Statistical Analysis Plan for Maths Counts

Durham University

INTERVENTION	Maths Counts
DEVELOPER	Mead School
EVALUATOR	Durham University
TRIAL REGISTRATION NUMBER	The trial was not registered*
TRIAL STATISTICIAN	Stephen Gorard
TRIAL CHIEF INVESTIGATOR	Stephen Gorard
SAP AUTHOR	Beng Huat See
SAP VERSION	1.0
SAP VERSION DATE	16/01/2018

** The evaluators consider that post hoc registration of the trial is not necessary since the protocol and this SAP are published. The report will be published in its entirety on the EEF website and the findings will be in the public domain. The reasons for registering a trial are to inform the field that a trial has been conducted, and to ensure that all results (both positive and negative) are published and that the trial protocol stating the main outcome measures is written before the trial begins to avoid dredging of results or changing the main outcomes. Since this trial already conforms to all these requirements, there is no need to register the trial.*

Introduction

The project to be evaluated is an intervention called 'Maths Counts developed by The Mead Community Primary School, a part of The Mead Academy Trust and Teaching School based in Trowbridge, England. Maths Counts draws on some of the key principles of the Every Child Counts 'Numbers Count' programme, developed by Edge Hill University, working in partnership with Lancashire County Council

The intervention is delivered by teaching assistants, also referred to as 'Learning Partners' (LPs) These LPs are supported by their Maths Leads, MLs (the school's maths co-ordinator or specialist teacher). After detailed diagnostic assessments conducted by MLs, the Maths Counts sessions are delivered three times a week over 10 weeks on a one-to one basis by trained Learning Partners; each lesson lasting 30 minutes. Maths Counts is facilitated by the use of a digital platform called the 'Digital Maths Tool', . The Digital Tool is specifically designed to record progress, identify strengths and areas for improvement and create bespoke lessons for individual learners struggling with basic number skills. The digital platform is populated with evidence-based resources, games and activities for LPs and teachers to support children in areas where they require help. It also includes home learning, so that parents can practice with the children at home.

Study design

The study is a one-year efficacy trial involving primary schools in England. The planned sample size is 30 schools, but 35 schools have been recruited. This is to increase the sample size as project developers have recommended that, in order to ensure fidelity to the programme, each Learning Partner (LP) should work with a maximum of two children and

each Maths Lead should undertake detailed diagnostics with a maximum of four children and support two LPs. Since some schools recruited are large primaries and have more than one ML, they are therefore able to support more learners. Schools recruited are based in four hubs in the South and South West of England. The four hubs are: London, Somerset, Bristol and Wiltshire. The targeted schools are those with above the national average levels of pupils eligible for free school meals.

Randomisation is at the pupil level, meaning that all schools are intervention schools, so reducing post-allocation demoralisation, and thus dropout. The Mead developers have also assured schools that they can continue the intervention with their control children after the post-test since the LPs and maths leads have been trained and they now have the teaching resources. This will minimise demoralisation of control children and perhaps reduce the John Henry effect.

Eligible pupils are those in Year 3 to Year 6. As each LP supports two pupils, an average of four pupils per school will be considered. Eligibility was assessed prior to randomisation using a combination of teacher judgements of which pupils are deemed to be unlikely to meet the Year 2 Programme of Study and a list of criteria spelt out in the Ofsted framework grade indicators for pupil outcomes. Priority was given to:

- Pupils at risk of not achieving the nationally expected levels
- Lowest attaining pupils
- Younger Key Stage 2 pupils will also be given priority as they are deemed to have most to gain from earlier intervention
- Pupil Premium pupils

Eligible pupils, once identified, were individually randomised within the school either to receive the Maths Counts intervention or to teaching business as usual.

The way the intervention is used ensures that there is little potential for contamination. First the programme is designed with the digital tool being password protected, so only treatment pupils' progress and the appropriate activities as ascertained by the tool can be accessed by LPs. The programme also begins with a diagnosis of needs and the appropriate level and activities to be used with individual child. Since control pupils were not diagnosed, their learning needs are not determined. There are therefore no identified activities for LPs to use with them. The process evaluation also assesses the possibility of contamination either by friendship groups or inadvertently by LPs sharing the bespoke Maths Counts' teaching activities.

