURCE	N/A

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]	24/1/24	*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

Introduction

This Statistical Analysis Plan describes the planned impact evaluation analysis for an efficacy trial of Counting Collections, developed by the University of Nottingham. The intervention is a hands-on approach to developing early years pupil number sense (understanding of number and quantity) which supports subitising, comparing numbers and composition of numbers. It involves children using containers of objects (manipulatives) to find out how many are in the collection.

The weekly Counting Collections sessions are based on a four-part routine, with pupils working in pairs. A key factor in the success of the intervention is the role of the adult as the facilitator and their knowledge of teaching and learning in number, including developmental progressions (learning trajectories). A small-scale pilot study (Gripton and Pawluch, 2021) indicated that the programme supported the development of teacher subject knowledge (rooted in developmental progressions in number), which led to perceived increases in pupil attainment and interest in number. This requires further investigation, especially around the varying degrees of teacher knowledge and the training needed.

One teacher per school takes part in the professional development programme led by the developer. This includes an online environment with ongoing support throughout the programme. Teachers can revisit content, communicate with the developers, and discuss Counting Collections with other teachers. Trained teachers deliver the intervention to pupils in class. If Teaching Assistants (TAs) are allocated to support these sessions, it is expected that teachers train them so that they understand the intervention. However, TA participation in Counting Collections is optional for schools.

During the evaluation period, control schools continue teaching number to reception pupils using a business-as-usual approach. They receive an incentive payment of £250 for submitting pupil data to the evaluation team and completing the baseline assessments, and a further £250 payment on completion of the outcome assessments. The full list of exclusion criteria for settings and pupils in the trial are outlined in the trial protocol.

Design overview

Table 1: Trial design

Trial design, including number of arms		Two-arm, cluster randomised
Unit of randomisation		School
Stratification variables (if applicable)		Geographical area Existing use of Mastering Number intervention
Primary outcome	variable	Number attainment
	measure (instrument, scale, source)	Sandwell Early Numeracy Test (B), raw score (0-76), GL Assessment
Secondary outcome(s)	variable(s)	N/A
	measure(s) (instrument, scale, source)	N/A
Baseline for primary outcome	**variable**	Number attainment
	measure (instrument, scale, source)	Sandwell Early Numeracy Test (A), raw score (0-76), GL Assessment
Baseline for secondary outcome	**variable**	N/A

This evaluation aims to provide evidence of the impact of Counting Collections. Schools in the South West, North East and East Midlands regions of England were recruited by the developer. Recruitment was extended into South Yorkshire to broaden the pool of schools that could take part. Any school in these areas with at least 20 pupils in the 2023/24 reception cohort was eligible for the trial. The aim was to recruit at least 50% of schools from Education Investment Areas. By July 2023 a total of 180 schools were recruited, 125 in EIAs (69%).

Schools supplied details on all pupils in one of their reception classes to the evaluation team. As assessing more than 20 pupils per school would increase costs and burden on schools while bringing only minor improvements in statistical sensitivity (see sample size calculations below), it was agreed to limit baseline assessments to 20 pupils per school. The headline analysis sample will be that group of pupils, with a maximum of 20 per school. This sample of pupils was randomly selected prior to baseline testing in each school. There is no standby list for replacing absent pupils; such cases will be treated as attrition.

The research questions are:

1. What is the impact of Counting Collections on reception pupil attainment in number as measured by the Sandwell Early Numeracy Test?
2. What is the impact of Counting Collections on disadvantaged reception pupil attainment in number as measured by the Sandwell Early Numeracy Test?

Primary outcome

The Sandwell Early Numeracy Test (SENT) will be used as the baseline and primary outcome measure. It has two components (A and B) and is suitable for measuring the impact of classroom interventions on a pre- and post-test basis. SENT component A will be used as the baseline measure for this trial, with SENT component B as the outcome measure, so all pupils will do both assessments. The publisher has confirmed that it is appropriate to use the assessment in this way, and it has been used similarly in previous research (Torgerson et al., 2011:49). Both SENT components are scored on a 0-76 scale and marks are recorded by the assessment administrator, who conducts the assessment one-to-one with each pupil. The raw scores will be used in analysis.

The assessment explores five strands of basic numeracy skills: identification, oral counting, value, object counting and language, but these are not validated for use as standalone measures. As Counting Collections aims to improve number attainment overall, all aspects of the SENT assessment are relevant to the intervention. Baseline assessments were completed at 172 schools between 25 September and 24 November 2023. The completed booklets from two schools were lost in the post, leaving 170 schools.

