Flash Marking

Durham University
Dr Rebecca Morris, Prof. Stephen Gorard,
Dr Beng Huat See and Dr Nadia Siddiqui

Education Endowment Foundation

Evaluation Summary	
Age range	Secondary – Key Stage 4 (Years 10 and 11)
Number of pupils	c. 12,500
Number of schools	100
Design	Randomised controlled trial with randomisation at school level
Primary Outcome	English Language and English Literature GCSEs
Protocol date	July 2019
Version	2
Changes to the protocol	- Updated the fidelity section following agreement in the SAP - The timing of the follow up teacher questionnaire was changed from Autumn 2019 to Spring 2020 as this better reflects the normal workload of teachers and the implementation of the programme.

Evaluation Protocol
Independent Evaluation of 'FLASH marking'

Background

The concept of providing high-quality formative feedback which encourages students to identify and develop their strengths and areas of improvement has had considerable attention in relation to research, policy and practice (see e.g. Black and Wiliam, 1998; Christodoulou, 2017; Hattie and Timperley, 2007). To date, however, there have few large-scale, robust studies that have examined the use of written feedback approaches on pupils' work (Elliott et al., 2016).

A number of studies cited here have explored the use of grades and/or comment-based feedback on students' written work. The evidence seems to suggest that grade-only marking does not have a positive impact on attainment (Elliot et al., 2016). Comment-only feedback (with no grades) was also found to be less beneficial when compared with both a comments and grades approach (Gorard and Smith, 2005). Another study, however, found that detailed, descriptive feedback was more effective for students' progress when given alone, unaccompanied by grades (Lipnevic and Smith, 2009).

The use of a combination approach (of comments and grades) has produced less clear research results. Work by Klapp (2015) showed that boys and lower attaining pupils did less well when receiving grades and comments whereas there was a positive long-term effect for girls. This study, based in Sweden, using data from 30 years ago, focused on the feedback students received in primary school and their latter attainment in secondary school. While the findings are important, it is not clear how applicable the findings are to a different context and a shorter term approach for understanding the impact of grades/comments on academic outcomes. A smaller scale study of middle school science achievement in America (Zhang and Misiak, 2015) found that a combination of both written comments and grades was most effective for promoting student attainment. The authors concluded, however, that in order to move students' learning forward, the comments must be highly relevant to the academic standards being assessed.

In addition to considering the benefits of different forms of formative assessment, it is also important to consider the costs too. Concerns surrounding the workload implications of extensive Assessment for Learning (AfL) practices combined with very little evidence that they work have led to calls for a pared back approach, particularly in relation to the amount of time that teachers must spend marking written work (Independent Teacher Workload Review Group, 2016). There is a renewed interest, supported by government, Ofsted and the unions, in finding more efficient and effective methods for providing valuable feedback. While this has led to some interesting and innovative approaches, more rigorous testing of these is needed in order to assess their effectiveness and scalability.

It is also important to situate this evaluation within the current policy context regarding assessment and schools. The removal of National Curriculum levels has provided additional autonomy to schools, allowing them to develop their own methods for supporting and measuring attainment and progress (Gibb, 2015). Many in the profession have welcomed this removal of a prescriptive approach but there is also an acknowledgement that many schools are not clear about the best approaches for replacing levels. The government have suggested a range of alternative methods (Lilly et al., 2014) although there is often a lack of rigorous evidence underpinning these. Changes to the English GCSE grading system and the exam specifications (in relation to content and a shift to a linear rather than modular model) (Ofqual, 2015) mean that many schools are looking for improved ways to develop students' knowledge and skills in preparation for the exams and later life. The use of good quality formative assessment of written work may be an important element in supporting this.