Protocol changes

Feedback from the pilot suggests that the GL Progress Test in Maths (originally proposed in the protocol) was too difficult for the kind of pupils that the programme was meant to support. For this reason the InCAS Assessment has been chosen by the developers (and supported by EEF) for the main trial as it is thought to be more in line with what the programme wants to measure. As InCAS includes a test of maths attitude, there will be no bespoke pupil attitude survey as originally proposed in the protocol.

Furthermore, during the trial it became apparent that baseline KS1 results were not homogenous for all pupils. Specifically, pupils in Year 4, Year 5 and Year 6 had point scores available, while pupils in Year 3 had ordered categorical outcomes. This is because the approach to describing achievement of pupils in England changed from levels to the use of 4 descriptive categories, which has changed the nature of the data available for Year 3 pupils.

Randomisation

Pupils identified as eligible are randomised to one of 2 groups: Maths Counts or business as usual. This was done by the evaluator using a random number generator (random.org programme) in the presence of two colleagues in the School of Education.

A total of 305 pupils have been identified and 152 are randomised to Maths Counts and 153 to control. This is because four schools were able to provide four LPs, and so were able to support 8 pupils.

Calculation of sample size

The sample size calculation is based on the assumption that there would be 30 schools and four year groups (Years 3, 4, 5 and 6). Ideally the trial would include an average of 3 eligible pupils per class. Assuming 1.5 classes per year group, and 3 eligible pupils per class, there would be 18 pupils per school, giving a total sample of 540 or 270 per arm.

Traditional power calculations are based on the approach of significance testing (Gorard et al. 2017). They are not included here. Instead, we calculate the sample size needed for any 'effect' size to be considered secure by considering *a priori* the number of 'counterfactual' cases needed to disturb a finding (Gorard and Gorard 2016). This number needed to disturb (NNTD) is calculated as the 'effect' size multiplied by the number of cases in the smallest group in the comparison (i.e. the number of cases included in either the control or treatment group, whichever is smaller). This approach allows for estimating ES and sample size using the formula as shown.

$$NNTD = ES*n$$

Therefore, $n = NNTD/ES$ and

$$ES = NNTD/n$$

This is a useful measure of the scale of the findings to chance (and their variability as represented by the standard deviation used to compute the 'effect' size), taking into account the scale of the study. It can then be extended to compare this sensitivity directly to other more substantial sources of error such as the number of missing values/cases. The number of cases actually missing a value can be subtracted from the NNTD to give an estimate of how large the 'effect' size would be even in the extreme situation that all missing cases had the "counterfactual" score hypothesised in the NNTD calculation. Here the 'counterfactual' score is one standard deviation away from the mean of the group with the largest number of cases. The standard deviation would be added if the mean of the smaller group (in scale) were smaller than the mean of the larger group, and subtracted if the mean of the smaller group was the largest. (Gorard et al. 2017).

Based on Gorard et al. 2016, NNTD of 50 can be considered a strong and secure finding. Using this as a working assumption, the number of cases needed in each group (assuming equal size) to detect an 'effect' size of 0.2 (which is typical for an education intervention) will be 250 (or 50/0.2). This is assuming no attrition. In this trial, 35 schools and a total of 305 pupils were recruited with an average of 8.7 eligible pupils in each school. The trial is therefore underpowered at the outset. The EEF was aware of this and the developers were encouraged to recruit more and bigger schools. The developers had worked extremely hard to recruit schools from a wide range of areas using their own professional liaisons. Although a lot more schools have expressed interest, the requirement was to include schools with a high proportion of free school meal pupils and also to focus on the more committed schools to reduce the possibility of dropouts.

With a sample of 270 per arm, we would expect to detect an effect as small as 0.19 (rounded to two decimal places). In this trial, the number of cases per arm is 152 (treatment) and 153 (control). Assuming NNTD of 50, we would expect to confidently detect an effect of 0.33 (rounded to two decimal places).

The NNTD calculation concerns the security of a difference, and so is relevant to internal validity only. Issues such as clustering, concerned with whether the result may also occur among cases not in the RCT, are therefore irrelevant. In addition, as pupils are individually randomized within schools and analysis would be of all pupils in the two groups and not by schools, clustering effects, if there are any, should be evenly spread between the two groups across all schools.

