



Education
Endowment
Foundation

Embedding Formative Assessment

Evaluation report and executive summary

July 2018

Independent evaluators:

Dr Stefan Speckesser, Johnny Runge, Francesca Foliano, Dr Matthew Bursnall, Nathan Hudson-Sharp, Dr Heather Rolfe and Dr Jake Anders



The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.



For more information about the EEF or this report please contact:

Danielle Mason
Head of Research

Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP
p: 020 7802 1679
e: danielle.mason@eefoundation.org.uk
w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from the National Institute of Economic and Social Research: Dr Stefan Speckesser, Johnny Runge, Francesca Foliano, Dr Matthew Bursnall, Nathan Hudson-Sharp, Dr Heather Rolfe and Dr Jake Anders.

The lead evaluator was Dr Stefan Speckesser.

Contact details:

National Institute of Economic and Social Research

2 Dean Trench Street, Smith Square London, SW1P 3HE

Email: enquiries@niesr.ac.uk

Tel: 0207 222 7665

Contents

Executive summary.....	4
Introduction	6
Methods	9
Impact evaluation	17
Implementation and process evaluation	24
Conclusion.....	41
References.....	43
Appendix A: EEF cost rating.....	44
Appendix B: Security classification of trial findings.....	45
Appendix C: Consent Letter	46
Appendix D: School Information Sheet.....	47
Appendix E: Memorandum of Understanding with schools.....	50
Appendix F: Details of Analysis Code.....	56
Appendix G: Multiple Imputation	57
Appendix H: ICC and Total Variance	61
Appendix I: Statistical Analysis Plan.....	62
Introduction	62
Aims and Objectives	63
Study design.....	65
Randomisation	68
Follow-up	69
Analysis	71
Report tables	79
Appendix J: Execution of Randomisation	80

Executive summary

The project

Embedding Formative Assessment (EFA) is a whole-school professional development programme aiming to embed the use of effective formative assessment strategies. Dylan Wiliam and Siobhan Leahy, who are experts in formative assessment, designed the intervention and associated materials. The Schools, Students and Teachers network (SSAT), an independent membership organisation, delivered the project.

Schools received detailed resource packs to run monthly workshops known as Teacher Learning Communities (TLCs). Each TLC was expected to last 75–90 minutes. All teaching staff were involved and split into groups comprising 8–14 people. TLC agendas and materials focused on five key formative assessment strategies: ‘clarifying, sharing and understanding learning intentions’; ‘engineering effective classroom discussions and activities’; ‘providing feedback that moves learning forward’; ‘activating learners as instructional resources for one another’; and ‘activating learners as owners of their own learning’. Within each of these high-level concepts, the TLC handouts introduced multiple formative assessment techniques for teachers to consider.

In-between workshop sessions, teachers were expected to conduct peer lesson observations and provide feedback to each other. Each school appointed a lead teacher who attended an initial training day and received ongoing implementation support from an SSAT Lead Practitioner. This included a mixture of visits, phone calls, e-mails, and access to an online community.

The project was a randomised controlled trial. It was an effectiveness trial, which tested whether the intervention worked under everyday conditions in a large number of schools. One hundred and forty secondary schools participated during the 2015/2016 and 2016/2017 academic years. The primary outcome was Attainment 8 GCSE scores for the 25,393 pupils who were in Year 10 (aged 14–15) at the start of the trial. The process evaluation involved a combination of methods, including interviews, focus groups, surveys of intervention and control schools, and observations of the launch day, some TLCs, and the celebration day.

Key conclusions

1. Students in the Embedding Formative Assessment schools made the equivalent of two additional months’ progress in their Attainment 8 GCSE score, using the standard EEF conversion from pupil scores to months progress. This result has a very high security rating.
2. The project found no evidence that Embedding Formative Assessment improved English or Maths GCSE attainment specifically.
3. The additional progress made by children in the lowest third for prior attainment was greater than that made by children in the highest third. These results are less robust and have a lower security rating than the overall findings because of the smaller number of pupils.
4. Teachers were positive about the Teacher Learning Communities. They felt that these improved their practice by allowing valuable dialogue between teachers, and encouraged experimentation with formative assessment strategies.
5. The process evaluation indicated it may take more time for improvements in teaching practices and pupil learning strategies to feed fully into pupil attainment. Many teachers thought that younger students were more receptive to the intervention than their older and more exam-minded peers.

EEF security rating

These findings have a very high security rating.

This trial was an effectiveness trial, which tested whether the intervention worked under everyday conditions in a large number of schools. The trial was a well-designed, two-armed randomised controlled trial. The trial was well powered. Relatively few pupils (1.9%) who started the trial were not included in the final analysis of Attainment 8 GCSE score. The pupils in the Embedding Formative Assessment schools were similar to those in the comparison schools in terms of their prior attainment.

Additional findings

The impact on Attainment 8 scores was 0.10, when measured as an effect size. This is roughly equivalent to an improvement of one GCSE grade in one subject.

The findings about Attainment 8 are statistically significant at the 10% level. The effect sizes for all specifications and sub-groups were consistently positive, which adds further weight to the conclusion that the intervention had a positive impact.

The subgroup analysis of students eligible for Free School Meals in the previous six years found a small effect size (0.07). This result needs to be treated with caution due to the small sample size, but does not suggest a different impact on this group.

The process evaluation found that schools generally achieved the broader aim of increasing the use of formative assessment and facilitating dialogue between teachers, which was thought to feed into improvements in teaching quality. However, implementation varied significantly as schools adapted the programme to fit their specific context and past experiences. Variation occurred particularly in relation to the format and structure of TLCs as well as the use and frequency of peer observations.

The formative assessment content was usually considered to be very similar to existing approaches already being implemented in schools, but what set the intervention apart was the sustained focus on reinforcing existing practices, and the TLC mechanism was considered a novel approach in many schools. However, some schools had been involved in programmes which affected implementation and impact. In particular, previous involvement in another intervention, Teacher Effectiveness Enhancement Programme (TEEP), strongly influenced the delivery of EFA. Subsequent exploratory analysis for schools not previously exposed to TEEP found evidence that EFA had an impact on Attainment 8, with a positive effect size of 0.13 which was significant at the 5% level. Taken together with Key Conclusion 1, this provides good evidence of the effectiveness of EFA.

Cost

The average cost of Embedding Formative Assessment for one school was around £3,895, or £1.20 per pupil per year when averaged over three years. This cost covers several components, including the cost of the SSAT resource package (£295), attendance to training days (£350), and support from SSAT Lead Practitioners during the two years (£3,250).

In terms of staff time, teaching staff were required to commit around two hours each month.

Table 1: Summary of impact on main outcomes

Outcome/ Group	Effect size (95% Confidence Interval)	Estimated months' progress	EEF security rating	No. of pupils	p-value	EEF cost rating
Attainment 8, all pupils	0.10 (-0.01 to 0.21)	2		25,393	0.09	£ £ £ £ £
Attainment 8, FSM pupils	0.07 (-0.04, 0.19)	1	N/A	7,470	0.22	£ £ £ £ £

Introduction

Intervention

'Formative assessment', often used interchangeably with the term 'Assessment for Learning' (AfL), refers to any assessment activities undertaken by teachers—and by students themselves—that provide feedback, which are then used to adapt teaching methods to meet student needs and improve learning outcomes (Black and Wiliam, 1998). The co-developer of the intervention, Dylan Wiliam, popularised this notion of formative assessment in the 1998 book *Inside the Black Box*, co-written with Paul Black, and he has since published extensively on formative assessment and AfL. This intervention builds on his research and experiences with implementing formative assessment programmes elsewhere. Broadly, the intervention aims to support teachers to successfully embed formative assessment strategies in their teaching practice in order to improve pupil learning outcomes and attainment.

The Embedding Formative Assessment (EFA) programme is a two-year intervention; it was delivered during the 2015/2016 and 2016/2017 academic years. All classroom teachers participated in the intervention and were expected to implement the strategies in lessons to pupils in all year groups across the school. The intervention consists of 18 monthly Teacher Learning Communities (TLCs) workshops (nine each year) and monthly peer observations. Dylan Wiliam and Siobhan Leah designed the intervention and the programme materials, and it was delivered by the Schools, Students and Teachers' network (SSAT). SSAT is an independent membership organisation of schools and academies which provides support and training to help improve outcomes for young people and drive school improvement and innovation.

The main element of EFA is the monthly Teacher Learning Community (TLC) workshops, which were usually arranged during normal CPD time. Each TLC workshop involves a group of teachers feeding back on their use of techniques, sharing new formative assessment ideas to try, and personal action planning for the coming month. The resource pack advises schools to have cross-curricular groups with ideally 10–12 teachers in each, but no fewer than 8 and no higher than 14. Each workshop lasts around 75 minutes and follows a similar pattern:

- introduction including the learning intentions for the session (5 mins);
- a starter activity (5 mins);
- feedback from all teachers on techniques they have attempted since last session (25 mins);
- formative assessment content (20 mins);
- action planning (15 mins); and
- summary (5 mins).

In addition, teachers are asked to pair themselves for monthly peer lesson observations in-between each TLC workshop. The peer observations can be for entire lessons or for 20 minutes at the start, middle, or end of a lesson. Pairs will then need to find 15 minutes to provide feedback to each other after each observation.

The intervention materials are provided to support teachers to deliver and guide themselves through the TLC workshops and conduct peer observations. The electronic resource pack included:

- TLC agendas;
- TLC leader's agendas;
- TLC handouts including role of challenger;
- personal action plans;
- peer lesson observation sheets;
- AfL materials including booklet, presentation slides and films of Dylan Wiliam, interviews with teachers, and videos of teachers implementing the techniques in their classrooms; and
- classroom materials.

TLC workshop agendas and materials covered a variety of topics revolving around five key formative assessment strategies: clarifying, sharing and understanding learning intentions; engineering effective classroom discussions and activities; providing feedback that moves learning forward; activating learners as instructional resources for one another; and activating learners as owners of their own learning. Within each of these strategic concepts, the workshop handouts introduced a number of formative assessment techniques for teachers to try.

The broad aim for the TLC workshops and peer observations is to improve teaching and learning by embedding formative assessment strategies in teaching practices. Teachers were required to attempt to address all five broad formative assessment strategies in their classroom, but the specific techniques that they used within each strategy was up to the individual teacher.

Within each school, a lead teacher was responsible for implementing the programme and appointed the required number of teachers to lead/facilitate each monthly TLC group. The main support mechanism was the resource pack. In addition, the lead teacher attended an initial training day by Dylan William (though his involvement is specific to this trial and not necessarily any future rollout) and received ongoing support from a designated SSAT Lead Practitioner. Most of the Lead Practitioners were currently school-based in a middle manager or senior leadership position, with a track record in delivering EFA in schools. They were also trained and supported by SSAT to ensure a consistent structure to their support. Support from Lead Practitioners involved a face-to-face meeting at the start of the project and at the end of the first year. The SSAT Lead Practitioner was also available to be contacted on phone and email throughout the two-year programme. Additionally, schools had access to an online forum to share resources.

Optimal treatment fidelity was emphasised during the initial training day and in the intervention materials. The resource pack does suggest some possibilities to adapt, mainly the possibility of having same-subject TLC groups and reducing the length for smaller groups to one hour. In addition, the materials emphasise that teachers are free to choose which techniques to implement and experiment with, as long as they attempt to address elements of the five broad formative assessment strategies in their classroom. The materials advise that any whole-school policies on preferred techniques should be deferred until the second year of implementation. After the intervention, SSAT acknowledged that it should have made it more explicit to schools exactly what changes and adaptations were permitted as part of the programme. This will be discussed in more detail in the section on fidelity in the process evaluation chapter. After the intervention, SSAT provided the evaluation team with a further list of minor permitted changes such as choosing between the starter activities, choosing to share learning materials in different ways (for instance in advance of a TLC), making changes to groups in Year 2 due to staff changes and movement to improve group dynamics, adopting minor language changes such as referring to peer observations as 'peer support', and using electronic formats of materials and handouts.