Intervention

FLASH (Fast Logical Aspirational Student Help) marking is a school-developed feedback approach in which teachers give skills-based comments rather than grades in Key Stage 4 English. The approach has been devised by staff at Meols Cop High School in Southport. The FLASH feedback uses language from the highest grades descriptors in GCSE English Language and English Literature. They are presented on students' written work using a code system which identifies where pupils have done something well but also where there are areas for improvement and development.

The codes are broken down into sections depending on whether they are describing skills focusing on reading or writing skills. During the development stages, they have been used by both teachers and pupils in English lessons. They are used to inform target-setting for future pieces of work.

The developers have outlined a number of reasons for devising and implementing FLASH marking. First, it is hoped that in having a clear skills-focus during teacher, peer and self-assessment that students' will be able to clearly understand how they can improve and develop in English. Second, by removing the use of grades, it is hoped that pupils will focus on the skills and knowledge that they need to develop in order to improve their attainment. Third, the use of codes is believed to 'speed-up' the marking process and ensure that children receive feedback sooner. Closely linked to this is the aim of reducing teacher workload in relation to assessment. Finally, tracking of pupils and classes through the coding system can highlight where there are skills/knowledge gaps and these can be addressed. The intervention has been piloted in a small number of schools so far.

Two English teaching representatives from each 'intervention' school will attend three training sessions (run by the Meols Cop development team). These staff will then be responsible for cascading the training to all other staff within their department. The school subject leader for English will be one of the staff members responsible for attending the training. They will be asked to select one other English teacher to attend too.

Details of the three training sessions are below:

- **Training session 1** will run for a full day in June 2018. It will focus on the principles of FLASH marking, lesson planning, modelling the assessment approach, using demonstration videos and the use of the web portal as a support mechanism for trial participants.
- **Training session 2** will be a three-hour session in November 2018. This session will focus on moderation of assessments and the use of the codes in future planning. Support will be offered through the use of demonstration videos and group discussion.
- **Training session 3** will run in September 2019, as schools enter the second year of the trial. This session will run from between two-four hours and will have two parts. The first part will be used as 'refresher' session for schools that need this or who have new staff who need initial training. The second part will be for the two nominated English staff from participating schools and will provide an opportunity to discuss progress and to receive further support from the delivery team.

Delivery of the training will take place in geographical 'hubs'. This will allow groups of schools from the same area/region to come together and receive the training as a group. It is anticipated that there will be approximately six hubs across England, each accommodating 8-10 participating schools. Locations for the training will be determined once recruitment is underway and the development team have a clearer picture of where participating schools may be located.

Additional support will be available to English departments throughout the period of the trial. This will be provided by the development team throughout the duration of the trial, and will include: frequent contact between schools and the development team; visits from the development team if needed; and the use of the web portal as a way of providing videos, models of assessed work and curriculum resources. It may involve communication between the subject leader and development team and/or might involve the whole English department or specific teaching staff, depending on what the school's needs are.

Research Plan

Research questions

The primary research question is:

- How effective is FLASH marking in improving the GCSE English outcomes for Key Stage 4 pupils?

We will also be seeking to address the following question:

- How effective is FLASH marking in reducing the marking and assessment workload for teachers of Key Stage 4 English?

Design

This will be a two-arm RCT with randomisation occurring at the school-level. This will reduce the possibility of diffusion which could occur with an in-school design given the cascade training model that will be used. A total of 50 schools (assuming that we reach the target number of 100 participant schools) will be allocated to either the 'intervention' group or the 'control' group. Those in the 'intervention' group will receive the FLASH marking programme, including the training and cascading of training and will implement the approach with their Year 10 classes from September 2018; the intervention will continue with these same groups until the end of their GCSE course in Year 11. Those

in the 'control' group will operate on a 'business as usual' approach, continuing with their usual methods of marking and feedback.

Randomisation

Schools will only be eligible for randomisation after:

- Signing a Memorandum of Understanding
- Providing pre-test data requested in the Memorandum of Understanding (including pupil UPNs, prior attainment and demographic data and providing teacher contact details).
- Completion of the first staff workload survey. All English department staff will be asked to complete this short online survey.