Secondary outcomes

As mentioned above, the SENT assessment targets five strands of basic numeracy skills that are all relevant to the intervention but not validated for use as standalone measures. It was decided not to use any of them as secondary outcomes. As no other suitable assessments were identified, it was decided to use a primary outcome only.

Randomisation

Randomisation was at school level and took place on 14 July 2023. The 180 schools (89 intervention, 91 control) were informed of their allocation immediately. It was decided to randomise before collecting baseline data to give schools sufficient time to plan for teacher release for full day of training in October. Randomisation was conducted at school level to minimise spillover risk. Geographical area (South West, East Midlands and South Yorkshire, North East) was used as a stratification variable, along with whether the school is already using the Mastering Number intervention. This is to reduce the risk of allocation imbalance in certain geographical areas undermining the viability of the training, and to mitigate against the use of other relevant interventions confounding the results of this trial. The allocation was completed using the ‘stratarand’ command in Stata 17.

As randomisation was undertaken before baseline data collection, there has been some attrition. This has mainly been in the control group, which has decreased from 91 schools at randomisation to 83 schools at baseline. The intervention group had 89 schools at randomisation with 87 remaining at baseline. In total, 170 schools are still in the trial at the time of writing. The distribution of the stratifying variables between the intervention and control groups is shown in Table 3 (below).

Sample size calculations overview

The design is a 2-level clustered RCT. In calculating the Minimum Detectable Effect Size (MDES), the smallest effect size that could be detected as statistically significant (often set as p<0.05) with a statistical power of 80% or higher, our estimates at the protocol stage were based on the following assumptions:

$M_{j\text{-}k\text{-}2}$ - T-distribution multiplier assuming a two-tailed test with a statistical significance of 0.05, statistical power of =0.80 and J-K-2 (175) degrees of freedom
R_i – Participant (pupil) level pre/post-test correlation of 0.6 (R_i^2 = 0.36)
R_c – Cluster (school) level pre/post-test correlation of 0.2 (R_c^2= 0.04)
ρ – Intracluster correlation (ICC) 0.17
j – number of schools (180)
m – number of pupils per school (20)
k – number of cluster level covariates (3)
P - Proportion of schools allocated to intervention group (P=0.5)

The ICC and participant correlation values are taken from the evaluation of Maths Champions (Robinson-Smith et al., 2018), the only early years maths trial published by EEF at the time of writing. The ICC reported at the analysis stage of that trial was 0.17, and the pupil pre/post-test correlation was 0.59. While these figures are from an evaluation using a different outcome measure with slightly younger children, this is still recent evidence from a trial of a programme of the same length and on the same subject. We have adopted the same ICC value (0.17) and a similar pupil pre/post-test correlation (0.60). The ICC is lower than the default ICC recommended for attainment outcomes by the IES What Works Clearinghouse (2022:171). Increasing the ICC to 0.20 would result in a higher MDES.

Cluster level correlations were not supplied in Robinson-Smith et al. (2018) but are conservatively estimated here as 0.20. Calculations were performed in Excel using the formula set out in Bloom et al (2007), which relates to two-level clustered randomised controlled trials. This allows covariates to be included at both individual (pupil) and cluster (school) level, which in turn increases sensitivity.

Equation 1: Minimum Detectable Effect Size in a two-level clustered RCT

$$\mathrm{MDES} = M_{j-k-2}\sqrt{\left(\frac{\rho(1-R_c^2)}{\mathrm{P}(1-\mathrm{P})\mathrm{J}}\right)+\left(\frac{(1-\rho)(1-R_i^2)}{\mathrm{P}(1-\mathrm{P})\mathrm{Jm}}\right)}$$

Based on the assumptions above, an MDES of 0.20 standard deviations would require 150 schools with 20 pupils each. However, to mitigate against potential attrition, it was agreed that more schools should be sought. The protocol states a target of 176 schools, and ultimately 180 schools were recruited. This is the sample size at randomisation and gives an overall MDES of 0.18. To illustrate the robustness of this sample, attrition of up to 20% at school and/or pupil levels results in indicative MDES estimates[2] of 0.20 or lower. These figures are presented to give an idea of the sensitivity that would be expected with this

[2] Attrition is likely to undermine randomisation and hence the validity of estimating an 'MDES'. Instead, these are indicative MDES estimates that indicate the sensitivity of a trial of a smaller size as per the different attrition scenarios described in the text

relatively high level of attrition. It is also worth noting that the overall MDES for the sample size at baseline (170 schools) is 0.19, rising to 0.21 for the FSM sample.