Follow-up

No schools have dropped out.

To minimize attrition, the developers are offering the destination schools of school leavers an incentive payment of £200 to administer the InCAS assessment.

Outcome measures

Primary outcome

The primary outcomes will be the general age-standardised maths scores on the digital CEM InCAS Assessment for All pupils combined. The only scores available are the age-standardised ones as the test is adaptive (according to the supplier). This will be the headline finding.

Pupils' prior KS1 point scores in maths from NPD will be used as the pre-test score and to establish baseline equivalence between the two groups.

We originally planned to have a pre-test and the EEF advises the use of KS1 results as pre-test scores partly to reduce cost but also to minimise the burden of testing[1]. As there are clear differences between the baseline values of Year 3 pupils compared to the other year groups in the trial we will analyse the results for the Y3 pupils separately from the other year groups as well as combined. The headline finding will be the combined results.

Secondary outcomes

The secondary outcomes will be mental maths and attitude towards maths measured using the subscales on the digital CEM InCAS test. Attitude to maths will be collected via the maths only questions in the attitudes subscale of the CEM InCAS test. This use of the maths-only attitudes was approved by CEM (the test developer).

Other data

Pupils' EverFSM status will be obtained from NPD and used for sub group analyses. Other background characteristics such as age, date of birth, sex, ethnicity, first language, SEN are also collected, where possible, from schools to establish equivalence between groups.

Analysis

Analysis is conducted independently of process evaluation results.

Primary intention-to-treat (ITT) analysis

Comparisons will be made between Maths Counts and the business-as-usual-group. Initial analysis will be based on the 'effect' size difference between groups on post-test scores only, and presented with pre-intervention KS1 scores. In addition, the results will be presented as 'effect' sizes based on gain scores calculated using the difference in the mean gain scores made between KS1 maths point scores and the InCAS general maths test by the two groups.

[1] https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Pre-testing_paper.pdf

KS1 maths scores and InCAS general maths scores will first be converted to Z scores for comparability. If there is little or no imbalance at pre-intervention then the substantive results will be the same for post-only (headline finding) and gain scores. If there is a substantive difference (an effect size of 0.05 or more) then the gain scores will form the basis of the headline finding.

However, while KS1 point scores are available for Years 4, 5 and 6, KS1 scores for Year 3 pupils are in descriptive categories. For this reason, we will analyse the two cohorts separately first and then analyse the two cohorts combined by converting the descriptive measures into scores equivalent to the NC levels.

For Years 4, 5, and 6 – a simple pre- post-test comparisons of mean scores will be used to determine the effect size, using KS1 maths point scores for the pre-test. For the Y3 cohort - because the pre-test scores are 4 skewed categories and the post-scores are normal interval scores, the results will be shown as the mean post-scores for each initial category. Two of the lower band categories (BLW and PKF) contained few pupils, so these are combined into one category.

For the combined analysis we will convert the descriptive measures for the Y3 cohort to a score equivalent to the National Curriculum levels. This is the system used by some of the schools in the trial in making comparisons between the old and new grading system. For example:

If Level 2b is the expected level for Y3 pupils, the new grading WTS (working towards expected standard) will be equivalent to Level 2c and the new PKF (pre-key stage foundation for the expected standard) will be equivalent to Level 1 (achieved Level 1) and so on (See table below). These grades will then be converted to the point score equivalent for each grade. This is the system used by some of the trial schools in making comparisons between the old and new grading system.

Table 3: Mapping of new and old KS1 point scores to levels

OLD NC level	New	Point scores
A = absent	A	
D = disapplied from NC	D	
W (Working towards level 1)	BLW = Below – corresponds with P-scales or NOTSEN	3
1	PKF = Pre-Key stage – Foundations for the expected standard	9
2c	WTS = Working towards expected standard	13
2b	EXS = working at the expected level	15
2a	GDS = Working at a greater depth within the expected standard	17

Imbalance at baseline

Presentation of ‘effect’ sizes for each measurement at outset.
Presentation of characteristics of schools in each group.