Randomisation will be conducted at the school level once all of the above data has been collected. The process of randomising schools to each group will be carried out independently by the evaluation team at Durham University, and observed by at least one colleague. Each school will be assigned a random number, created in Excel, and the list of schools will then be sorted into number order. The first half will be the treatment schools. If there is an odd number of schools then the last school will be assigned to treatment if a further random binary digit is 0 (rather than 1).

Incentives

Schools allocated to the intervention group will receive £700 and those allocated to the business-as-usual group will receive £1000. Incentives have been offered to both groups in order to try and reduce the potential for schools to 'drop out'. For the intervention schools, it is hoped that the incentive can facilitate the release of two English staff to attend training. For both the intervention and business-as-usual groups it is anticipated that the incentives will also encourage schools to continue in supporting the evaluation with the provision of survey data, and permitting visits and communication with participating schools. The sums will be paid in instalments following completion of certain stages/requirements of the evaluation. For both groups, 50% of the financial incentive will be paid following randomisation and a further 50% following completion of the final staff workload survey (in Spring 2019).

Participants

All state-funded mainstream secondary schools (with a Key Stage 4 cohort) in England are eligible to participate in this trial. Recruitment will be carried out by the developers, Meols Cop High School, but will be supported by the evaluation team where needed. It is expected that schools from across the country will wish to participate and the developers intend to create a number of 'hubs' in different geographical regions in order to support training and facilitation of the intervention. Schools that are interested in being part of the trial are currently being asked to contact lead developer, Sarah Cunliffe at Meols Cop High School.

Year 10 cohorts (in September 2018) in the participating schools will be the focus of this study. These same children will continue to be part of the study for two academic years i.e. until the end of their GCSE English Literature and Language courses. It is expected that teachers in schools will be using the FLASH marking approach with the children over two academic years. This should allow enough time for teachers to familiarise themselves with the approach in the first year, and embed the approach in their practice in their second year. With the exception of those who opt-out of the study, all children within these cohorts will form part of the trial.

Sample size calculations

The pupil sample size calculation is based on the assumption that there will be 100 schools participating in the project. The developers have suggested having six regional 'hubs' to provide training and support for the intervention schools. If spread equally, this would mean that there would be an average of about eight schools represented in each hub. All pupils in one-year group will take part in the trial (Year 10 in September 2018) – with the exception of those who opt-out. Experience suggests an average year group of 125 pupils, meaning that there will be an approximate total of 12,500 pupils involved in the trial as a whole. Each arm of the trial will include approximately 6,250 children.

Whatever the final number of schools involved, it is vital that all cases are retained. The evaluation team are happy to attend and address recruitment events in order to explain how the trial will work and to stress the importance of committing for the duration of the evaluation even if schools do not continue with the intervention.

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 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, 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.

Outcome Measures

The primary outcome for this evaluation is students' attainment in GCSE English Literature and GCSE English Language. Both of these measures are currently 100% examination based and will be taken by the vast majority of students in all participating schools. The external nature of these assessments means that we do not have to rely on teacher assessment within schools. Equally, the fact that the students would be taking these examinations anyway (whether or not they were part of the evaluation) reduces the burden of over-testing. For a baseline measure we will use pupils' Key Stage 2 English scores. Correlation with GCSE results has been calculated by the evaluation team to be at around 0.69.

A secondary outcome from the project involves the focus on teacher workload, in particular workload that is related to marking and assessment. In order to measure these we will administer a short (approximately 10-15 minute) online teacher workload survey prior to randomisation and to the intervention being delivered. This will be completed by all English teachers within the 100 participating schools. A second survey will be given to the same group of staff in the Spring term of the second year of the trial (Spring 2020). While numbers of English teachers per school will vary, we anticipate that there will be approximately 600-700 working across the 100 schools. As completion of the first survey is a condition of eligibility for randomisation (and is needed prior to receiving the first incentive payment), we anticipate a good response rate.