For the subgroup of pupils eligible for free school meals (FSM), (estimated at eight per school, as the number is likely to be higher than the national average in Education Investment Areas and the aim is to recruit at least 50% of schools from these districts), the MDES is 0.20 with 180 schools. With 150 schools, the FSM MDES would be 0.22.

Table 2: Sample size calculations

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.18	0.20	0.18	0.20
Pre-test/ post-test correlations	level 1 (pupil)	0.60	0.60	0.60	0.60
	level 2 (school)	0.20	0.20	0.20	0.20
Intracluster correlations (ICCs)	level 2 (school)	0.17	0.17	0.17	0.17
Alpha		0.05	0.05	0.05	0.05
Power		0.80	0.80	0.80	0.80
One-sided or two-sided?		2	2	2	2
Average cluster size		20	8	20	8
Number of schools	intervention	90	90	90	90
	control	90	90	90	90
	total	180	180	180	180
Number of pupils	intervention	1800	720	1800	720
	control	1800	1800	1800	720
	total	3600	1440	3600	1440

Analysis

Multilevel linear regression models will be constructed for the SENT (B) primary outcome, with pupils clustered within schools. For each model, the coefficient of the treatment allocation variable, which distinguishes intervention group pupils from control group pupils, will be converted into Hedges' g effect size statistics with 95% confidence intervals.

The first model will only include the treatment allocation identifier (an outcome only model). The second model will also include the baseline test score as a covariate at the pupil and school levels[3]. SENT (A) will be used as the baseline covariate for analysis of the primary outcome. The final model will also include the stratifiers used in the randomisation process

[3] These will be centred so that the school level will be centred on the mean for all schools and the pupil level will be centred around the school mean.

(geographical area, whether the school is using the Mastering Number intervention) and will form the headline intention to treat (ITT) impact analysis for the SENT (B) primary outcome.

Follow-on ITT analyses will focus on the impact of Counting Collections on number attainment among disadvantaged pupils, as defined by the NPD variable EVERFSM_6 (which in this instance simply refers to FSM eligibility in the study year given the pupils are in the reception phase). The same three model stages used for the headline ITT analyses will be used.

Subgroup analyses

Subgroup analysis will be conducted on pupils eligible for free school meals as identified by the 'EVERFSM_6' indicator obtained from the NPD, as is required for all EEF trials. This data will be collected from NPD for consistency with other EEF trials. It will be available in Autumn 2024, when the outcome assessments have been completed. The second research question for the trial is based on this subgroup analysis, although it is exploratory as the study is not powered for meaningful subgroup analysis. The analysis will be run on a sample including only FSM pupils and will also be run on the whole study sample, with an interaction effect between FSM status and treatment allocation included in the model.

Additional analyses

None planned.

Longitudinal follow-up analyses[4]

None planned.

Imbalance at baseline

To monitor imbalance between treatment and control groups at baseline following baseline assessment, descriptive analysis was undertaken at the setting and child level. Table 3 shows the breakdown of the baseline sample according to a series of key variables. As discussed above, two stratifiers were used in the randomisation, geographical region and whether the school uses the Mastering Number intervention, and both variables were well balanced at randomisation as a result.

Attrition between randomisation and baseline has disrupted the balance created by the stratified randomisation, yet the sample remains reasonably well balanced on the stratifiers. At baseline, the intervention and control groups both have 20 schools in the North East region, and in the other regions the number of intervention schools is slightly higher. The percentage of schools in the intervention group using the Mastering Number programme (47%) is almost the same as the control group (46%).

The Ofsted rating of the schools in the baseline sample is also well balanced. A very similar percentage of schools are classed as outstanding (Intervention 9%, Control 11%), good (Intervention 74%, Control 71%), and requiring improvement (Intervention 5%, Control 6%) across both treatment conditions. It is worth noting that 24 schools do not have an Ofsted rating although these are evenly split across intervention and control schools.

