Statistical Analysis Plan
FLASH Marking
Evaluator: Durham University
Principal investigator(s): Rebecca Morris

Template last updated: March 2018

PROJECT TITLE	FLASH Marking efficacy trial
DEVELOPER (INSTITUTION)	Meols Cop Research School, Southport
EVALUATOR (INSTITUTION)	Durham University
PRINCIPAL INVESTIGATOR(S)	Rebecca Morris
TRIAL (CHIEF) STATISTICIAN	Stephen Gorard
SAP AUTHOR(S)	Stephen Gorard and Rebecca Morris
TRIAL REGISTRATION NUMBER	Trial is not registered
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/flash-marking/

** 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.*

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	14 November 2018	Original version

Table of contents

Contents

Statistical Analysis Plan
FLASH Marking
Evaluator: Durham University
Principal investigator(s): Rebecca Morris

Education Endowment Foundation

Template last updated: March 2018

Introduction

FLASH marking is a two-year trial running across 103 secondary schools in England. Schools have been randomly assigned to either intervention (n=52) or business-as-usual (n=51) groups. The trial involves all Year 10 pupils in these schools, and will continue with the same cohort as they move in to Year 11 in 2019, and complete their GCSEs in 2020.

The intervention involves using a set of codes to mark and provide feedback on students' written work in English. Two teachers from each school are provided with training on FLASH marking by the developers. They then cascade this to their departments and the intervention is implemented by all English teaching staff. Follow-up training is provided at two more points across the trial along. Resources to support implementation are also provided by the developers. The study will measure both students' attainment at the end of Key Stage 4 (using English GCSE results) and teachers' views on whether FLASH marking has an impact on the time they spend marking.

The intervention is based on research in the field which highlights the importance of high-quality formative feedback for promoting students' attainment (Black and Wiliam, 1998; Christodoulou, 2017). There are, however, very few robust, large-scale studies which examine the impact of written feedback (Elliott et al., 2016). The FLASH marking trial aims to start developing this evidence base, exploring whether code marking and reducing the frequency of grading work can have an effect on students' outcomes and teachers' workload.

Design overview

Trial type and number of arms		Two-arm randomised control trial
Unit of randomisation		School
Stratification variables (if applicable)		n/a
Primary outcome	variable	Attainment scores in GCSE English Language and English Literature
	measure (instrument, scale)	GCSE English Language and English Literature (Grade 0-9)
Secondary outcome(s)	variable(s)	Teachers' reported workload
	measure(s) (instrument, scale)	Teacher questionnaire – number of hours reported on workload tasks

Sample size calculations overview

Please ensure all details are in line with the latest version of the protocol.

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Pre-test/ post-test correlations	level 1 (pupil)	0.69		0.69	TBC
	level 2 (class)				
	level 3 (school)				
Average cluster size		125		183	TBC
Number of schools	intervention	50		52	TBC
	control	50		51	TBC
	total	100		103	TBC
Number of pupils	intervention	6,250		9,233	TBC
	control	6,250		9,639	TBC
	total	12,500		18,872	TBC

** At present we do not have access to NPD data on this cohort of students meaning that we do not have baseline attainment/demographic background information and are therefore unable to complete some aspects of this table (e.g. number/percentage of FSM pupils in each arm). The application for NPD data was submitted in April 2018; as soon as the data are received, we can update this section. For school level data (see tables in appendix), information from the January 2018 Census has been used.*

*** The pre-/post- correlation at protocol and randomisation is actually based on prior studies using KS2 and KS4 attainment measures.*

We have calculated 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:

$$NNTD = ES*n$$

$$\text{Therefore, } n = NNTD/ES$$

$$\text{and } ES = NNTD/n$$

Based on Gorard (2016, 2018), NNTD of 50 can be considered a strong and secure finding. Using this as a working assumption for the FLASH evaluation, we would expect to detect an 'effect' size as low as 0.01 or 50/6,250 (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 practice, following randomisation, each arm of the trial now includes over 9,000 students, providing a very strong sample size for the detection of an 'effect' of almost any size.