The survey will use items similar to those found in the DfE workload survey. This was most recently used in 2016 and had a total of 3,186 respondents (DfE, 2016). The survey asks teachers to self-report the amount of time that that they spend on the different activities that form part of their job. The focus of the survey is on issues of marking, assessment and feedback but it is important to also gain a sense of what other activities teachers spend time on in order to understand the balance of their workload, and for us to see if this alters following involvement in the FLASH marking study. A copy of the questions used in the DfE's survey can be found in the Workload Survey report (DfE, 2016).

To supplement the survey data, semi-structured interviews will be conducted with a range of English-teaching staff during our visits to the 15 case study schools. Information collected from these interviews will give further qualitative insight in to teachers' views on workload in relation to assessment and feedback, and will also provide detail of shifts in workload following the introduction of FLASH marking.

Fidelity measure

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 Spring 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.

Other data

Data on pupils' prior attainment (KS2 results) and background characteristics such as age, date of birth, sex, current FSM eligibility and Ever FSM will also be collected as a routine part of the pre-testing. These data will be uploaded for all pupils at the outset from each school's SIMS or similar. These will eventually be linked via UPN to the individual post-test scores.

Analysis plan

Primary outcome

Assuming the two groups are reasonably balanced (in terms of 'effect' size) at the outset in terms of pre-intervention scores, all headline analyses will be based on the difference between groups on post-test scores only - the Age Standardised scores of PTM. Comparisons will be made between the FLASH marking group and the business as usual group, using Hedge's effect sizes. If the two groups are not reasonably balanced (initial ES of 0.05 or greater) at the outset then gain scores from pre to post-test would form a more useful basis for calculating the effect sizes. Either way, the report will include the pre-test, post-test and gain score analysis so that readers can compare them.

Attrition

There is no remedy for missing cases or values, and all attrition can lead to bias. Therefore, attrition must be as little as possible and we take great care to prevent school dropout and to chase up missing data, even where pupils have moved schools since randomisation. We will then assess the strength of any effects founds by comparing them with the number of cases missing data – using the procedure of number needed to disturb the finding or NNTD (Gorard et al. 2017).

Subgroup analysis

Subgroup analyses will also be conducted for only those pupils selected in terms of prior attainment (at or above, or below, average), gender and FSM eligibility. These will each take the same form as the headline figures. A second model will analyse first stage subgroup variables using interaction terms.

Further analyses

Multivariate regression analyses will also be conducted using post-test scores as the dependent variable and prior test scores as predictors, in the first stage. In the second stage, a binary variable representing the treatment or control group will be added to see how much variation in post-test scores is explained once prior attainment is accounted for. This will yield the same substantive result as the gain score analyses (above). The effect sizes for each stage will be the amount of variation explained. This is done to be in line with the EEF's preferred analytical approach.

Fidelity measure
See section above.

Process Evaluation

The process evaluation will enable us to collect important information about the delivery of the training and the implementation of the intervention in schools.

The data collection for this part of the project will form the majority of the fieldwork carried out overall. We aim to provide detailed and formative evidence on all phases and aspects of the intervention from the selection and retention of schools through to the initial training and delivery of the intervention, and finally to evaluating of the outcomes. The information collected will help us to assess the fidelity to treatment, and the perceptions of participants involved including enthusiasm or resistance, and to advise on improvements and issues for future implementation and evaluation.