Visits to the ten case study schools found a high level of variation in how schools implemented the intervention. The process evaluation explored fidelity to the intervention design using an end-of-project survey of lead teachers to provide a better understanding of this variation and reasons for adaptations. Overall, the qualitative findings demonstrate that while schools generally achieved the broader aim of facilitating dialogue and reflection, sharing of practices, and trialling of formative assessment techniques through the use of monthly workshops, implementation of the programme varied significantly. Most case study schools had made adaptations to the programme, some of which were substantial. In particular, variation was found in relation to the format/structure of TLCs and the use and frequency of peer observations. It was also identified that schools often adapted the programme significantly ahead of the second year of implementation. This element of the evaluation is covered in more detail in the section on fidelity.

Background evidence

There is pre-existing evidence that feedback is effective in improving students' learning, as summarised in the EEF toolkit (EEF, 2018b). The existing evidence is based on a number of relatively small studies with committed teachers, supported by the close involvement of a team of researchers and recognised experts in the field. For example, Wiliam et al. (2004) find an effect size of 0.32. This trial differs from the previous studies in that it includes a much larger group of schools and delivery that is self-administered by schools with less intensive day to day engagement by experts.

To be effective and to be formative, pupils need to change what they would have done as a result of the feedback, spend their efforts differently and often in very specific areas (Black and Wiliam, 1998). The evidence, however, also suggests that teachers find it hard to implement feedback consistently and in ways that respond to students' individual learning barriers and needs, and some studies suggest it can have little or negative impact if not implemented effectively (EEF, 2018b). The findings presented here extend existing evidence in the EEF toolkit by providing evidence of an impact on Attainment 8 that is statistically significant at the 10% level, with an effect size equivalent to two months additional progress in GCSEs (EEF, 2016).

Evaluation objectives

The trial protocol was published in February 2016 (EFA, 2016b).

The trial was designed to identify whether use by schools of 'Embedding Formative Assessment' will improve children's performance in academic tests at age 16.

The primary research question was 'How effective is the Embedding Formative Assessment programme compared to usual practice in terms of improving overall GCSE examination performance?'

The secondary research question was 'How effective is Embedding Formative Assessment compared to usual practice in terms of improving examination performance in Maths and English GCSEs, i.e. subjects which are of high importance in terms of progression to employment?'

Ethical review

The trial was subject to ethical review by the ethics committees at NIESR. SSAT sent schools an information sheet and opt-out form (Appendix C and D) so that schools or individual pupils could opt out. Consent was sought from parents or guardians on behalf of subjects, including the use of a control group. Schools then provided SSAT with a list of pupils after removing those who opted out.

Project team

The development team in SSAT included Fie Raison, Corinne Settle and Anne-Marie Duguid. The process evaluation team included Johnny Runge and Nathan Hudson-Sharp and was led by Heather Rolfe. The impact evaluation team included Francesca Foliano and Matthew Bursnall and was led by Stefan Speckesser and initially by Jake Anders (now UCL Institute of Education). Assistance from Zoe Davison in the Department for Education (DfE) National Pupil Database (NPD) team and Elena Rosa Brown and Guillermo Rodriguez-Guzman from the EEF is gratefully acknowledged.

Trial registration

The trial is registered as ISRCTN ISRCTN10973392 at <https://www.isrctn.com/ISRCTN10973392>.

Methods

Trial design

Randomisation took place at the school level. EFA is a whole-school intervention so within-school randomisation was not appropriate. Schools participating in the trial were randomly assigned to one of two trial arms, either:

- the intervention group, which received the EFA pack, one day's training from Dylan William (the programme developer) at the launch event, and ongoing support by SSAT; or
- a control group, which received a one-off payment of £300 at the start of the trial (September 2015).

The control group was a 'business as usual' control in that there were no restrictions placed on how control schools took forward formative assessment techniques as part of their usual teaching and learning activities. Some treatment and control schools may have accessed the pack prior to the intervention but SSAT prevented the control group from buying the pack if they were on the trial.

Since registration of the protocol there have been four changes to the trial design:

- Originally, the trial design specified that the impact evaluation would be based on GCSE attainment at the end of 2016/2017, comparing the intervention group—after exposure to the treatment during 2015/2016 and 2016/2017—with the control group. This was to be supplemented with interim analysis of pupils that were in Year 11 when the trial began, in terms of their impact GCSE results in 2015/2016, after one year of exposure to EFA. The design was revised in the published Statistical Analysis Plan stage (see Appendix I) to exclude this interim analysis after one year's exposure because the programme was expected to have less impact on attainment for pupil's mid-way through their GCSE programme. See, for example, evidence from the process evaluation suggesting that older pupils were less receptive to the techniques than younger pupils.
- The original protocol was based on the previous GCSE system that used grades A* to F. The analysis presented here uses the new system where GCSE results are reported on a new scale between 1 and 9 points.
- The protocol specified that a subgroup analysis would be conducted for schools in which there was high fidelity to the planned intervention defined as the SSAT-employed Lead Practitioner for a school at no point expressing concern about implementation of the project in their monitoring reports to SSAT. In discussion with the EEF and SSAT during the production of the Statistical Analysis Plan, it was decided that this data was related to potential for trial attrition rather than fidelity to the treatment and this would be replaced by subgroup analysis testing for an impact in the presence of full compliance using data from the SSAT Lead Practitioner implementation survey (see Appendix I). However, the evaluation team decided against using this measure of compliance as well as other self-reported measures of compliances such as the evaluation team's end-of-project survey. The main reason for this was the strong finding from the process evaluation that there existed substantial variation in how people involved in the intervention interpreted high and low fidelity (see more detailed description in the section on fidelity in the process evaluation chapter). Instead, a compliance analysis using a binary, categorical variable was used and an instrumental variables (IV) approach (see amendments to SAP in Analysis sub-section below).
- The Evaluation Protocol included additional indicator variables to be included in the model—including ethnicity, gender, school type, and whether the school has a 6th form. These were removed in line with EEF guidance stating that 'unless there are clear reasons otherwise, evaluations should only use the pre-test scores, the group status and design characteristics as covariates'.

- The process evaluation highlighted that some schools had previously been involved in the Teacher Effectiveness Enhancement Programme (TEEP) intervention which was built on the same collaborative and experimental approach and developed by SSAT. While other interventions and past approaches to formative assessment may also have influenced the delivery of EFA, the process evaluation found that previous involvement in TEEP had the potential to substantially change how lead teachers delivered and presented the EFA intervention in schools (see more detail in fidelity section in process evaluation chapter). As such, the Statistical Analysis plan added an additional subgroup analysis of those schools not previously exposed to TEEP.

Participant selection

The trial included secondary schools drawn from across England and focused on pupils in Year 10 when the intervention began in 2015/2016. In order to be considered, schools had to agree to the responsibilities outlined in the Memorandum of Understanding with Schools (Appendix E). These included:

- providing student data so the evaluation team could match to extracts from the National Pupil Database (NPD);
- committing to provide Continuing Professional Development (CPD) time of approximately 75 minutes per month and offer opportunities for peer observation for all teachers to participate in the Teaching and Learning Communities;
- appointing at least one lead teacher to be the main contact for the project and attend the launch day; and
- cooperating with the project and evaluation teams during the trial as specified in more detail in the Memorandum of Understanding with Schools (Appendix E).

The trial was open to all secondary schools in England. SSAT made all secondary schools in their network of members and non-members aware of the opportunity. Interest was expressed by 250 schools and 140 were chosen using a selection process that included an interview. The interview ensured that selected schools were committed to the four requirements in the above bullets and used specific questions which were designed to show that the schools were committed and were not engaging in too many other new 'initiatives' at the same time as the trial.

SSAT and NIESR provided schools with a joint consent letter and opt-out form for pupils. Consent was sought on an opt-out basis from parents or guardians on behalf of subjects, including the use of a control group, prior to randomisation. The consent letter is provided in Appendix C.

Outcome measures

Both primary and secondary outcomes are GCSE scores provided by the DfE through the National Pupil Database and as such are externally validated and widely recognised measures.

The primary outcome was student's GCSE Attainment 8 score (DfE, 2017). This uses the new GCSE numerical grades introduced in 2016/2017 which range from 0 to 9 using the NPD variable KS4_ATT8. The two secondary outcome measures were student's numerical grades for Maths and English using NPD variables KS4_APMAT_PTQ_EE and KS4_APENG_PTQ_EE, which again range from 0 to 9. Prior attainment was controlled for using Key Stage 2 results using the NPD variable KS4_VAP2TAAPS_PTQ_EE. GCSE invigilation is blind and independent and because they are high stakes tests the pupils will be equally motivated to perform well in each arm.

They are also the preferred EEF measures to be used where possible to ensure comparability across its trials.

The cohort involved in the evaluation studied the new Maths and English GCSEs, taught from September 2015, and the old GCSEs in other subjects. Any new initiative requires adaptations to teaching practice so the requirements of the new GCSEs may have meant that teacher's capacity to focus on EFA was lower during the trial period than it might have been at other periods when the GCSE examination landscape was more stable. As such, the introduction of the new GCSE might have led to a conservative estimate of impact on Attainment 8 and may have contributed to no significant effect being detected for impact on Maths and English scores.

Although not part of the study design, it is also worth mentioning that the impact of EFA could vary by subject. Subjects where teaching and learning involves reflection and discussion of the relative merits of different arguments (such as English and the humanities) might have more to gain from the introduction of formative assessment than subjects where the answer is either correct or incorrect, with less scope for nuanced feedback and reflection (such as mathematics and science). In contrast, formative assessment may be better suited to Maths and science due to the prevalence of student misconceptions. Another argument commonly advanced is that the former type of subject may already have formative assessment embedded to a greater extent than the latter and hence there is less scope for the trial to demonstrate the benefit of the approaches as this would already be captured in the scores for both the treatment and control group.

Sample size

The sample size was chosen in relation to a Standardised Mean Difference of 0.20 standard deviations; this equates to an improvement of approximately one third of a GCSE grade, considered by the SSAT and the EEF advisory panels to be an acceptable level of improvement from a policy perspective to roll out the intervention more widely. This Minimum Detectable Effect Size (MDES) was based on an expectation that 120 schools would be allocated randomly: 60 into the treatment group and 60 into the control group, with an expected average of 100 students in Year 10 at each participating school at the start of trial. The total expected sample size was therefore 12,000 students (100 students per cluster for 0.05 significance level, 0.8 power, and 0.20 intra-cluster correlation; EEF, 2015). The trial eventually recruited 140 schools to account for possible attrition.

Randomisation

Schools were identified as belonging to blocks based on the proportion of students in each school to achieve five A*–C grades in the 2014 GCSE examinations (low, medium, high—where these thresholds are chosen to achieve equal sized groups), and the proportion of students in each school to be eligible for Free School Meals (FSM, low, medium, high) using DfE sources; thresholds were chosen to achieve equal sized groups. Blocking was undertaken to minimise bias at baseline by factors of particular relevance to the study. FSM was chosen as a factor because EEF guidance is for the subgroup analysis to include FSM as a minimum. GCSE score was used because of the potential for differential impact of EFA by ability. Because of correlation between FSM and GCSE performance, a block with fewer than six schools would have been combined with the block with the same level of students achieving five A*–C at GCSE, but a higher proportion of FSM students (unless it is the high FSM block, in which case it would be combined with the medium block instead). However, this was not implemented in practice as all blocks were sufficiently populated.