Missing data

Dong and Lipsey (2011) demonstrated that any missing values can create bias, even if attrition is balanced between comparator groups. And where such attrition is not random (as is most often the case) it can bias the estimate of the treatment effect, and the bias can still be large even when advanced statistical methods like multiple imputations are used (Foster & Fang 2004; Puma et al. 2009). Such bias can distort the results of statistical significant tests and threaten the validity of any conclusion reached (Shadish, Cook & Campbell 2001; Campbell & Stanley 1963; Little & Rubin 1987).

We cannot use existing data to substitute for data that is missing, since we have little or no knowledge of the missing cases, and missing data/cases are seldom random. Doing so will simply increase the potential for bias. We therefore present differences in pre-test scores (KS1 maths) between cases dropping out from both groups (where these are available).

In addition, we will report any missing data and compare the level of missing data to the number of hypothetical counterfactual cases needed to disturb the finding (Gorard et al 2017). The number of counterfactual cases will help determine whether the number of missing cases is large enough to alter/explain the findings (see explanation in section on Sample Size).

Fidelity analysis

The fidelity to the intervention will be assessed by comparing the outcomes of pupils with the number of sessions they attended (dosage). The number of sessions will be used as a continuous variable in the analysis. This will be zero for all cases in the control group.

In addition, we will perform Complier Average Causal Effect (CACE) analysis to estimate the effects for the subgroup of treatment students who comply with their treatment assignment. Specifically, compliance will be measured using the threshold of 30, which is the minimum number of sessions recommended. Essentially this will be the treatment group who complied vs controls who would have complied if given the treatment.

Data on the number of sessions conducted is collected from the Digital Tool, and is provided by the developers who have access to the Tool.

Secondary outcome analyses

The same ITT analysis as described above will be conducted for the secondary outcomes (Mental Arithmetics and Attitude to Maths) as for the primary outcomes.

Additional analyses

Three separate regression analyses will also performed: one for Y3 and one for the other year groups, and one combined. For Years 4, 5 and 6 a one-step multiple regression analysis will be conducted using KS1 scores and treatment group membership as the predictor, with post-test scores (InCAS general maths assessment) as the dependent variable.

For the Year 3, regression 3 dummy input variables representing the 4 categories of pre-test and treatment group will be used as predictors with post-test scores (InCAS general maths assessment) as the dependent variable.

A one-step multiple regression analysis will also be conducted using the combined scores as the converted KS1 scores and the treatment group as predictors and the InCAS general maths scores as the dependent variables.

Subgroup analyses

The main analyses will be repeated with only those pupils designated as EverFSM. Analysis will be performed for All pupils combined. In addition we also analyse the results for the Y3 and the other year groups separately.

Effect size calculation

'Effect' sizes will generally be calculated as Hedges' *g* for each variable based on the difference between mean post-test (and gain scores) for each variable. We will not report 'confidence intervals', but an interested reader can compute them if they wish as we will report the number of cases per group, and the effect size for each comparison.

Any 'effect' sizes for categorical variables will be based on post- odds ratios – or changes in odds where the groups are clearly unbalanced at the outset ('effect' size of 0.05 or more). All will be presented with the number of counterfactual cases needed to disturb the results.
.

Report tables

Executive Summary

Key conclusions
1. Impact of 10-week Maths Counts on the CEM digital InCAS general maths test scores
2. Impact of 10-week of Maths Counts on the CEM digital InCAS mental maths and attitude towards maths scores
3. Important factors for implementation
4. Main barriers to implementation
5. Possible further research question

Summary of impact on primary outcomes

Group	Effect size	Estimated months' progress	EEF security rating	EEF cost rating
Treatment vs. control – InCAS general maths				
Treatment FSM vs. control – InCAS general maths				

Comparison of trial schools and all primary schools in England (based on 2015 School Performance tables)

Variable	All primary schools (N= 16,766)		Trial schools (N=35)		
School-level categorical variables	**n**	**%**	**n**	**%**	
Academy converter	1,590	9.5	10	29	