The implementation and process evaluation will predominantly focus on 15 case study schools. These will include approximately 12 intervention schools and three control group schools. Schools will be selected with a view to achieving a range in terms of region, size, and school intake demographics. Of course, schools' willingness to be involved in this part of the evaluation is key too. The process evaluation will be explained to schools during the recruitment process with an expectation that schools will be willing to participate wherever possible. We intend to visit each case study school twice per year during the course of the trial (four times in total). During these visits the evaluation team will interview a range of staff and students about marking and assessment practices, carry out observations of teaching and feedback in Key Stage 4 English lessons, and look at evidence of marking practices in students' exercise books. We would anticipate interviewing both English staff that received the FLASH training plus one other member of staff who has received the cascade training. In business-as-usual schools we will speak to at least two English-teaching staff from each department. We will conduct one focus group interview in each case study school in order to speak to Year 10/11 students about their experiences and perceptions of marking and feedback in English.

All aspects of the data collection for the process evaluation will be conducted in a simple, straightforward and non-intrusive way. Our interviews with staff and students will be short and focused, ensuring that high-quality data are collected but that we do not take up too much valuable time. Our interview schedule will be developed in due course, and will be finalised following the start of the trial. It is likely that some aspects from the training sessions and beginning of the trial period may inform some of the areas that we wish to explore via the school-based interviews. The dates and timings of visits will be agreed in advance with the developers and the schools. Our observations of teaching and marking in exercise books will be conducted in a non-judgemental, supportive and purposeful way. We will be focused on issues linked to assessment and feedback, and the FLASH marking trial, and will not be looking to comment on other aspects of teaching and learning.

The process evaluation will also include a brief staff attitudes survey which will explore English teachers' experiences of and views of marking and feedback in their subject. It is anticipated that we will combine this staff attitudes survey with the second workload survey; this will mean that it can be completed by staff in both intervention and business-as-usual schools although will specifically target teachers of the treatment year group (i.e. Year 11 at the time of the survey). It is difficult to know exactly how many teachers in each school might have a Year 11 English class in the second year of the trial; however, it is likely that there will be approximately five teachers per school working with this year group which would yield around 500 staff attitude survey responses. There will also be a small number of additional survey items for Heads of Department in order to gain their specific views on key issues associated with assessment. Again some items will be drawn from the DfE Workload survey

in order to explore more general perceptions and attitudes about assessment. There will also be some items for teachers in intervention schools which specifically ask about the FLASH marking training and implementation. The staff attitudes survey will drawn-up in conjunction with the development team and sample questions will be published as part of this protocol. The survey will be completed by teachers in Autumn term 2019 along with the second workload survey.

Finally, a student survey will also be conducted in case study schools during the second year of the trial. This will explore issues of responsiveness, reach, quality and programme differentiation (in relation to what was done before). Through it we aim to develop an understanding of children's attitudes and experiences in relation to marking and feedback in English lessons during participation in the trial. As discussed above, we anticipate having 15 case study schools an average of 125 children per year group, approximately 1875 children will be asked to complete the survey. The survey will be paper-based and will only take a short amount of time to complete (approximately 15 minutes). The evaluation team will arrange for the printing, administration, delivery and collection of all surveys in order to ensure a completion rate of as near to 100% as possible. Exact details of the survey items are yet to be decided and will be developed in the coming months in conjunction with the Meols Cop team.

The data collected as part of the process evaluation aim to assess:

- teachers' response to training, including attendance rates
- the fidelity of training
- teachers' delivery of the intervention
- whether the teams understand the process and purpose
- the contents and use of any materials
- changes in teacher behaviour
- staff and students' views of the intervention
- any apparent impact on children's behaviour and attitudes
- possible indication of contamination or diffusion

Costs

The costs of the trial will be the amount that the school would incur if they were to implement the intervention in their school. This will be calculated per pupil using the following estimates:

Cost of setting up

- Cost of delivering training to English teaching staff
- Cover for teachers to attend the three training sessions
- Any other costs

Cost of delivery

- Any teaching materials and resources associated with FLASH marking
- Day rates for teaching staff
- On-going monitoring and support

Other non-monetary costs

Time taken away from regular lessons for attending training, organising the administration and collection of the training/cascading materials. Once set up, it is not anticipated that staff will lose time in the classroom. Indeed, it is possible that marking time will be reduced for staff and, if this is the case, we are interested in how this time is being used instead. Information on this will be collected

with input from the project team and via surveys and interviews with teaching staff.