Within the nine blocks combining the three dimensions of GCSE performance and FSM, schools were randomly allocated to treatment and control groups (half each). This was achieved using a random number generator.

Each school was assigned a randomly generated number between 0 and 1 using the Stata command 'runiform' with seed 2387427. The randomisation was automated by Stata and in this sense was blind. Schools were sorted by blocking variable and, within each block, by the random number. The first

school was randomised to treatment or control; each subsequent school was assigned to the opposite outcome of the previous school.

Randomisation was implemented ahead of the two-year development programme; no baseline data was collected other than GCSE results from the pre-intervention period, eligibility for free school meals and KS2 scores, from an NPD Tier 2 request. For a record of randomisation execution see Appendix J.

Analysis

This analysis follows closely the Statistical Analysis Plan provided in Appendix I. There are four deviations:

- Subgroup analysis for low and high GCSE and eligibility for FSM was supplemented with significance testing of the interaction between the subgroup indicator and the treatment variable. This is in line with the EEF's analysis guidance and does not affect the subgroup models themselves.
- Exploratory analysis of group means has been omitted because we are interested in the intercept for the treatment but not the separate intercepts for the schools.
- The planned compliance analysis based on the survey of SSAT appointed Lead Practitioners has been omitted because the process evaluation identified that people involved in the intervention, including Lead Practitioners, often had very different interpretations of what constituted high and low compliance. This led the evaluation team to conclude that any compliance measure based on self-reporting (including one based on the end-of-project survey) were unlikely to give much insight (see more detail in the section on fidelity in the process evaluation chapter). As such, the previously proposed compliance analysis is replaced by analysis in which compliance is defined in the most narrow sense as treatment schools that remained engaged until the end of the intervention (58 schools). The 12 schools that did not comply dropped out before the end of the programme.
- In the compliance analysis, rather than using the CACE approach originally proposed we have used an IV approach following Dunn (2005) to account for potential endogeneity between impact and compliance, in line with the latest EEF guidance (EEF, 2018).

The estimated impact is based on the difference in KS4 scores between treated and control schools for all schools where data is available, regardless of drop-out, but only those schools and pupils who consented to be included. This was in order to estimate the 'intention to treat' (ITT) effect. Analysis was conducted in Stata using a mixed model with school as a random component to account for school level clustering. Two models were fitted, in line with the latest EEF guidance:

- 'Precise model': including prior attainment, the allocation of dummy and indicator variables specifying membership of the randomisation blocks (all fixed effects) and schools as a random effect.
- 'Simplest model': including prior attainment and allocation dummy as fixed covariates, and school as a random effect.

The model of primary interest is the precise model because this aligns with the randomisation design. Results based on the simplest model are provided so that readers can undertake comparative and meta-analysis with other EEF evaluations.

KS4 scores were standardised by centring around the (treatment) group means but not around individual school means because we are interested in the intercept for the treatment but not the separate intercepts for the schools.

Impact was estimated by fitting the model in equation (1) and testing whether the coefficient of the treatment dummy (Z) was significantly different from zero.

Impact Model

$$y = \beta X + Z\mu + \varepsilon \quad (1)$$

Where

y = vector of centred outcome scores (KS4)

X = covariate matrix (KS2 scores in 'simplest' model, additional dummy variables for stratification groups in the 'precise' model)

Z = design matrix identifying which school (or cluster) an individual attended.

μ = vector of school random effects

β = fixed effect parameters

ε = residual error

with the covariance structure given by Σ , where:

$$\Sigma = (\sigma_a^2 + \sigma_e^2) \begin{bmatrix} I & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & I \end{bmatrix}$$

Where σ_a^2 is a measure of school level variation; σ_e^2 is a measure of student level variation and I is given by:

$$I = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{bmatrix}$$

and ρ is the intra-school correlation coefficient:

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

The effect sizes are calculated using model estimates of the difference between groups divided by the pooled standard deviation (Hedges g) as shown in equation 2

Effect Size calculation:

$$\frac{\widehat{\beta}_\delta}{\sqrt{\widehat{\sigma}_e^2 + \widehat{\sigma}_a^2}} \quad (2)$$

Where:

$\widehat{\beta}_\delta$ is the model estimate of the difference between groups,

$\widehat{\sigma}_e^2$ is the residual error variance and

$\widehat{\sigma}_a^2$ is the total between school variance.

Additional Exploratory Analysis

The process evaluation highlighted that some schools had previously been exposed to the related Teacher Effectiveness Enhancement Programme (TEEP) intervention and there were differences in how EFA was received between these and other schools. Additional exploratory analysis was therefore undertaken for schools which had not previously been exposed to the TEEP intervention. SSAT provided a list of schools that had been involved in TEEP in the years 2011 to 2016. This comprised 12 in the treatment group and four in the control group. This intervention built on the same collaborative and experimental approach as EFA though the strategies were less similar as Assessment for Learning (AfL) only accounted for one strand of the TEEP cycle. While many practices and interventions may influence the delivery and impact of EFA, the process evaluation found that previous involvement in TEEP strongly influenced the delivery of EFA as lead teachers tended to implement the EFA programme according to their previous experiences and practices with TEEP, and often presented the programme as a continuation of TEEP. This impact was picked up organically during evaluation visits, and only afterwards were SSAT asked to supply information about previous TEEP involvement. The impact was more apparent than the impact of any other interventions or existing approaches.

The process evaluation team established a number of hypotheses for the potential impact of being part of TEEP. This included the hypothesis of a more positive (or less negative) impact due to TEEP schools being more open to and experienced with collaborative approaches, or a more negative (or less positive) impact due to already having experienced the advantages of implementing a collaborative approach (see process evaluation for further details). As such, any change resulting from EFA in the treatment group and the estimated impact of business as usual from the control groups may be biased. To address this potential bias, a subgroup analysis of schools that had not previously been involved in TEEP was undertaken.

Subgroup analysis

To test the appropriateness of undertaking further subgroup analyses, the significance of the interactions between the subgroup indicators and the treatment indicator was tested. The subgroup indicators were created using 3 NPD variables:

- students who have ever received free school meals [NPD variable EVERFSM_6];
- students with low attainment scores in Key Stage 2 tests (bottom third as used for the randomisation) [NPD variable KS4_VAP2TAAPS]; and
- students with high attainment scores in Key Stage 2 tests (top third as used for the randomisation) [NPD variable KS4_VAP2TAAPS].

Missing data analysis

In line with the published Statistical Analysis Plan (SAP; EEF, 2018), Multiple Imputation (MI) was undertaken for all Maths and English impact model specifications because more than 10% of data was missing in all cases. It was not undertaken for Attainment 8 because none of the specifications were missing more than 5% of data (overall) or 10% for individual schools or variables. See Appendix G for further details.

Implementation and process evaluation

The overarching purpose of the process evaluation was to show how the EFA intervention was implemented by treatment schools, whether this differed from the intended treatment model, and the factors that informed this. In addition, the process evaluation of EFA monitored the activity of the control group to establish what was done in the absence of the intervention. The process evaluation also aimed to bring greater clarity to the quantitative research findings and to understand the reasons behind the quantitative findings to be estimated in the impact evaluation. It also explored the

perceived impact of the intervention in the eyes of implementers and participants, and to gather their views on how the intervention might be improved, to inform its future rollout.

The following research methods were used:

- visits to ten treatment schools (referred to here as ‘case study’ schools) conducted from May 2016 to September 2016;
 - all visits included interviews with lead teachers, focus groups with TLC leads and other teachers, observations of TLC sessions, and in some cases an interview with the headteacher;
- end-of-project survey of lead teachers in treatment schools, administered from June to July 2017;
- survey of control schools, administered from June to July 2017; the survey was sent to the lead contact for the original application, or alternatively to the headteacher; and
- attendance at the initial training day in September 2015 and celebration event in September 2017, as well as reviewing training content and materials.

All evaluation activities were carried out by NIESR, with support from the delivery team at SSAT. The delivery team provided the raw data collected for their end-of-year reports (Year 1 and Year 2) which were analysed and triangulated with fieldwork findings.

The end-of-project survey was completed by 40 schools, equivalent to 57% of all treatment schools (N = 70), or 69% of schools that finished the programme (N = 58). The control group survey was completed by 39 schools, equivalent to 57% of control schools (N = 70).

Steps were taken to ensure that the ten case study schools included a variety of delivery contexts. That is, treatment schools were selected that differed by Ofsted rating, proportion of pupils receiving free school meals, and geographical location—including whether located in an urban or rural setting. However, they are a relatively small proportion of the treatment group as a whole. The findings therefore may not necessarily reflect the views of the wider population of treatment and control schools. For instance, the fieldwork visits revealed that four case study schools had previously participated in TEEP, which was found to affect the delivery and implementation of EFA. This was a higher proportion than in the overall treatment group. Nevertheless, we believe the qualitative data collected through the visits provide useful insights into the range and diversity of views, and the experience of participants, in the EFA intervention. The findings of the process evaluation should be considered with these strengths and limitations in mind.

Costs

The EEF cost guidance was used as the method to calculate cost per pupil per year.¹ Cost information was collected primarily from the delivery team. They provided information about the cost of the resource pack, and the costs associated with arranging the training day and celebration event, as well as for providing support from Lead Practitioners including their expenses for visiting schools. This was supplemented by specific cost-related questions during fieldwork visits to the ten case study schools. In particular, Lead Teachers, headteachers, and teaching staff were asked about additional costs and time commitments associated with the intervention.

1

https://v1.educationendowmentfoundation.org.uk/uploads/pdf/EEF_guidance_to_evaluators_on_cost_evaluation.pdf

Timeline

Table 2: Timeline

Date	Activity
Summer 2012	Pre-test (KS2)
Spring 2015	Recruitment of participating schools
July 2015	Schools randomised to treatment or control group
July 2015	Names of participating schools provided to NIESR by SSAT, including school URN
9th July 2015	Schools Informed of their allocation
September 2015	Start of school year and beginning of programme delivery. Majority of control schools paid £300.
December 2015	Tier 2 NPD extract provided to NIESR by DfE. This included GCSE results for the 14/15 cohort (pre-intervention); KS2 scores and eligibility for free school meals. Used to estimate sample size
January 2016	Analysis to confirm no systematic bias in the randomisation
May–September 2016	Process evaluation case study visits
September 2015–July 2017	SSAT monitoring data of which schools dropped out
June–July 2017	Survey of treatment and control schools
July 2017	End of 2 year programme delivery
July 2017	Post-test (KS4)
Expected November 2017	Tier 2 NPD National Pupil Database administrative data provided to NIESR by DfE, including KS4 achievement and student and school characteristics (for groups taking GCSE's in 2015/16 and 2016/17).
Actual, January 2018	
March 2018	Draft evaluation report provided to EEF by NIESR