In terms of pupil numbers, the intervention and control groups appear to be of nearly equal size, with an average of 340 and 349 pupils respectively. Another area where the baseline

[4] Please see the longitudinal analysis guidance.

sample is balanced is the percentage of disadvantaged pupils in school, with both the intervention and control groups having 30% of pupils eligible for free school meals.

Finally, the scores from the pupil baseline assessment conducted in Autumn 2023 shows that the control group achieved slightly higher scores. The mean score for control pupils was 17.69 compared to 16.97 for intervention pupils. Dividing the difference by the pooled standard deviation produces an effect size of -0.09. It is worth noting that the ICC value in the baseline data is 0.12, which is lower than the ICC used in the power calculations and if replicated at the analysis stage would lead to a more sensitive trial design than anticipated.

Table 3: Imbalance at baseline

	Baseline (N Schools=170)		Analysis (N Schools=)	
	Intervention (N=87)	Control (N=83)	Intervention (N=)	Control (N=)
School level (categorical)	% (n)	% (n)	% (n)	% (n)
East Mids./South Yorks.	54%(47)	53%(44)		
North East	23%(20)	24%(20)		
South West	23%(20)	23%(19)		
Uses Mastering Number	47%(41)	46%(38)		
Does not use Mastering Number	53%(46)	54%(45)		
OFSTED ratings				
Outstanding	11%(10)	16%(13)		
Good	83%(72)	76%(63)		
Requires improvement	5%(4)	6%(5)		
Inadequate	0%(0)	0%(0)		
Missing	1%(1)	2%(2)		
School level (continuous)	Mean (SD)	Mean (SD)		
Total number of pupils (including part-time pupils)	340 (189)	349 (253)		
Percentage of disadvantaged pupils	29.57 (16.34)	30.30 (18.68)		
Pupil level (continuous)	Mean (SD)	Mean (SD)	Effect size	
Pre-test scores				
SENT-R	16.97(8.21)	17.69(8.67)	-0.09	

Missing data

The impact analyses will examine missing data in the outcome and explanatory variables and consider whether it is reasonable to assume that the missing data are random. A multilevel logistic regression model with a binary outcome denoting when outcome data is missing (=1) or not (=0) and the same covariates as the headline ITT model will be estimated to identify any patterns. This model will then be replicated with only pupils at schools that took part in the outcome testing, to focus on pupil level attrition in those schools.

In the instance of any missing outcome data, the (complete) baseline and ITT samples will be compared across all ITT variables and additional variables shown in Table 3 above. If over 5% of outcome data is missing, as part of the follow-on analyses a multilevel logistic regression model estimating when outcome data is missing (=1) or not (=0) will be constructed. The ITT

variables and additional school level variables will be used to identify whether the missing outcome data can be assumed to be missing at random.

If none of the explanatory variables are found to account for a statistically significant amount of variation in the missing data outcome, we will cautiously assume that the data is missing at random, otherwise multiple imputation will be used and the results compared with the headline ITT analysis for the primary outcome.

If one or more explanatory variables are found to account for a statistically significant amount of variation in the missing data outcome, we would undertake a sensitivity analysis to repeat the ITT analysis with these variables included. The potential bias introduced by missing outcome data on the ITT estimate will be illustrated by comparing the estimated ITT effect size with the effect size estimated from the ITT model including the additional variables.

Compliance

Compliance will be measured at the school level. Full details of the three indicators are provided below. Each relates to activities undertaken by the participating teacher, yet as only one teacher per school is taking part in the trial, these are effectively school level indicators. The measures will be combined to create overall full and part compliance indicators at the school level, and then used to estimate the Complier Average Causal Effect (CACE). The purpose of the CACE analysis is to estimate the impact of Counting Collections for pupils deemed to have 'complied' with the intervention, though as no pupil level compliance indicators are being used, pupil compliance is simply being at a school classed as compliant.

Table 4: Compliance indicators

Activity	Full compliance	Part compliance
Materials	Graded manipulatives and supporting tools must be present in the classroom	
Delivery	Delivering at least 5 weekly Counting Collections sessions per half term	
Training	Attending all five sessions, including the initial in-person session. Attending online sessions 2-5 synchronously (live) or one of these sessions asynchronously (catch-up) if all other criteria are met	Attending all five sessions, including the initial in-person session. Completing online sessions 2-5 asynchronously (catch-up)

CACE will be estimated using two-stage least squares (2SLS) regression (Gerber and Green, 2012). The first stage will model the compliance variables using the same explanatory variables used for the headline ITT analyses along with additional school level items that are available via the school census as included in Table 3 (above). This will be a multilevel logistic regression model used to generate predicted compliance (1 or 0) for use in the second stage model. The second stage models will use predicted compliance in place of the group identifier variable in the ITT analyses specified above to generate the CACE estimates. This process will be undertaken twice, for full and part compliance.