Academy sponsor	757	4.5	6	17	
Community	8,124	48.5	8	23	
Voluntary controlled	2,233	13.3	7	20	
Voluntary aided	3,270	19.5	3	9	
Foundation	699	4.2	1	3	
Total			35		
***Ofsted Rating**					
Outstanding	93/1,034	9	3	9	
Good	641/1,034	62	25	71	
Requires improvement	268/1,034	26	4	11	
Inadequate	41/1,034	4	0	0	
No information	-	-	3	9	
School-level (continuous)	**n**	**[Mode]**	**n (missing)**	**[Mean/mode]**	
Size of schools	16,677	201-300	35	318 (101-200)	
Pupil-level (categorical)	**All Primary schools**	**Percentage**	**Trial schools**	**Mean (%)**	
Proportion achieving level 4 and above in reading, writing and maths	16,766	80	34 (1 school has no data)	76	
Proportion of pupils eligible for FSM	16,766	15.6	35	17.2	
Proportion of pupils with SEN	16,766	13.4	35	17.1	
Proportion of pupils with EAL	16,766	19.4	35	14.4	

Data for all school characteristics relates to January 2015 and was downloaded from the Department for Education 2015 Performance Tables (http://www.education.gov.uk/schools/performance/download_data.html). Ofsted ratings for intervention schools are taken from the latest inspection reports.
*National *data for Ofsted ratings is based inspections completed between 1 Jan 2015 and 31 March 2015. (https://www.gov.uk/government/statistics/maintained-schools-and-academies-inspections-and-outcomes-january-2015-to-march-2015)*

Comparison of pupil baseline characteristics

Variable	Intervention	Control	Odds ratio	Total
Characteristics of pupils at randomisation	n = 152	n = 153		
Proportion of boys	52	50.3	1.09	305

Proportion of pupils eligible for FSM	36.2	40.5	0.9	
Proportion of pupils with SEN	52.6	54.9	0.8	305
Proportion of pupils whose first language is not English	18.4	13.7	1.4	305
Proportion of pupils who are not White British (figures provided by schools)	26.3	29.4	0.86	305
Year group	%	%		n=
Proportion in Y2	1.3	0.7		3
Proportion in Y3	53.9	49		157
Proportion in Y4	27	35.3		95
Proportion in Y5	13.8	9.2		35
Proportion in Y6	3.9	5.9		15
Age in months	7.66	7.67		305
KS1 Maths				

Impact on general maths attainment

	N	KS1 points	SD	InCAS general maths	SD	Gain score	SD	Post-test 'Effect' size
Treatment								
Control								
Overall								

Impact on mental maths

	N	InCAS mental maths	SD	Post-test 'Effect' size
Treatment				
Control				
Overall				

Impact on attitude towards maths

	N	InCAS attitude towards maths	SD	Post-test 'Effect' size
Treatment				
Control				
Overall				

Impact on general maths attainment FSM eligible pupils

	N	KS1 points	SD	InCAS general maths	SD	Gain score	SD	Post-test 'Effect' size
Treatment								
Control								
Overall								

Impact on mental maths FSM eligible pupils

	N	InCAS mental maths	SD	Post-test 'Effect' size
Treatment				
Control				
Overall				

Impact on attitude towards maths FSM eligible pupils

	N	InCAS attitude towards maths	SD	Post-test 'Effect' size
Treatment				
Control				
Overall				

Regression results for the headline outcome

Model	InCAS general maths
R at step 1 (prior score and FSM status)	
R at step 2 (treatment group)	

Standardised coefficients for variables used in models in the regression analysis

Variable	InCAS general maths
Prior score	
FSM	
Treatment	

References

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.

Dong, N. and Lipsey, M. (2011) *Biases in estimating treatment effects due to attrition in randomised controlled trials: A Simulation study*. SREE Conference, 2011.

Foster, M. E. and Fang, G. Y. (2004). Alternatives to Handling Attrition: An Illustration Using Data from the Fast Track Evaluation. *Evaluation Review, 28*:434-464.

Gorard, S. and Gorard, J. (2016) What to do instead of significant testing? Calculating the 'number of counterfactual cases needed to disturb a finding'. *International Journal of Social Research Methodology*, 19, 4, 481-490.

Gorard, S. and See, BH and Morris, R. (2016) *The most effective approaches to teaching in primary schools*, Saarbrucken: Lambert Academic Publishing

Gorard, S., See, B.H. and Siddiqui, N. (2017) *The trials of evidence-based education: The promises, opportunities and problems of trials in education.* London: Routledge

Little, R. J. A., and Rubin, D. B. (1987) *Statistical analysis with missing data.* New York: Wiley

Puma, M.J., Olsen, R.B., Bell, S.H., and Price, C. (2009) *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.