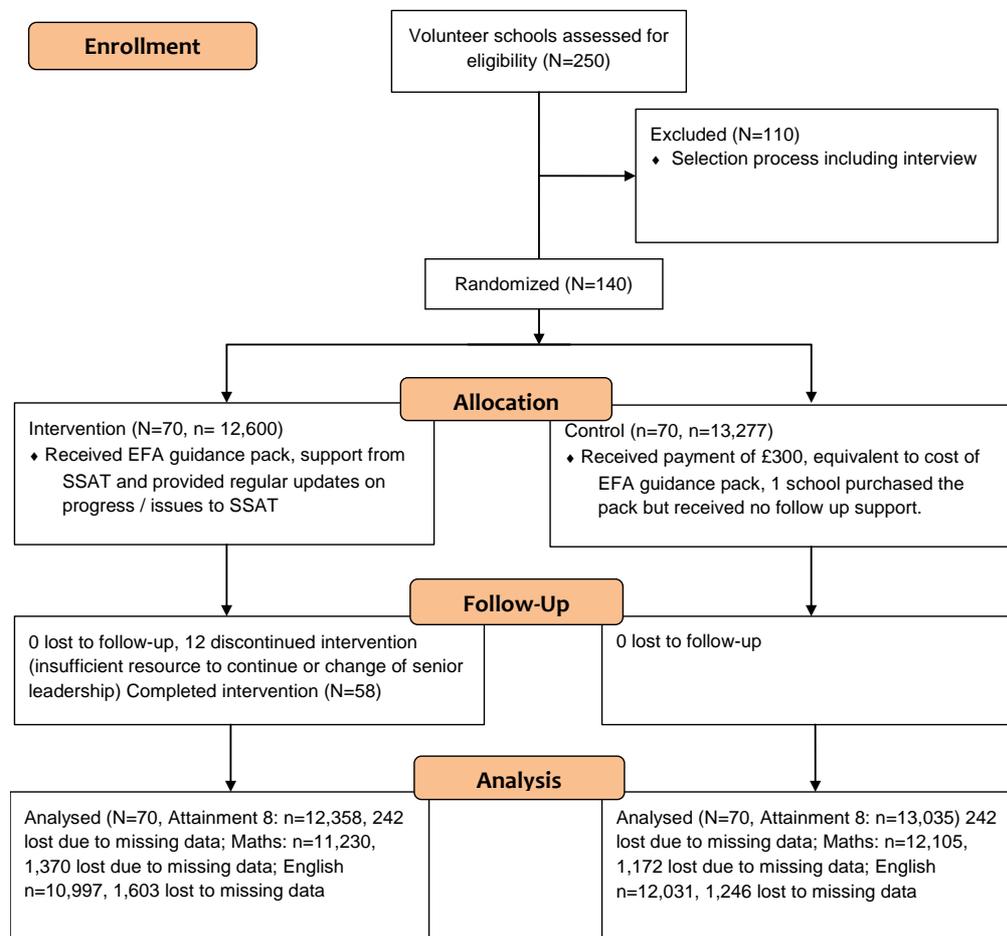
Impact evaluation

Participant flow including losses and exclusions

SSAT made all secondary schools in their network of members and non-member aware of the opportunity. Interest was expressed by 250 schools and 140 were chosen using a selection process that included an interview. The interview ensured that selected schools were committed to the four requirements (see page 10) and used specific questions which were designed to show that the schools were committed, were not engaging in too many other new 'initiatives' at the same time as the trial.

The participation flow diagram is provided in Figure 1 below and the Minimum Detectable Effect Size at each stage of the evaluation is provided in Table 3. School level attrition is relatively low with 12 schools dropping out of the treatment group due to having insufficient resource to continue or a change in senior leadership. Pupil level attrition following randomisation is 0.2% due to missing data only.

Figure 1: Participant flow diagram



For a breakdown of missing values by KS2 and KS4 see Appendix G

Table 3: Minimum detectable effect size at different stages

Stage	N [schools/pupils] (n=intervention; n=control)	Correlation between pre-test (+other covariates) & post-test	ICC	Blocking/ stratification or pair matching	Power	Alpha	Minimum detectable effect size (MDES)
Protocol, EEF (2016b)	60, 60 (6,000; 6,000)	*	0.20	School blocking, 9 blocks	80%	0.05	0.20
Randomisation (precise model)	70, 70 12,600; 13,277	n/a	0.19* *	School blocking, 9 blocks	80%	0.05	0.18
Randomisation (simplest model)	70, 70 12,600; 13,277	n/a	0.15&	School blocking, 9 blocks	80%	0.05	0.17
Analysis (precise model)	70, 70 (12,358; 13,035)	n/a	0.19* *	School blocking, 9 blocks	80%	0.05	0.18
Analysis (simplest model)	70, 70 (12,358; 13,035)	n/a	0.15&	School blocking, 9 blocks	80%	0.05	0.17

*Assumes a within-school R-squared of background characteristics to outcome measure of 0.44 and a between-schools R-squared of background characteristics to outcome measure of 0.32.

**Based on the precise model as reported in Appendix H, with associated total standard deviation of 0.88.

& Based on the simplest model as reported in Appendix H, with associated total standard deviation of 0.93.

Pupil and school characteristics

Table 4 provides a comparison between treatment and control at baseline and suggests the trial is well balanced in terms of inspection grade, broad school type, cohort size, and eligibility for FSM. Most importantly it is well balanced in terms of the KS2 score pre-test. In terms of effect sizes, the difference between treatment and control groups for FSM eligibility and KS2 scores is -0.016 and 0.07 respectively. Pupil level attrition following randomisation is very low, 0.2% for the primary outcome analysis due to missing data.

Table 4: Baseline comparison

Variable	Intervention group		Control group	
	N (missing)	Percentage	N (missing)	Percentage
School-level (categorical)				
Ofsted rating of Good or Outstanding	70 (0)	77%	70 (0)	76%
Academy	70 (0)	71%	70 (0)	69%
School-level (continuous)	N (missing)	Median	N (missing)	Median
Number of Y11 pupils	70	175.5	70	179
Pupil-level (categorical)	n (missing)	Percentage	n/N (missing)	Percentage
Eligible for FSM	12,600 (0)	29%	13,277 (0)	30%
Pupil-level (continuous)	n (missing)	Median	n (missing)	Median
Pre-test score	12,358 (242)	29.05	13,035 (242)	28.91
Pupil-level (continuous)	n (missing)	Mean (SD)	n (missing)	Mean (SD)
Pre-test score	12,358 (242)	27.0 (2.97)	13,035 (242)	26.81 (2.31)

Outcomes and analysis

Table 5 summarises the ITT analysis of the primary outcome measure and Table 6 summarises the associated subgroup and compliance analysis. Table 7 covers the secondary outcomes (maths and English GCSE) and associated subgroup and compliance analysis.

All analysis followed the SAP specification (EEF, 2018) apart from the four deviations explained in the methods section above (page 12). No additional exploratory analysis was undertaken. All effect sizes take clustering into account as specified in the model section above and inter-cluster correlations and total variances are reported in Appendix H. In all cases, the analysis of maths and English based on imputed data, summarised in the Methods section above, provided very similar estimates to the main analysis. These are reported in Appendix G.

The ITT analysis using the primary analysis based on the precise model gives a modest effect size of 0.10 on Attainment 8 which is significant at the 10% level. Compliance analysis, where compliance is defined as remaining engaged for the two years of the intervention (58 schools) gave very similar results. This is a relatively low threshold for compliance so it is not surprising that the effect sizes is no higher than the effect size for the Intention to Treat analysis. In addition, exploratory analysis for schools not previously exposed to TEEP found evidence that EFA had an impact on Attainment 8 with an effect size of 0.13 which was significant at the 5% level.

Interrogation of the subgroups showed a slightly higher effect on lower attainers than higher attainers and although the subgroup analysis itself did not find a significant relationship (at the 10% level) the interaction between higher attainers and the treatment group was significant at the 1% level (see penultimate column in Table 6).

The effect size described above of 0.10 on Attainment 8 can be translated into an impact on GCSEs. Multiplying the parameter estimate for the treatment by the pooled standard deviation of the outcome in the control and treatment groups gives an impact of roughly one GCSE grade in one subject.

Neither the full sample analysis nor the subgroup analyses found significant evidence of EFA having an impact on performance in GCSE Maths or English. The cohort involved in the evaluation studied the new Maths and English GCSEs, taught from September 2015, and the old GCSEs in other subjects. Any new initiative requires adaptations to teaching practice so the requirements of the new GCSEs may have meant that teacher's capacity to focus on EFA was lower during the trial period than it might have been at other periods when the GCSE examination landscape was more stable. As such, the introduction of the new GCSE might have led to a conservative estimate of impact on Attainment 8 and may have contributed to no significant effect being detected for impact on Maths and English scores.

Details of analysis code are provided in Appendix F.

Table 5: Primary outcome analysis

Model/ Outcome Group	Raw means				Effect size		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p- value
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
Precise model, Attainment 8, All pupils	12,358 (242)	47.70 (45.32; 50.07)	13,035 (242)	45.66 (43.81; 47.50)	25,393 (12,358; 13,035)	0.10 (-0.01, 0.21)	0.088*

* significant at 10% level

Table 6: Additional analysis for Attainment 8

Model/Outcome Group	Effect size (95% confidence Interval)	No. of pupils	p-value (subgroup/ treatment interaction)	p-value
Simplest Model, Attainment 8, All pupils	0.11 (-0.04, 0.25)	25,393	n/a	0.16
Precise model, Attainment 8, FSM	0.07 (-0.04, 0.19)	7,470	0.23	0.22
Precise model, Attainment 8, Bottom Tertile	0.1 (-0.04, 0.24)	8,346	0.82	0.17
Precise model, Attainment 8, Top Tertile	0.02 (-0.09, 0.13)	8,522	0.01***	0.71
Precise model, Attainment 8, Non-TEEP	0.13 (0, 0.25)	22,709	n/a	0.04**
Precise Model, Attainment 8, compliance analysis	0.10 (-0.017, 0.212)	25,393	n/a	0.096

* significant at 10% level

** significant at 5% level

*** significant at 1% level

Table 7: Secondary outcome analysis

Model/Outcome Group	Effect size (95% confidence Interval)	No. of pupils	p-value
Precise Model, English, All pupils	0.02 (-0.08, 0.11)	22,628	0.72
Precise Model, Maths, All pupils	0.02 (-0.08, 0.12)	22,935	0.74
Precise Model, English, FSM	0 (-0.11, 0.12)	6,489	0.96
Precise Model, Maths, FSM	-0.01 (-0.12, 0.1)	6,564	0.89

Cost

The cost of this whole-school intervention as delivered in the trial is estimated at around £1,300 per year per school over three years. Based on an average of 1,086 pupils per school in the treatment arm, this is equivalent to an estimated average cost of £1.20 per pupil per year. This cost includes several components: the SSAT resource package is estimated at £295; costs associated with running and the lead teacher attending the initial training day and end-of-programme event is estimated at £350; and the cost of ongoing support throughout the two years from a designated Lead Practitioner is estimated at around £3,250. The total financial cost to the school of participating in the programme is estimated at £3,895. These costs are all incurred during the two years of programme implementation. For purposes of comparability to other EEF interventions, the annual cost estimates are given over a three-year period. It is assumed that schools would incur no financial costs in the third year of programme implementation as teachers continue to use the formative assessment strategies in their planning and teaching practices but they are not participating in other programme elements such as any further TLCs or other activities.

In terms of future roll-out of the intervention, the cost will vary depending on a number of factors, including the size and location of the school, training dates, collaboration with other schools to share training events as well as the membership status and prior involvement with the SSAT and so on. It is also possible to purchase different types of the intervention, including without support from a Lead Practitioner (although different versions of the programme have not been evaluated here). Costs associated with the evaluation itself are not included in these estimates.