Randomisation

Randomisation took place in Spring 2018 following recruitment of 103 schools to the project. As per the protocol, a simple randomisation process was used with an online randomisation programme (randomiser.org). As 103 schools were eligible for randomisation, a decision was taken over how to allocate the 103rd school (due to there being an odd number). The evaluation team decided to allocate 52 schools to the intervention group and 51 to the control group (as summarised in the consort diagram below). Over 9,000 students were allocated to each arm of the trial (see table above).

A national sampling frame was used and all 103 schools were randomised in a single batch. There was no stratification by region. The eight regional hubs are purely for training and convenience purposes and were determined after randomisation.

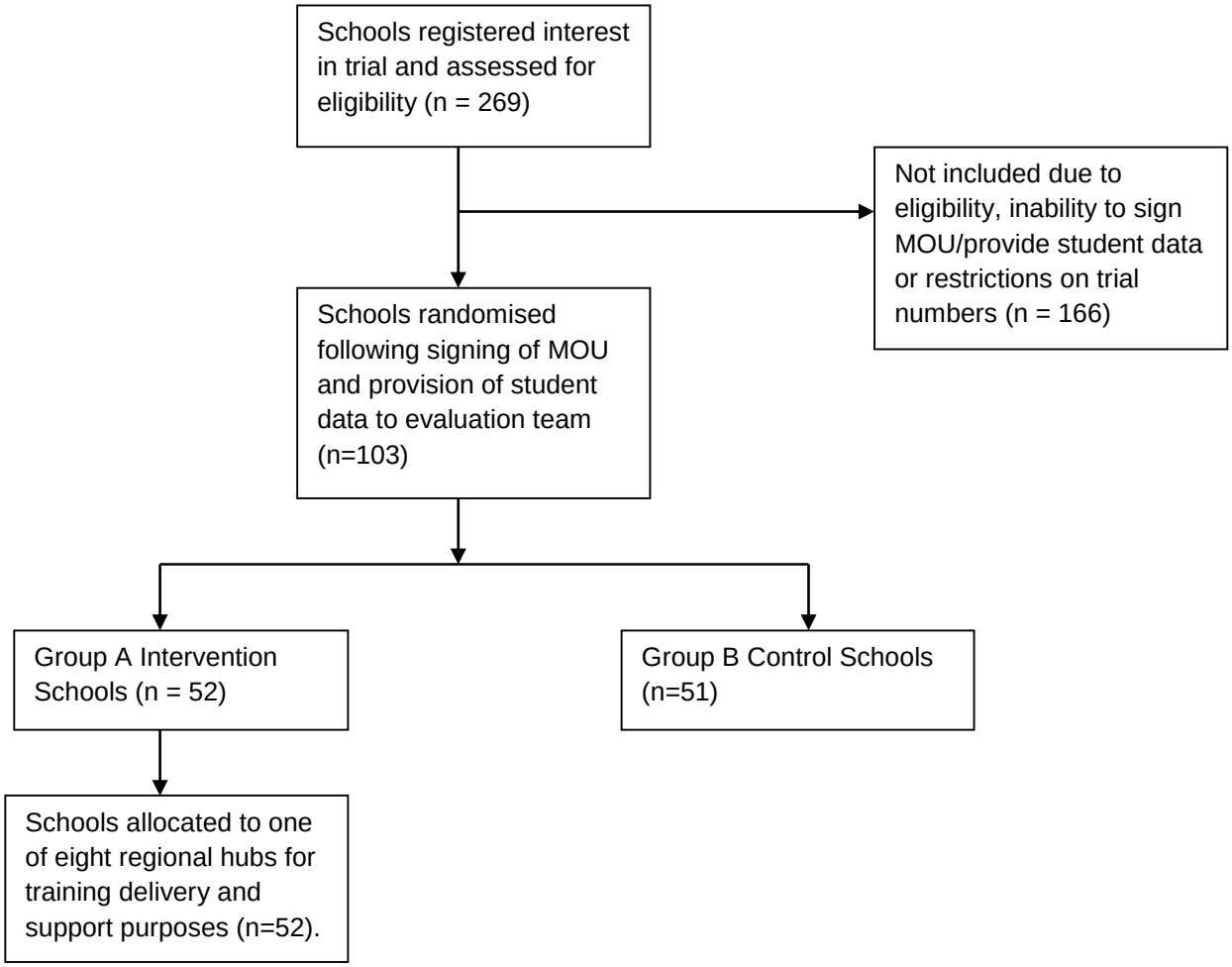

Figure 1: Consort diagram showing number of schools involved at each stage

Analysis

Primary intention-to-treat (ITT) analysis

The two types of schools included in the trial are:

A. Intervention schools that deliver FLASH marking
B. Control schools (using 'business-as-usual' approach to marking/feedback)

The analyses for the impact evaluation will be based on the difference in GCSE English Language and Literature scores between groups A and B for all schools where data is available. This will include schools that have dropped out of the trial and only pupils who have not withdrawn, in order to estimate the 'intention-to-treat' effect. The results will be presented as 'effect' sizes (see below) by dividing the difference in the means of the GCSE scores (using the compare means option in SPSS) between treatment and control, by the overall standard deviation of the exam scores.

If there is a substantive imbalance in the pre-intervention scores (an 'effect' size of 0.05 or more), then gain or progress scores will form the basis of the headline findings. They will be presented as 'effect' sizes based on gain scores calculated using the difference in the mean gain scores made between KS2 English point scores and the GCSE English Language/English Literature results by the two groups. For comparability in creating fair scores, the KS2 English scores and descriptive measures and GCSE English scores will be converted to standardized scores (Z scores).

All key results will be presented with a simple sensitivity analysis – the number of counterfactual cases needed to disturb the finding, or NNTD (Gorard and Gorard 2016). This can be computed by multiplying the achieved 'effect' size by the number of cases in the smallest group, and then comparing it to the number of missing cases. If the answer is clearly greater than the number of missing cases, then the finding cannot be due to biased missing data. The larger the answer is the more secure the finding is.

All analyses will be conducted using SPSS.

Secondary outcome analysis

For secondary outcome analysis, we will be focusing on estimating the effect of the intervention on teacher's workload. A baseline workload questionnaire was completed (before randomisation) by English teaching staff in trial schools. Questions were closely linked to the recent DfE workload survey (DfE, 2016) and included items asking teachers to report the number of hours that they spend on a different activities and a total for their last full working week. They are also asked to report perceptions of their workload. A second questionnaire will be administered during the Autumn term of the second year of the trial (October-November 2019) with a view to examining whether time spent on marking/feedback has altered for those teachers within FLASH marking schools.

The analysis here will focus on the differences between teachers in groups A and B at the two measurement points (first questionnaire and second questionnaire). Again, 'effect' sizes will be used along with gain scores (if there is imbalance between the two groups at the outset) to examine the numbers of hours that teachers state they are spending on different aspects of their job (including assessment and marking). There is likely to be a degree of missing data here due to some teachers leaving their schools between the two data collection points. We will examine these missing cases in order to establish whether there are differences in the pre-test scores of missing cases between the two groups. Categorical variables (i.e. items about teachers' attitudes to their workload) will also be analysed using odds ratios to examine changes between pre and post measurements.

Interim analyses

No interim analyses are planned during this trial.

Subgroup analyses

Estimates will examine the differential effects in four main subgroups. These are:

- Students who have ever received Free School Meals (FSM Ever6).
- Students with low attainment scores in Key Stage 2 attainment tests (bottom half of scores)
- Students with high attainment scores in Key Stage 2 attainment tests (top half of scores)
- Students' gender (male/female)

The first of these analyses reflects the EEF's commitment to improving outcomes for disadvantaged students. The final three subgroups are related to findings in the literature which suggest that some grading, marking or feedback practices may be more likely to benefit higher or lower attainers, or boys/girls (Elliott et al., 2016; Klapp, 2015).

Subgroup analyses will be conducted in the same way as the primary ITT analysis (see above).

Additional analyses

Two separate one-step multiple regression analyses will also performed. The first will be conducted using KS2 scores and treatment group membership as the predictors, with post-test scores in GCSE English Language as the dependent variable. The second will use KS2 scores and treatment group membership as the predictors, with post-test scores in GCSE English Literature as the dependent variable.