Ethics

A number of ethical considerations will be taken in to account during the planning, implementation and reporting stages of the trial. The BERA ethical guidelines (2011) and Durham University's strict research Code of Practice will be adhered to at all times. Ethical approval will be gained from Durham University's Research Ethics Committee prior to the trial beginning.

Parental consent will be sought in relation to the use of students' data in the analysis and reporting of the trial outcomes. This will involve an 'opt-out' consent form and information letter whereby parents and carers can choose to have their child's information withdrawn from the final analyses and reporting.

All schools, staff and pupils involved in the trial will remain anonymous. Participant information will be treated confidentially and all participants in interviews and observations will be informed of their right to withdraw at any stage. No person or school will be identifiable in the reporting of this trial.

The GCSE exams will be occurring irrespective of this trial and so will add no additional burden in terms of testing.

The data used from the National Pupil Database will contain potentially sensitive information about the children involved in the trial. For the purposes of this work, however, we will only need Tier 2 data (rather than the more sensitive Tier 1 information) in order to conduct the analyses detailed above. No pupils will be identifiable and all secondary data will be kept in secure environments, following the strict data protection requirements of the NPD and Durham University.

Durham University's data protection policy is publicly available at:
http:www.dur.ac.uk/resources/data.protection/dataprotectionpolicy.pdf

"Durham University is committed to protecting the rights and freedoms of individuals in accordance with the provisions of the Data Protection Act 1998. The requirements to which University staff and student who process personal data must adhere are set out in the University's Data Protection Policy"

Personnel

Evaluation team - School of Education, Durham University

Dr Rebecca Morris will be responsible for the final delivery of all outputs and meeting deadlines. She will lead in the day-to-day organisation of the study, arranging fieldwork, communicating with the EEF, collecting data for the impact and process evaluations, analysis and report writing. Her role at Durham University is dedicated to EEF evaluations and similar opportunities.

Dr Beng Huat See will support with all aspects of the fieldwork, data collection, analysis and report writing. She has led a number of previous EEF evaluations and her work is dedicated to such projects and other similar opportunities.

Dr Nadia Siddiqui will assist with fieldwork, data collection and cleaning, arranging fieldwork, and assist with communicating with the study researchers, analysis and report writing. Her role is dedicated to EEF evaluations and similar opportunities.

Professor Stephen Gorard will be responsible for the design and analyses of the impact evaluation, and will assist with all other elements, including report writing.

Research assistants will be employed as and when needed for parallel fieldwork and to relieve pressure on the principal researchers, including cleaning and preparing data, entering the staff workload data, coding, parallel fieldwork, and literature searches.

Project team – Meols Cop High School, Southport

The intervention team from Meols Cop High School will be responsible for school recruitment, collecting opt-out consent from parents, on-going relationships with schools and keeping parents informed of the intervention (if necessary). They will be responsible for the training of staff, supporting the cascading of training and the delivery of the intervention. They will work with schools to collect baseline data. These tasks will be conducted with the support of the Durham evaluation team.

Risks

As a team, we have conducted scores of evaluations, and have always completed them successfully and on time. The biggest risk to the evaluation probably stems from temporary unavailability of any of the evaluating team, through illness, for example. To a very great extent they can substitute for each other. If Rebecca Morris was unavailable for any reason, then Stephen Gorard or Beng Huat See would lead, and Nadia Siddiqui would take on more responsibility for this project. Prof. Carole Torgerson, a senior member of the evaluation team at Durham is also available if needed. A pool of researchers with the relevant skills and experience is available within the School of Education, Durham University.