In terms of staff time, teaching staff are required to participate in 18 Teaching Learning Communities (TLCs) approximately every four weeks (nine TLCs per year). Each lasts around 75–90 minutes and typically takes place during CPD time. Over the course of the two years, this is equal to between 22.5 and 27 hours for attending TLCs. The preparation time for TLC facilitators (of which each school needs around one for every ten teaching staff) is estimated at 20 minutes per session, equivalent to three hours per year. The process evaluation indicated that the preparation time for the Lead Teacher varies substantially according to how many adaptations are made to the intervention, but a rough estimate would be 20 hours per Lead Teacher per year. In addition, each staff member is required to spend at least 35 minutes on a monthly basis on peer observations. This is divided into at least 20 minutes of observing another teacher's lesson and 15 minutes to provide feedback in pairs. Schools in the intervention generally did not use supply cover to accommodate for peer observations, though some schools may choose to do so if they consider this as usual and best practice.

As part of the process evaluation fieldwork, case study schools were asked if they had incurred additional costs related to the programme. Most participants did not have such information, though a few lead teachers reported that additional costs were fairly minimal. Where incurred, the biggest additional cost was photocopying, though several said this had been a 'drop in the ocean' compared to the overall photocopying budget, and that the intervention wasn't substantially different from business as usual. At the same time, other teachers observed that it was a fairly paper-heavy intervention. In one

school, teachers noted that the materials had come out of their personal photocopying budgets rather than the CPD budget, which meant they had not favoured paper-heavy techniques. Some schools argued that the provision of resources could be better adapted to digital formats such as Google Drive and iPads, which would be more cost effective, though some also highlighted that the paper-based approach had been positively received by schools and teachers as they had sometimes become too reliant on electronic resources. SSAT have informed that they are currently updating the delivery format of the pack to make it suitable for both approaches.

Many schools said they had paid for refreshments for TLC sessions, but that this was a normal feature of their CPD sessions and wasn't an additional cost as such.

Implementation and process evaluation

The overarching purpose of the process evaluation of Evaluating Formative Assessment (EFA) was to explore how the intervention was implemented by treatment schools, whether this differed from the intended treatment model, and the factors that informed this. In addition, the process evaluation of EFA also monitored the activity of the control group to establish what was done in the absence of the intervention. This was done through a survey only. The below description of the process evaluation also aims to bring greater clarity to the quantitative research findings and to understand the reasons behind the marginally significant evidence of impact in terms of increasing Attainment 8 GCSE scores. It also explored the perceived impact of the intervention in the eyes of implementers and participants, and to gather their views on how the intervention can be improved to inform its future rollout.

This section is mainly based on findings from the end-of-project survey and visits to ten case study schools which included interviews and focus groups with teachers and senior leaders as well as observations of TLC sessions (see methodology section for more information).

Implementation

Experience of training and preparedness for delivering the intervention

All treatment schools received a one-day launch in September prior to implementing the intervention, attended by the lead teacher and sometimes accompanied by one of the designated TLC leads. The experiences of the launch day were explored in interviews with school leads and through the qualitative survey. Overall, participants were very positive about attending the launch. In particular, most described the trainer as ‘inspirational’, highlighting the powerful and engaging presentation of the programme’s strategies and the underlying research, which contributed to a high buy-in and enthusiasm among participants. Some lead teachers said this had given them the required confidence to sell the project to teachers in their school, and some regretted that more teachers had not been able to attend the training session. In this regard, it should be noted that Dylan William, the co-presenter at the launch and co-developer of the programme, would not normally be part of the launch of EFA, but that schools will be offered a bespoke launch at the start of programme implementation.

Similarly, the survey of lead teachers in treatment schools, administered at the end of the two-year project, explored lead teachers’ experiences of attending the training day. In total, 32 people in the sample had attended the training day which was attended by all schools (46% response rate). The findings demonstrated that lead teachers were largely in agreement that the training day had made them suitably prepared for delivering the programme (Table 8).

Table 8: To what extent did the launch event and training day prepare you for delivering EFA? (N=32)

Response	Number
To a great extent	17
Somewhat	14
Very little	1
Not at all	0

However, both fieldwork interviews and survey responses indicated that the training day could have been improved (Table 9).

Table 9: Do you think the training could have been improved? (N=32)

Response	Number
Yes	17
No	15

Those who responded that improvements could have been made gave their reasons in a subsequent open-ended answer, which was further elaborated during the fieldwork visits to case study schools. Overall, while some lead teachers said the training had provided them with clear guidance into the expectations of programme delivery, lead teachers more commonly felt that the training day had focused too much on the research and theory behind formative assessment, rather than the practicalities of implementing the programme in schools. School leads suggested that the training could have included mock TLCs and examples of how schools had previously modelled the programme. Typical comments were:

'They didn't go into materials that much. I was thinking they would explain how they wanted sessions to be run and stuff. It was more, this is the holistic [Assessment for Learning], this is what we're aiming for, which was very motivational and inspirational, but I personally didn't come away thinking, "right, I'm clear as to how I'm leading this".'

Lead Teacher, School 9

'The training event focused largely on why we're doing this and why it is good, which was brilliant for understanding that, but wasn't necessarily helpful in terms of logistically how they expected us to implement this within schools... There was no real guidance, apart from the TLCs have to happen once a month, that was about it.'

Lead Teacher, School 5

School leads suggested that the training day could have benefited from being more interactive, with greater sharing of ideas of how to implement the project and overcome specific barriers faced in schools. Lead teachers said a more interactive approach would have helped in understanding the practicalities of implementing the intervention including identifying strategies for possible adaptations. Some particularly expressed regret that the project team had not engaged in discussions around possible adaptations. Lead teachers commonly observed that the trainer had been very prescriptive about the structure and implementation of the programme:

'He gave the impression that you have to do it in this way, and he was very prescriptive about the way he wanted to keep it. I guess that is good for a research project, as clean as possible, but as soon as I got back to the school and talked to my bosses, I realised it was not going to work in this format.'

Lead Teacher, School 10

Finally, some lead teachers strongly criticised the timing of the training day, which was held in September, immediately before rolling out the intervention in schools. They emphasised that most CPD preparation for the upcoming school year was usually done by May-June. In the case of the EFA intervention, they recommended that the training day should have been held in April or May to give lead teachers the necessary time to go back and discuss implementation with senior management, brief TLC facilitators, and potentially arrange an in-service training (INSET) day to launch the programme. This could potentially have affected the impact of this trial, but could be addressed in a wider roll-out. It should be noted that the programme delivery team was restricted by the timeline for when the grant for the trial was agreed, and as such were not able to hold a training day before September.

Aside from the training, the delivery team provided support through designated Lead Practitioners. The support was generally described as good. Most lead teachers noted that, aside from an introductory visit to help setting up the programme, the support had been relatively light-touch and mostly on an ad hoc basis if and when schools encountered any challenges for which they needed assistance. All schools emphasised that there had been open lines of communication, and that they had regularly kept in contact via emails and phone conversations, and the schools knew where to go if there had been any problems. A few schools reported specifically that they had asked Lead Practitioners for advice ahead of making significant adaptations to the programme, often reporting that the Lead Practitioner had accepted and often encouraged this. As such, the process evaluation identified a mismatch between the advice given on the initial training day to follow the intervention guidelines strictly and the more relaxed advice given by Lead Practitioners during programme implementation. The issues around varied implementation and adaptations are further discussed in the section on fidelity.

Experiences of delivering the intervention

The overall experience of participating in the intervention was good. School leads and teachers reported relatively high responsiveness and engagement from their teaching teams. Some lead teachers noted that the buy-in was higher than for the average CPD initiative. Both school leads and teachers often emphasised that it was not an onerous programme and that it did not place undue additional pressure on staff. They repeatedly said *'this was not just another initiative'*, but about embedding good practice and enhancing what teachers were already doing. Engagement was also facilitated by the fact that the programme was perceived as flexible and allowed the individual teacher to take charge of their own personal development rather than being required to deliver a 'one-size-fits-all' scheme. These factors increased the extent of buy-in:

'People haven't felt threatened by this one. [Other initiatives] are like the emperor's new clothes, the next thing comes along and you kind of forget. So that has been the main seller for us is that it is nothing new, it is just really good practice and sometimes the old things are still the best things.'

School Lead, School 3

While some teachers described the project as providing a useful toolbox of teaching techniques, others said that the techniques were nothing 'ground-breaking', but helped re-focusing staff attention on their application and in reflecting on their use. This was perceived as useful both for newer teaching staff who could use the programme as inspiration, or more experienced teachers who may not have reviewed and updated their teaching styles and strategies. One teacher noted:

'There hasn't been a week where I've read the information and been "this is ground-breaking..." But it actually prompts you to think about it and gives you the time to reflect on it and reminds you to do it.'

Teacher, School 5

Some, however, had expected the techniques to be more revolutionary and criticised the materials for being out-dated. They were also seen as too 'paper-heavy' rather than adaptable to digital devices. In relation to the first of these observations, one lead teacher noted:

'I'm surprised that they almost republished what already existed and put together in a neat package, because there is nothing new on that disc and it would be nice for schools to actually have tried something new in terms of formative assessment, because ultimately I would imagine every school will have developed those strategies already.'

Lead Teacher, School 7

Another common observation was that the monthly occurrence of the TLC workshops over a two-year period meant that there had been a sustained focus on formative assessment techniques. Unlike other initiatives which come and go, the EFA programme had been high-profile in the school for a long period, with regular recapping of practices, allowing them to become embedded:

'With any learning, the more you can go back to something you can reflect on, the more embedded it can be.'

School Lead, School 10

However, whilst teachers and leads appreciated the opportunity to embed techniques and practices and understood the rationale for repetition within the programme, many leads and teachers stated that the programme was at times overly repetitious, particularly in its second year. In many cases, this caused schools to adapt elements of the programme, which will be explored in more detail in the section on fidelity.

The TLC workshop format was highlighted in most schools as the key element of the EFA programme. Participants explained that the TLC format provided a formalised forum for effective sharing and reflection of practices. Most recognised that the EFA programme had given the school the required time to have a dialogue about teaching and learning:

'It's actually just half an hour when you're all just talking openly about what worked and what didn't that week, and I think you don't normally get time to just sit there and think about that.'

Teacher, School 6

Teachers valued this time as it allowed them to engage in discussions in small and diverse groups, describing the conversations as non-judgemental, non-threatening and relaxed, solely focused on developmental outcomes. In particular, teachers valued the focus of TLCs on interaction between staff members, with teachers often remarking that *'there is nobody talking at you and no death by PowerPoint.'* However, the process evaluation also identified that the exact TLC structure and the activities within the sessions varied quite substantially from school to school, and, importantly, from the descriptions in the intervention specifications. This will be further explored in the section on fidelity.

The cross-curricular element of the TLC workshops was also highlighted as a positive feature by case study schools, and often described as the real success of the project by senior leaders. Teachers reported that they had enjoyed learning about techniques/strategies from different departments and subjects, acknowledging that many could be easily adapted and be applied to their own lessons, and sometimes complemented their own teaching methods very effectively. In many case study schools, the cross-curricular element was also appreciated for more everyday reasons. In particular, many teachers said the mixed-departmental TLCs gave them an opportunity to meet and speak to teachers outside their own departments. They explained that their school was mainly split along subject lines with separate staff rooms, and also that teachers tended to work in their own rooms:

'We don't congregate in the staff room like we used to... Sometimes a problem shared is a problem halved.'

Teacher, School 8

It should be noted that in a few cases teachers in case study schools said they would have liked to have had more time to go back to departments after TLCs to discuss how techniques and strategies could be applied specifically in their departments. This was a common theme in the survey findings, conducted at the end of the second year, in which some schools reported that they had moved to departmental groups during the second year of the programme, often due to requests from teachers and departments. This will be discussed further in the section on fidelity.