The developer has specified three criteria that schools must meet to achieve full compliance:

- Teaching at least 5 weekly Counting Collections sessions per half term

- Teacher attendance at all five professional development sessions (first session in person, then synchronous online sessions)
- Graded manipulatives and supporting tools must be present in the classroom

If one of the online professional development sessions is attended asynchronously, full compliance is still achieved provided that the other criteria are met. Part compliance can be achieved through teachers completing the online PD sessions asynchronously. The unit of analysis for compliance is the school, although as mentioned above only one teacher per school is taking part in the trial.

Intra-cluster correlations (ICCs)

For the primary outcome at both pre and post-test, ICCs at the school level will be estimated using a null (empty) 2-level multilevel variance components model. Within the analyses, a table will present the variance decomposition for the two levels (school and pupil) along with the ICC estimates.

$$\mathrm{ICC} = \frac{\mathrm{Variance}_{\mathrm{school}}}{\mathrm{Variance}_{\mathrm{school}} + \mathrm{Variance}_{\mathrm{pupil}}}$$

Effect size calculation

The effect size measure to be used will be Hedges' g. This will be calculated using the following equation.

$$ES = \frac{(T-C)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{pup}}}$$

Where:

δ^2_{sch} is the school level variance and δ^2_{pup} is the pupil level variance for the language outcome from the empty/null multilevel model.

$(\mathrm{T}-\mathrm{C})_{\mathrm{adjusted}}$ is the mean difference between the attainment of pupils in treatment schools and pupils in control schools. This is obtained from the coefficient for the school level 'group' variable from the final (headline) analyses.

The coefficient standard error and the upper/lower 95% confidence intervals will also be converted into units of standard deviations using the above formula.

Timeline

Table 5: Timeline

Dates	Activity	Staff responsible/ leading
Oct 22	Set-up meetings and IDEA workshop	All

Dates	Activity	Staff responsible/ leading
Nov-Dec 22	Ethical approval Draft MoU, consent and information forms Design IPE instruments Evidence review	SHU
Feb 23	Protocol Trial registration	SHU
Feb-Jul 23	Recruitment Data collection from schools	UoN
Jul 22	Randomisation	SHU
Sep-Nov 23	Collect pupil data Baseline testing	SHU
Oct 23	Day 1 teacher training/observations	SHU/UoN/schools
Nov 23- May 24	Intervention delivery	Schools
Mar-Apr 24	Conduct IPE school visits	SHU/schools
Nov 23- May 24	Day 2-5 teacher training/observations	SHU/UoN/schools
Jan 23	Statistical Analysis Plan first draft	SHU
Mar 23	NPD application	SHU
Jun-Jul 24	Outcome testing Post-intervention teacher survey	SHU/schools
Sep 24	Data analysis	SHU
Jan 25	Report first draft	SHU
Apr 25	EEF to receive the final report	SHU

References

Bloom, H, S., Richburg-Hayes, L. and Black, A. R. (2007) 'Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions' *Educational Evaluation and Policy Analysis* 29(1) 30–59

Gerber, A., and Green, D. (2012). Field Experiments: Design, analysis and Interpretation. W.W. Norton & Company

Gripton, C. and Pawluch, D. (2021) 'Counting Collections in the Early Years'. *Mathematics Teaching* 275, 6-10

Hedges, L.V. and Hedberg, E.C. (2013) Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster randomised experiments in education. *Evaluation Review* 37(6) pp445-489

Ofsted (2021) *Research review series: mathematics.* Available online at: https://www.gov.uk/government/publications/research-review-series-mathematics/research-review-series-mathematics

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.* Education Endowment Foundation

Torgerson, C., Wiggins, A., Torgerson, D.J., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hodgen, J., Hulme, C., Tymms, P. (2011) Every Child Counts: The independent evaluation, Technical Report

What Works Clearinghouse (2022) *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0*, Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). Available online: https://ies.ed.gov/ncee/wwc/Handbooks