Imbalance at baseline

This will be assessed using NPD data and will be based primarily on the proportion of pupils eligible for FSM, and means for Key Stage 2 results (achieved in 2014-15 academic year) for the cohort participating in this trial. 'Effect' sizes will be presented. To cater for any initial imbalances between groups (i.e. if initial 'effect' size is 0.05 or more) we will present a gain scores analysis as well as post-test only. For the benefit of readers we present the pre-, post, and gain scores regardless of imbalance. An application to the NPD for this data was submitted in April 2018. In addition, other variables such as school type, Ofsted performance and geographical settings are presented in the tables below (See Appendix 1), based on school level information that is currently available.

Missing data

We will report and summarise the level of missing data in the primary and secondary outcome analyses. Missing data (even if attrition is balanced between groups) can bias the estimate of treatment effect (Dong and Lipsey, 2011). As such, we will not use existing data to substitute for data that are missing, since we have little or no knowledge of the missing cases, and they are seldom random. We will therefore present differences in pre-test scores (KS2 English) and any other available indicators of context (such as FSM), between cases dropping out from both groups (where these are available) to see whether these missing scores are imbalanced between groups. We will also run a sensitivity analysis, the number of counterfactuals needed to disturb the finding, and compare this with the level of missing data.

Compliance

In addition to the above, fidelity to the intervention will be assessed by comparing the outcomes of pupils with adherence to three key elements of the programme. These will be:

1. Number of training sessions (out of three) that staff from intervention schools attended

2. Confirmation that cascade training was delivered to Year 10 English teachers in each school prior to trial start in September 2018.
3. Reported compliance with FLASH marking elements across department and for first 15 months of trial - to be asked in a question to heads of department on the teacher questionnaire in Autumn 2019

We will run a correlation analysis using each measure of compliance with the student treatment outcomes. For (1), compliance will be assessed using number of training sessions attended as a continuous measure (with control schools having zero sessions by definition). For (2), confirmation that cascade training has been delivered to Year 10 English teachers will indicate 'compliance'; no confirmation or confirmation that it has been delivered only to *some* teachers will indicate 'non-compliance'. For (3), heads of department (HoD) will be asked to report the extent to which their teachers/departments have fully committed to the FLASH marking project and the implementation of the intervention. This will be done using a five point Likert scale question. These three correlation analyses will illustrate the extent to which the level of compliance is linked to any subsequent level of impact.

Achievement of 'baseline compliance' or not in relation to element (2) above will also be the variable used within a Complier Average Causal Effect (CACE) analysis in order to estimate the effects for the subgroup of treatment students whose schools complied with their treatment assignment (Nicholl, undated). Comparison is made of the average outcome of treatment pupils who were in compliant schools with control pupils in schools it is assumed would have complied if given the treatment (assuming same rate of compliance as for the actual treatment group). The effect sizes are recalculated using only the average results for cases deemed to be compliers in both groups. This is the same as scaling up the ITT 'effect' size using the Wald estimator.

Effect size calculation

As per current EEF guidance 'effect' sizes will be calculated using 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, standard deviations, and the effect size for each comparison.

For ease, the Hedge's g 'effect' size formula is written out as follows:

$$\boldsymbol{Effect\ size} = \frac{[\boldsymbol{mean\ of\ treatment\ group}] - [\boldsymbol{mean\ of\ control\ group}]}{\boldsymbol{standard\ deviation\ (pooled)}}$$

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). Headline results will be presented with the number of counterfactual cases needed to disturb the results.

Appendix 1

School-level information following randomisation

Regional spread of schools

Region	No. intervention schools	% of intervention schools	No. control schools	% of control schools	National % secondary schools (n = 3,436)
East Midlands	5	9.6	5	9.8	8.5
East of England	3	5.8	3	5.9	11.4
London	5	9.6	2	3.9	14.7
North East	0	0	2	3.9	5.2
North West	19	36.5	13	25.0	13.6
South East	6	11.5	9	17.6	14.9
South West	3	5.8	8	15.7	10.0
West Midlands	5	9.6	6	11.8	12.3
Yorkshire and Humber	6	11.5	3	5.9	9.4
Total	52	100 (rounded)	51	100 (rounded)	100