Another key element of the EFA programme was peer observation, in which pairs of teachers observed each other's lessons in-between TLC sessions. The case study visits and the survey findings found that peer observation, as a principle, was highly valued. However, all case study schools pointed out that there had been significant issues relating to organising and timetabling, leading schools to substantially reduce the length and frequency of peer observations compared to the intervention specification. The fidelity section will explore these issues in more detail.

Summary

Overall, the above discussion indicates that EFA was a well-received programme among teachers and senior leaders. In particular, schools reported positive experiences with the TLC format, which was found to facilitate valuable dialogue between teachers and encouraged experimentation with formative assessment techniques.

Fidelity

The online survey of lead teachers in treatment schools, administered at the end of the two-year project, explored to what extent schools had followed the programme and what adaptations they had made. The end-of-project survey was completed by 40 schools, equivalent to 57% of all treatment schools (N=70), or 69% of schools that finished the programme (N=58).

Respondents reported that they had followed the EFA programme 'to a great extent' or 'somewhat', compared to what was outlined by SSAT (Table 10).

Table 10: To what extent did you follow the EFA programme as outlined by SSAT? (N=38)

Response	Number
To a great extent	19
Somewhat	18
Very little	1
Not at all	0

But at the same time, a significant proportion of schools reported that adaptations had been made to key elements of the intervention including to TLCs (Table 11), peer observations (Table 12), and materials/resources (Table 13).

Table 11: To what extent did you make changes to TLCs (frequency, duration, structure, activity plan, chronology, group size etc.)? (N=40)

Response	Number
To a great extent	4
Somewhat	18
Very little	11
Not at all	7

Table 12: To what extent did you make changes to peer observations (frequency, way of organising etc.)? (N=39)

Response	Number
To a great extent	4
Somewhat	20
Very little	11
Not at all	4

Table 13: To what extent did you make changes to materials and resources? (N=39)

Response	Number
To a great extent	3
Somewhat	20
Very little	11
Not at all	5

The fieldwork visits identified that most case study schools had made adaptations to elements of the programme, including TLCs, peer observations and materials/resources. These were often substantial. While schools were found to have achieved the broader objectives—facilitating dialogue and reflection, sharing of practices, and trialling of formative assessment techniques through the use of monthly workshops—implementation of the programme varied significantly, often away from the programme specification. Variation was found particularly in relation to the format and structure of TLCs and the use and frequency of peer observations. School leads frequently said that they had contacted their Lead Practitioner ahead of making adaptations, and the Lead Practitioner had typically accepted or even encouraged the adaptations, often explaining that other participating schools had made similar changes. As a result, many lead teachers who had adapted elements of the programme insisted that their school had still kept to the ‘spirit’ and ‘essence’ of EFA by maintaining a focus on dialogue, sharing, and experimenting with formative assessment techniques. This means that caution should be used when interpreting the findings reported in Table 10 as school leads may have had quite different interpretations of what constitutes ‘following the programme’. This is likely to be an issue with any attempt to establish a compliance measure, particularly one based on self-reporting.

Another issue was that the implementation of the programme varied within schools over the course of the two-year programme. School leads reported that adaptations had typically been made gradually throughout the programme, or at certain points (see Table 14). Lead teachers sometimes said they had followed the TLC structure and activities ‘to the letter of the law’ during the first couple of workshops, but then responded to feedback from teachers and TLC facilitators. Lead teachers also explained that they would adapt, or had already adapted, the intervention for the second year due to staff feedback that the materials were repetitive. Sometimes, staff turnover also meant that the programme was significantly adapted in Year 2.

We asked in our end-of-project survey at what stage changes were made to the delivery of the programme. The most common answer was ‘start of Year 2’, though a large proportion of respondents also answered that they had made the changes ‘during Year 1’ or ‘gradually throughout the two-year programme’ (Table 14).

Table 14: At what stage did you make changes to the delivery of the EFA programme? (N=40)

Response	Number
Start of programme	2
During Year 1	8
Start of Year 2	12
During Year 2	4
Gradually throughout	8
Didn't make any changes	3
Other	3

This is an important finding for the subsequent analysis. The fieldwork findings reflect the experiences of schools around the transition between Year 1 and Year 2 since the majority of visits to case study schools (eight out of ten) were conducted at the end of Year 1 and the remaining two visits at the start of Year 2. The survey findings reflect the whole two-year programme.

Below we will outline some of the adaptations that the process evaluation identified in regard to three key elements of the programme:

- Teaching Learning Communities (TLCs);
- use of formative assessment techniques; and
- peer observations.

Teaching Learning Communities (TLCs)

Among the ten case study schools, there were a wide range of different delivery formats for the TLC sessions, both regarding the frequency, structure, and group composition. For purposes of overview, the case study schools can be broadly (but very simplistically) categorised into the following typologies according to their implementation of EFA during the first year of the programme:²

- Completion of nine, or almost nine, monthly sessions in the first year of the programme, with TLC groups mixed by curriculum, experience, and seniority. The sessions typically lasted around 60–75 minutes (compared to the prescribed 75–90 minutes). The TLC agendas had typically been adapted, including removing/changing the starter activity and the role of the challenger, and made the core sessions more varied and interesting by adapting and changing activities and delivery. One school had split each TLC into a 30-minute all-staff introductory session in the morning, followed by an hour-long TLC at the end of the day **(5/10 schools)**.
- Other schools had differentiated the TLC groups by themes or areas of improvement which teachers opted into. Resources had been adapted accordingly **(3/10 schools)**.
- One school differentiated TLC groups by experience. One TLC group consisted of 20 curriculum leads that were tasked to subsequently disseminate to departments. Groups of other teachers were also differentiated by experience, consisting of around seven to ten members. Resources were adapted, including removing starters, the role of the challenger etc. **(1/10 schools)**.

² The case study schools are not necessarily a representative sample of treatment schools, and as such this list is only included for purposes of providing an overview.

- One school had only one TLC group, consisting of one or two members of each department (60–75 minutes) who then disseminated the information the following week to their department (15–30 minutes). Resources were not shared with teachers to reduce workload (**1/10 schools**).

As indicated in the bullet points above, four of the ten case study schools had differentiated the groups by focus area, which could potentially signify a change away from a key feature of the model, which is that teachers were free to choose any of the techniques covered in the programme. Most commonly, schools had chosen a number of focus areas which were deemed important across the school. Teachers then opted into an area in which they wanted to improve (and TLC facilitators typically chose an area in which they felt they had particular strengths). The aim of this approach was to increase buy-in by personalising the content rather than randomly assigning members to specific groups. The school leads recognized that this required substantial time commitment to undertake adaptations to the TLC agendas and handouts. One school lead acknowledged that this had been overambitious, and that teachers fed back that the TLC agendas were not sufficiently adapted to their chosen theme.

Most case study schools had generally managed to complete the nine TLC workshops during Year 1, though some schools reported problems during specific times of year (for example exam term) and some had had to use an INSET day to complete the required sessions. As adaptations had been made in Year 2, the frequency of sessions was sometimes reduced. Some schools reported having shorter TLC sessions than envisaged by the programme specifications. But typically, TLCs lasted around 60–75 minutes compared to the required minimum of 75 minutes. Some school leads who had had shorter sessions said it was not feasible to ask teachers to spend longer because of their other responsibilities. Some leads also felt that the shorter sessions had not compromised the quality of the sessions because they had removed activities that they perceived as unnecessary, repetitious, and sometimes unconstructive.

Most importantly for fidelity, most schools had adapted the TLC agendas. While some teachers said that the TLC agendas provided TLC facilitators with a nice ready-made structure for delivering the TLC sessions, most described the agendas merely as a good starting point to plan sessions. Many school leads emphasised that the main body of the TLC agendas were considered rigid, dry and repetitive, and that they had found it necessary to increase the pace, or ‘jazz it up’, to increase buy-in. This included the ‘core’ agenda points for reflection and the formative assessment content. In addition, lead teachers felt there were too many unnecessary agenda points, and that the focal point should be on discussion, reflection, and sharing. This meant that many schools had removed or adapted specific activities. In particular, many had removed the ‘three-minutes moan’ which was one of the starting activities. This was considered unproductive, as well as unnecessary in starting the session with a negative slant. Typical comments were:

‘The “have a little rant” thing didn’t work for us. Our staff didn’t actually want to rant, they just wanted to get on with whatever it was that they needed to talk about.’

Lead Teacher, School 9

‘We don’t necessarily need ice-breakers or three minutes to get things off your chest every time, because we work with each other on a daily basis.’

Teacher, School 5

Many case study schools had also formally removed the role of the challenger, which they felt represented a step backwards from their school journey of developing reflective and critical practitioners. Instead, many let it happen naturally, with everyone allowed to pitch in at any time.

There seemed to be mixed approaches among schools regarding the involvement of senior leadership team (SLT) members in TLC groups. Some school leads spoke of the importance of SLT involvement, arguing that the participation of senior leaders was crucial to show that this was seen as an important

project in the school, helping to increase buy-in. One lead in a school with no SLT input argued that the lack of participation from a very senior person had restricted the success of the programme:

'Every member of staff needs to realise that everyone stops for it.'

Lead Teacher, School 9

A few schools, however, had taken a different approach. While there was support from SLT to the overall project, senior leaders did not participate in TLC groups. This was said to have contributed to the feeling that this was a developmental project, with focus on teachers taking charge of their own learning, based on dialogue and interaction between teaching staff.

'Nobody has put anybody under any kind of pressure. They haven't tried to tell me what to do.'

Teacher, School 4

All surveyed treatment schools indicated that the TLC workshops had been implemented as a whole-school intervention. In the fieldwork visits, one school was found to be using a cascading approach where only departmental leads participated in a TLC group and then disseminated the findings in their departments. This signifies a substantial change from the programme specifications. Across schools, the qualitative interviews found that both teachers and school leads agreed that successful implementation should involve the whole school. They emphasised the importance of building a common language among teachers and creating a whole-school culture of self-reflection. Most importantly, teachers felt pupils more quickly became familiar and comfortable with techniques when exposed to them across all subjects. This meant that individual teachers spent less time on giving instructions. A typical comment was:

'You couldn't just throw in one lesson, it wouldn't work, you'd need to build [pupils] up and get them prepared for it to work and understand the basis of what they are doing.'

Teacher, School 6

Techniques

As previously discussed, for classroom teachers, having a toolbox of ideas to choose from was one of the main attractions of the programme. Most school leads said they had not been prescriptive in techniques but let teachers choose those they found most interesting and fitted their teaching style and class requirements. This was in line with the recommendations in the resource pack.

'There was never pressure to say you've got to try this. There were always options.'

Teacher, School 2

'We're not saying everybody has to do x, but we're exposing everybody to everything.'

School Lead, School 1

This meant that techniques applied varied substantially across schools, within schools, and across TLC groups. A typical comment was:

'I can't think of any techniques that leap out at me as ones that we've used across the board.'

Teacher, School 5

The intervention resource pack recommended that any potential whole-school policies related to the use of specific techniques should be deferred until the second year of implementation. This seemed to

have been largely kept, though the process evaluation did identify that some case study schools had developed a specific focus within the programme during the first year of implementation, such as the development of a new whole-school marking policy, which had become almost synonymous with EFA.

Peer Observations

As part of the programme, pairs of teachers were required to observe each other's lessons in-between the monthly TLC workshops. This involved observing a lesson and filling out a feedback sheet, which would then be discussed in pairs. According to the intervention materials delivered to schools, the peer observations can be for entire lessons or for 20 minutes at the start, middle, or end of a lesson. Pairs will then need to find 15 minutes to provide feedback to each other after each observation.