Diffusion of the intervention is a potential risk to the study as a whole. If some control schools are using or choose to adopt similar practices to the intervention group during the trial period then the comparison between groups is weakened. It will be important, therefore, to be aware of the assessment practices of the control schools prior to the trial starting and monitor these throughout.

It is possible that it may be challenging to recruit 100 schools to this project. However, to counter this issue, the development team have stressed flexibility in being able to work across the country, increasing the pool of potential participant schools. Equally both the development team and evaluation team are willing and able to stress the potential benefits to schools of involvement in a trial such as this. The intervention has also been designed in a way that it will not create too much disruption or additional work for staff in schools; indeed, it can be emphasised that the FLASH marking process could help to reduce time spent on assessment/feedback for English teachers.

The most likely risk is schools dropping out or not cooperating in providing data or conducting the survey. In order to try and prevent this from happening incentives are being offered to both intervention and control group schools. The evaluation team will aim to develop strong working relationships with both the developers and schools involved in the project. They will offer support where needed to schools and maintain close communication in order to ensure that they are aware of their role and any requirements (e.g. submission of data, taking of tests, distribution of surveys) at each stage of the trial. Contact information will also be kept up-to-date so that we can intervene and offer support and reassurance should it be needed. A regular newsletter will go out to participating schools with updates, project requirements and contact details for the developers and evaluation team.

Indicative Timeline

	2017-2018				2018-2019						2019-2020						2020-2021			
	Spring term		Summer term		Autumn term		Spring term		Summer term		Autumn term		Spring term		Summer term		Autumn term		Spring term	
Recruit schools																				
Collect pre-test data from schools																				
English teachers complete first workload survey																				
Randomise schools																				
Confirm teachers and class lists																				
First training session and cascading																				
Teachers use FLASH marking with Year 10 classes																				
Second training session and cascading																				
Classroom observations and interviews																				
Teachers continue using FLASH marking with Year 11 classes																				
Third training session																				
Second teacher workload survey																				
Classroom observations and interviews																				

Year 11 students take GCSE exams																				
Collect post-test data																				
Analysis and report writing																				
Final report complete																				

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

DfE (2016) Teacher Workload Survey 2016, DfE: London. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/592499/TWS_2016_FINAL_Research_report_Feb_2017.pdf

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

Gibb, N. (2015) Assessment after levels. Speech, 15th February 2015. Available from: https://www.gov.uk/government/speeches/assessment-after-levels

Gibson, S., Oliver, L. and Dennison, M. (2015) Workload Challenge: analysis of teacher consultation responses research report. London: Department for Education. Available from: www.gov.uk/government/uploads/system/ uploads/attachment_data/file/401406/RR445_-_Workload_ Challenge_-_Analysis_of_teacher_consultation_responses_ FINAL.pdf

Gov.uk (2016) Ofsted Inspection Myths https://www.gov.uk/government/publications/school-inspection-handbook-from-september-2015/ofsted-inspections-mythbusting

Hattie, J. and Timperley, H. (2007) The Power of Feedback, Review of Educational Research, 77(1): 81-112

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

Lilly, J., Peacock, A., Shoveller, S. and Struthers, D. (2014) Beyond Levels: alternative assessment approaches developed by teaching schools Research Report, London: DfE. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/349266/beyond-levels-alternative-assessment-approaches-developed-by-teaching-schools.pdf

Lipnevich, A. and Smith, J. (2009). Effects of differential feedback on students' examination performance. Journal of Experimental Psychology: Applied ,15(4), pp. 319–333

Ofqual (2015) GCSE changes: a summary. Available from: https://www.gov.uk/government/publications/gcse-changes-a-summary

Smith, E. and Gorard, S. (2005) 'They don't give us our marks': the role of formative feedback in student progress, Assessment in Education, 12 (1), pp.21-28

Zhang, B. and Misiak, J. (2015). Evaluating three grading methods in middle school science classrooms, Journal of Baltic Science Education, 14(2), pp.207-21