Geographical setting

Setting	No. intervention schools	% of intervention schools	No. control schools	% of control schools	National % secondary schools
Rural hamlet, village or town	9	17.3	7	13.7	14.5
Urban city or town	24	46.2	30	58.9	46.7
Major/minor urban conurbation	19	36.6	14	27.5	38.5
Total	52	100 (rounded)	51	100 (rounded)	100 (rounded)

Performance as judged by Ofsted

Most recent Ofsted Grade	No. intervention schools	% of intervention schools	No. control schools	% of control schools	National % secondary schools
Outstanding	16	30.7	12	23.5	24.0
Good	25	48.1	27	52.9	56.3
Requires Improvement	7	13.5	7	13.7	15.5
Inadequate	2	3.8	2	3.9	4.2
Information not available	2	3.8	3	5.9	N/A
Total	52	100 (rounded)	51	100 (rounded)	100 (rounded)

****Information taken from Ofsted.gov (September 2018).*

****Where schools have converted to academies recently, we have used the Ofsted grade from pre-academy status (n=18; 14 of these were previously Outstanding schools). It is also important to note that a number of these schools have not been inspected for over five years.*

Socioeconomic disadvantage

Percentage of Free School Meals eligible pupils	No. intervention schools	% of intervention schools	No. control schools	% of control schools	National % secondary schools
0-5%	11	21.2	6	11.8	17.5
5.1-10%	15	28.8	13	25.5	29.3
10.1-20%	16	30.8	23	45.1	32.4
20.1-30%	6	11.5	3	5.9	14.6
30.1-40%	2	3.8	3	5.9	4.6
40.1+	2	3.8	3	5.9	1.6
Total	52	100 (rounded)	51	100 (rounded)	100

****Data from January 2018 DfE School Census. National figure for FSM eligibility in secondary schools in 2018 is 12.4%.*

School type

School type	No. intervention schools	% intervention schools	No. control schools	% of control schools	National % secondary schools
Academy Converter	26	50.0	20	39.2	44.8
Academy Sponsor-Led	11	21.2	9	17.6	22.2
Community School	7	13.5	12	23.5	13.2
Foundation School	1	1.9	6	11.8	6.3
Free School	2	3.8	1	2.0	5.0
Voluntary Aided/Controlled	5	9.6	3	5.9	8.4
Total	52	100 (rounded)	51	100 (rounded)	100 (rounded)

****Data from January 2018 DfE School Census.*

Average prior attainment and KS4 English scores

	Intervention Schools	Control Schools
KS2 APS scores	28.3	28.3
Attainment 8 English element	9.9	9.8

****Data from 2016-2017 School Performance tables website*

Appendix 2

Tables for presentation of headline findings

Below are the tables to be used for presentation of headline findings (outcomes in GCSE English Language and GCSE English Literature).

Table 1: Post-intervention analysis of GCSE English Language outcomes

Group	N	KS2 English points	SD	Pre-'effect' size	GCSE English Language Score	SD	Post 'effect' size
Treatment							
Control							
Overall							

Table 2: Post-intervention analysis of GCSE English Literature outcomes

Group	N	KS2 English points	SD	Pre-'effect' size	GCSE English Literature Score	SD	Post 'effect' size
Treatment							
Control							
Overall							

Table 3: KS2 – KS4 progress in English

Group	N	English gain z-score	SD	'Effect' size
Treatment				
Control				
Overall				

References

Black, P., and Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. London: Granada Learning.

Christodoulou, D. (2017) Making good progress? The future of assessment for learning, Oxford: Oxford University Press

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

Elliot, V., Baird, J., Hopfenbeck, T., Ingram, J., Thompson, I., Usher, N., Zantout, M., Richardson, J., Coleman, R. (2016) A marked improvement? A review of the evidence on written marking, London: EEF

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

Gorard, S., 2018. Do we really need confidence intervals in the new statistics?. *International Journal of Social Research Methodology*, pp.1-11.

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. Assessment in Education: Principles, Policy and Practice, 22(3), pp. 302-323

Nicholl, J. (undated) *Complier Average Causal Effect analysis*, Available from: https://www.sheffield.ac.uk/polopoly_fs/1.418711!/file/JNicholls.pdf