All case study schools pointed out that there had been significant practical issues relating to the organisation and timetabling of peer observations. Many indicated that while peer observations had happened, they had been much less frequent than envisaged in the programme outline. In the survey and visits, many school leads remarked that *'peer observations had been the hardest part of the project to do'*. Teachers and school leads explained that teachers struggled to find space in their timetables to complete the observations, and were generally not able to ask for cover during lessons. The scheduling problems were often compounded for part-time teachers or when schools operated two-week timetables. In some instances, teacher turnover also meant that observation pairs were frequently split up.

As a result, case study schools invariably adapted the length, frequency, and structure of peer observations. As a result, school leads reported very varied practices among teachers. Many school leads explained that they had scaled down this aspect of the programme by not routinely expecting teachers to complete monthly peer observations and asking teachers to observe only a small part of a lesson. This might include, for example, a starter or plenary lasting around ten minutes. Some teachers felt this put pressure on them to force formative assessment techniques into that specific part of the lesson rather than it happening organically. In some schools, there was a specific focus on peer observations during 'Open Door' weeks, rather than regular monthly peer observations. In some schools, teachers had recorded lessons through Lesson Box and IRIS, which enabled both peer observations and self-reflection. Finally, some teachers reported meeting in pairs outside TLC sessions to discuss and reflect on a specific lesson, without having observed the lesson itself. Some had preferred this approach, finding the dialogue more useful than the observation itself.

Summary

While schools achieved the broader aim of increasing the use of formative assessment and facilitating dialogue and sharing of practices, the actual implementation of the intervention varied significantly across case study schools. This led to variation in the nature of the intervention, often away from the programme specification. This occurred particularly in relation to the format and structure of TLCs and the use and frequency of peer observations.

Existing practices—was it business as usual?

Overall, both the survey and qualitative findings indicate that EFA, in many cases, did not signify a major change in schools' approaches to formative assessment and feedback. The majority of survey respondents answered that the intervention was 'quite similar' to previous approaches (Table 15). However, this question may have been interpreted as relating to the concept of formative assessment through approaches to student feedback, rather than the Embedding Formative Assessment programme which includes these techniques plus other elements such as the TLC format.

Table 15: To what extent was EFA similar to your school’s approach to formative assessment and feedback prior to starting the project? (N=40)

Response	Number
Very similar	5
Quite similar	26
Quite different	7
Very different	2

With regard to the use of techniques, some schools said that the formative assessment techniques and strategies were not ground-breaking. Some schools argued that they had already had a strong focus on Assessment for Learning (AfL) prior to the intervention, including on formative feedback, classroom questioning, peer assessment, self-assessment and different schemes focused on marking such as DIRT marking (‘Directed Improvement & Reflection Time’). As such, it was commonly noted that schools were *‘not starting from a low baseline’*. At the same time, schools noted the intervention had been more focused, reflective, and sustained than anything they had previously done, and it had ensured a whole-school approach to formative assessment rather than having pockets of expertise and good practice. One school lead said:

‘Formative feedback has always been on our agenda, we were not going in as novices... But rather than being subtle, it is now at the forefront.’

Lead Teacher, School 3

Previous exposure to the TLC format was more mixed among case study schools. For some schools it was a new and novel approach, whilst some schools reported that they had previously used similar collaborative approaches as part of their CPD programme, and often referred to them as Professional Learning Communities (PLCs).

Typically, these had been less focused, less structured, and not sustained as regularly over such a long period of time. Some schools with prior experience of using the TLC format reported that they had known from the outset where to adapt the format to suit their staff and school. Meanwhile, the cross-curricular element was something that the majority of case study schools described as very different from their previous approaches to Teaching and Learning, which usually happened solely within departments and subject areas.

With regard to the use of peer observations, the picture was mixed. Some schools already had an open-door classroom policy in place prior to the intervention, which meant that the peer observation element of EFA had not transformed the school environment. However, some respondents said that peer observations within EFA were on a larger scale and were across departments. Meanwhile, many other schools reported a culture shift to a more open-door policy, though this was not always attributed solely to EFA. In these schools, lesson observations had become a development tool which could be used to share good practice and used as a starting point for a reflective dialogue about teaching and learning. Some teachers compared this favourably to Ofsted lesson observations, which were reported as a less positive experience.

The survey also asked whether schools had used other materials or resources aimed at improving assessment or feedback during the course of the two-year intervention: 16 out of 40 schools answered ‘yes’ to this question, and provided a list of some of the materials and resources they had used including those developed within the school. From those responses, it was clear that schools are often very active in the use of materials and resources in this area. This is not necessarily a problem for evaluation purposes as most of the resources are focused on content while the rationale for the EFA intervention

is to reinforce existing practices and exactly to bridge the gap between content and implementation by providing a set of processes to sustain the use of formative assessment strategies.

The fieldwork visits identified four out of ten case study schools that had previously participated in the TEEP, implemented by the same organisation as EFA.³ Across the treatment population, 12 schools had previously been part of TEEP. Of the 60 schools that completed the programme, 11 had participated in TEEP, which is a smaller proportion than for the ten case study schools. The fieldwork visits clearly found that previous engagement with TEEP strongly influenced the delivery of the EFA intervention and the experience of delivery. While many other interventions and resources have the potential to impact the delivery of EFA, none seemed to have the same transformative impact on how lead teachers implemented EFA as TEEP. Lead teachers tended to implement the EFA programme according to their previous experiences and practices with TEEP. This may, in part, be due to the programme building on some of the same collaborative principles or simply because lead teachers saw the programmes as connected as they came from the same developer. In most of these schools, EFA was known and sold to teachers as TEEP Top Up. Senior leaders had described the intervention to teachers as a continuation of TEEP, albeit with a stronger focus on Assessment for Learning, which only accounted for part of the TEEP cycle.

It should be noted that this impact was picked up organically during evaluation visits, and only afterwards were the delivery team asked to supply information about schools' previous TEEP involvement. The observation helped inform the decision in the impact study to conduct a subgroup analysis based on prior exposure to TEEP. This showed a significant positive impact (5% level) for schools that hadn't been exposed to TEEP, whilst the impact on all schools was only marginally significant (10% level).

The process evaluation findings indicated that involvement in TEEP seemed to have the potential to affect delivery of EFA both positively and negatively. Generally, teachers and leads said that TEEP and EFA were built on the same collaborative and experimental approach, which meant that particularly the TLC format was already ingrained in the school and among teachers. School Leads in schools previously exposed to TEEP were more likely to state that EFA had not had the same transformative impact because TEEP had already changed the schools' working practices. This may partly explain why the impact analysis found a significant, positive impact for the subgroup excluding such schools.

'If we'd not done TEEP, then I can absolutely see that setting up those cross curricular groups would have been quite a revolutionary idea and a different way of doing it, but because we set those up in 2013 when we started TEEP, had a couple of years running them, I think that's why it didn't have the big wow effect that everyone was telling me it should have.'

'Because we are a TEEP school, we're quite used to the interactive strategy. A lot of the strategies that were in the packs, we had already explored as part of our TEEP journey, so for some it felt like we were recapping.'

Lead Teacher, School 9

However, it could also be argued that TEEP schools could benefit from their previous involvement with a similar programme. Some lead teachers noted that their previous experiences with a similar format meant they knew exactly what had worked and not worked, and had therefore been able to make adaptations accordingly. Some teachers also said that the mind shift they had experienced by participating in TEEP meant that they had been more receptive to trialling different formative assessment techniques:

³ <https://www.ssatuk.co.uk/cpd/teaching-and-learning/teep/>

'I think TEEP was necessary to give you a mind shift change because sometimes people are set in their ways... I think I had a mind shift and I've always been creative and I am always looking at new ideas and that opened up so many avenues for me. I do love to do things differently.'

Teacher, School 10

For lead teachers, previous participation in TEEP could make it more complex to plan the intervention. While many schools had invested substantial time commitments on adapting the intervention to fit into school specific requirements, this seemed to be more pertinent among TEEP school leads. In a few cases, the adaptations meant that the programme looked very different in the school. Combined with teachers' limited knowledge of the difference between TEEP and EFA, it was hard for the evaluation team to assess in interviews and observations to what extent these schools adhered to the programme. Given the impact evaluation estimated a significant, positive impact at the 5% level for the subgroups—excluding those schools previously exposed to TEEP, under both model specifications—it may be that the adaptations taken by former or current TEEP schools have removed implementation too far away from the original programme specifications. Alternatively, improvements already made under TEEP may have biased the impact estimated using the full dataset.

Summary

The intervention (though more intensive and collaborative than previous practices) may not be sufficiently different from existing practices in treatment schools for the impact assessment to detect any further positive impact. However, this is hard to assess, in part because the intervention is fundamentally about reinforcing existing practices.

The process evaluation also identified that previous engagement with TEEP strongly influenced the delivery of the EFA intervention and the experience of delivery. The process evaluation findings indicated that involvement in TEEP seemed to have the potential to affect delivery of EFA both positively and negatively.

Outcomes

The process evaluation explored the perceived outcomes of the intervention among those overseeing implementation in schools. It should be noted that the fieldwork visits to case study schools took place before implementation had finished at around the half-way point of the intervention,⁴ while the survey was answered at the end of the project. Moreover, the survey was completed by school leads or headteachers who would not necessarily experience the everyday impact in classrooms, but often rely on teacher accounts, feedback data, and lesson observations.

With these caveats in mind, the survey tentatively identifies improvements in teaching, and a belief among participants that the programme is likely to improve pupil attainment (Table 16 and Table 17).

Table 16: To what extent do you think the EFA programme has improved teaching? (N=40)

Response	Number
To a great extent	11
Somewhat	26
Very little	3
Not at all	0

⁴ Eight visits took place in May–June at the end of Year 1, and two visits took place in September at the start of Year 2.

Table 17: To what extent do you think the EFA programme is likely to improve pupil attainment? (N=39)

Response	Number
To a great extent	6
Somewhat	28
Very little	5
Not at all	0

During the case study visits, most interviewees (both senior leaders and teachers) were hesitant in concluding that the programme would lead to improvements in pupil attainment within the relatively short time span of two years. Some reasoned it would be hard to disentangle from other initiatives in their school which had worked in tandem with EFA, or the general upward trajectory of the school. Many wondered how it would be measured in the context of government reforms of KS4 and exam specifications. Finally, participants explained that the embedding of formative assessment principles and the improvement in pupils' approaches to learning was a long process. In addition, teachers frequently observed that the intervention had had more impact on younger pupils who were considered more receptive and less critical to trying new things. In contrast, older pupils were described as more exam-minded and frequently questioning the purpose of specific teaching activities. A common example was peer assessment where older pupils wanted the perspective of the teacher as a subject expert rather than input from a peer. Teachers also explained that older students had not had these types of teaching techniques embedded from a young age, predicting that the new cohort of students would be more receptive to formative assessment techniques. Therefore, some suggested that it would be more appropriate to evaluate the impact on pupil attainment a number of years after delivery, among pupils who were younger at the time of the intervention.

On the other hand, some participants believed the programme had more immediate impacts. In particular, some teachers and school leads focused on the positive impact on teaching practices and teaching quality, including increases in high quality feedback, more awareness of current student attainment and understanding of topics, the increased use of engaging teaching strategies, and more self-reflection about own teaching practices. It was generally believed that, ultimately, this would result in higher attainment:

"It's good teaching. Good teaching is always going to produce better results... If you are thinking about it more and you're helping them build on their learning, results are going to be better."

Teacher, School 10

School teachers often pointed towards other positive impacts on pupils other than attainment. Specifically, this included possible improvements in non-cognitive outcomes such as behaviour, concentration, confidence, and communication. Some teachers felt that, by encouraging pupils to become reflective learners and by using strategies to activate all learners including those who were under-achieving, some learners had improved their confidence and independence. It was also suggested that the intervention had increased pupils' engagement and enjoyment of lessons. Improvements in behaviour were also noted though sometimes attributed to other developments in the school. Impacts on higher-achieving pupils were also observed, with a teacher in a selective girls' school explaining:

'This is a selective school, so you've got girls who are coming in and who want to learn and who actually would soak up boring lessons, whereas actually if we're using more techniques and being reminded of those techniques, they're still going to get those results, but their journey to get them might be better, more pleasant.'

Teacher, School 3

Some school leads and headteachers often reported changes in school culture around formative assessment and dialogue, which they attributed mainly to the EFA programme. The increased dialogue between teachers was attributed to the TLC model, with leads and headteachers particularly pleased where they were aware of ad hoc conversations about teaching and learning during the school day. With regard to formative assessment, many schools reported that there had been a switch from having knowledge and occasionally using formative assessment techniques to becoming more conscious of formative assessment, developing an in-depth understanding of the area, and taking ownership of developing and trialling techniques. A typical comment was:

'I think the project will have that legacy in the school in the sense of formative assessment will always be there in the background happening, because we've had such a clear focus on it over the last couple years.'

Lead Teacher, School 9

The process evaluation did not identify any perceived unintended consequences or negative impacts.

Summary

Despite enjoying the programme and believing in its merits, teachers often argued that improvements in teaching practices and pupils' approaches to learning would take time to feed into higher pupil attainment. Regarding potential improvements to GCSE results, this was exacerbated by the perceived tendency of younger pupils to be more responsive to the intervention. In addition, school staff identified positive improvements on pupils other than attainment, such as non-cognitive outcomes and changes in the school culture around formative assessment and dialogue.

Control group activity

An online survey was administered to schools allocated to the control group at the end of the 2016/2017 academic year, coinciding with the end of the programme in treatment schools. Its purpose was to gather information on what activity had been undertaken in relation to assessment and feedback. Similar to the treatment group, the survey was completed by just over half of schools allocated to the control group (39 out of 70) and provides only a partial insight into their activities over this period.

Overall, control schools were asked to assess to what extent their current approach to feedback and assessment was similar to the one prior to the start of the programme. This is a useful explorative analysis as the impact model is based on the assumption that control schools are not contaminated by the treatment, and in particular do not implement any similar changes in their approach to formative assessment as the treatment schools. Most control schools said their approach was either 'quite similar' or 'quite different' (Table 18).

Table 18: To what extent is your current approach to feedback and assessment similar to the one you had two years ago? (N=39)

Response	Number
Very similar	2
Quite similar	15
Quite different	20
Very different	2

Control schools were also asked specifically whether they had made any changes in their approach to feedback and assessment during the previous year. A large majority (33 out of 39 schools) responded they had made some changes. Many schools reported they had had an increased emphasis on giving pupils more time to respond meaningfully to feedback, with more focus on formative comments, greater pupil involvement in acting upon feedback, and a focus on monitoring and improving progress and improvement areas. Some approaches were commonly mentioned, particularly the use of DIRT marking, peer-marking, and self-assessment. A few schools emphasised that the new GCSE and A-level specifications had been the main catalyst for change, leading schools to change revision and learning strategies.

Finally, control schools were asked whether they had used any materials/resources, or participated in any interventions, aimed at improving assessment or feedback: 13 out of 39 schools responded they had, with five of those saying they had used Dylan William's Embedding Formative Assessment resources. In addition, data provided by the SSAT shows that 4 out of 70 control schools had participated in TEEP, a smaller proportion than in the treatment schools (12 out of 70 schools). Apart from this, there were mixed responses, with most citing internally developed materials and one-off CPD sessions.

Conclusion

Key conclusions

1. Students in the Embedding Formative Assessment schools made the equivalent of two additional months' progress in their Attainment 8 GCSE score, using the standard EEF conversion from pupil scores to months progress. This result has a very high security rating.
2. The project found no evidence that Embedding Formative Assessment improved English or Maths GCSE attainment specifically.
3. The additional progress made by children in the lowest third for prior attainment was greater than that made by children in the highest third. These results are less robust and have a lower security rating than the overall findings because of the smaller number of pupils.
4. Teachers were positive about the Teacher Learning Communities. They felt that these improved their practice by allowing valuable dialogue between teachers, and encouraged experimentation with formative assessment strategies.
5. The process evaluation indicated it may take more time for improvements in teaching practices and pupil learning strategies to feed fully into pupil attainment. Many teachers thought that younger students were more receptive to the intervention than their older and more exam-minded peers.

Interpretation

The effect size was consistently positive under both model specifications for Attainment 8 and for all subgroups and the compliance analysis using the precise model. Taken together with the fact that the ITT analysis was significant at the 10% level and the subgroup analysis for schools not previously exposed to the Teacher Effectiveness Enhancement Programme (TEEP) was significant at the 5% level, the evaluation provides good evidence that EFA has a modest impact on Attainment 8 GCSE scores.

The process evaluation found that overall the intervention was well-received by schools. Treatment schools generally reported positive experiences with the TLC format. This facilitated valuable dialogue and sharing of good practices in relation to Teaching and Learning broadly, and formative assessment specifically. In addition, the programme brought formative assessment techniques and strategies to the forefront of teachers' minds for an extended period of time, which encouraged teachers to experiment with techniques and helped to embed already existing good practice. It was also reported to feed into an improvement in teaching quality. These positive elements may help explain why the ITT analysis shows a small positive effect on the Attainment 8 GCSE scores that was statistically significant at the 10% level.

However, while schools generally achieved the broader aim of increasing the use of formative assessment and facilitating dialogue and sharing of practices, implementation of the programme varied significantly, often away from the programme specification. This occurred particularly in relation to the format and structure of TLCs, as well as the use and frequency of peer observations. This resulted in variation in the nature of the intervention amongst treatment schools. This lack of consistency of implementation across schools means that caution should be used when interpreting why the programme has had a positive impact on pupil attainment. In particular, caution should be used in attributing the impact to the exact treatment model since the format of TLCs and frequency of peer observations may not have been followed per the programme specifications. Instead, the process evaluation findings indicate that the modest positive impact on Attainment 8 GCSE scores is more likely to have been caused by the general and sustained focus on formative assessment and Teaching and Learning through the use of monthly TLCs with emphasis on dialogue, reflection, and sharing of practices, as well as on experimenting and trialling techniques. However, due to difficulties in measuring compliance, which have been discussed throughout the report, the evaluation cannot make any firm

conclusions about whether this impact is felt across all schools or only among those that have adapted or not adapted the programme.

The TLC mechanism was considered a novel approach in many schools though some schools had been involved in similar programmes which affected implementation and impact. This included, in particular, participation in another intervention called TEEP. Meanwhile, the formative assessment content was usually very similar to existing approaches already being implemented in control and treatment schools. However, this was part of the project design as the intervention is fundamentally about reinforcing and sustaining existing good practices by providing an effective mechanism to implement the formative assessment approach consistently across the whole school.

Finally, despite enjoying the programme and believing in its merits, teachers often argued that improvements in teaching practices and pupils' approaches to learning would take time to feed into higher pupil attainment. This was reinforced by the frequent observation among teachers that younger pupils were more receptive to the intervention than their older and more exam-minded peers. As such, the intervention may be more suited for a long-term evaluation in which the impact on pupil attainment is assessed a number of years after delivery, among pupils who were younger at the time of the intervention.

Limitations

The principal source of bias is previous exposure of schools to similar interventions. This was addressed and adjusted to some extent by SSAT providing data on schools' previous involvement in their TEEP intervention which more than any other intervention impacted how lead teachers delivered EFA (see introductory sections and fidelity section in process evaluation chapter). However, the process evaluation indicates that both treatment and control schools often already used similar strategies and techniques, and may have been involved in interventions other than TEEP with similar aims, and there was perhaps not enough systematic collection of data on these sources of 'contamination' including a usual practice survey at the baseline and examining in more detail any potential compensation rivalry in the control group. Future such studies might benefit from collecting this information systematically.

As discussed earlier, the data collected from treatment schools as part of the process evaluation (either via survey or fieldwork visits) only represent the views and experiences of a subset of the larger treatment population. Whilst visited schools were selected to include a variety of delivery contexts, the qualitative findings are not necessarily representative, but provide an insight into the range and diversity of views and experiences in the treatment population.

Fieldwork visits to case study schools were undertaken at the end of the first year and the beginning of the second year of implementation. The fieldwork findings indicated that many schools had found the TLC agendas too repetitive for the second year, and had made substantial changes to delivery accordingly. While the end-of-project survey captured some of the nature of these adaptations during the second year of the programme, the qualitative fieldwork was not able to assess in full how these adaptations affected the second year of the programme. As such, the reported findings are related to the first year of the intervention where implementation already varied substantially across schools.

Future research and publications

The effect size for the impact on Attainment 8 was consistently positive under all specifications and for all subgroups. Combined with the process evaluation findings that teachers felt the techniques may take time to embed, and that younger learners were often observed to be more receptive to the techniques than their 'exam focussed' peers, this suggests the programme should be monitored further using a longitudinal analysis for all year groups. Specifically, given that this was a whole-school intervention, it would be sensible to analyse the GCSE attainment of subsequent Year 11 cohorts.

References

- Black, P. and Wiliam, D. (1998) 'Inside the Black Box: Raising standards through classroom assessment', London: School of Education, King's College, London.
- Education Endowment Foundation (2015):
https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf [accessed 20 August 2017].
- Education Endowment Foundation (2016) Trial Report Template:
https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0ahUKEwithfmF7qLaAhXCyaQKHUwhAcAQFggUAE&url=https%3A%2F%2Feducationendowmentfoundation.org.uk%2Fpublic%2Ffiles%2FEvaluation%2FWriting_a_Research_Report%2F2016_trial_report_template.docx&sg=AOvVaw012Yi8nRk6WQEeKIUDDTEW [accessed 15 February 2018].
- Education Endowment Foundation (2016b) 'Project Protocol, Embedding Formative Assessment':
https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_EmbeddingFormativeAssessment.pdf [accessed 7 March 2018].
- Education Endowment Foundation (2018) 'Statistical Analysis Plan, Embedding Formative Assessment': https://educationendowmentfoundation.org.uk/public/files/Projects/Round_7_-_Embedding_Formative_Assessment_SAP.pdf [accessed 7 March 2018].
- Education Endowment Foundation (2018b) Teaching and Learning Toolkit | Feedback:
<https://educationendowmentfoundation.org.uk/pdf/generate/?u=https://educationendowmentfoundation.org.uk/pdf/toolkit/?id=131&t=Teaching%20and%20Learning%20Toolkit&e=131&s=> [accessed 5 April 2018].
- Dunn, G., Maracy, M. and Tomenson, B. (2005) 'Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods', *Statistical Methods in Medical Research*, 14 (4), pp. 369–95.
- The Department for Education (DfE) (2017) 'Progress 8 and Attainment 8: Guide for maintained secondary schools, academies and free schools':
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/583857/Progress_8_school_performance_measure_Jan_17.pdf

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

Cost rating	Description
£ £ £ £ £	<i>Very low:</i> less than £80 per pupil per year.
£ £ £ £ £	<i>Low:</i> up to about £200 per pupil per year.
£ £ £ £ £	<i>Moderate:</i> up to about £700 per pupil per year.
£ £ £ £ £	<i>High:</i> up to £1,200 per pupil per year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

Rating	Criteria for rating			Initial score		Adjust		Final score				
	Design	Power	Attrition*									
5 	Well conducted experimental design with appropriate analysis	MDES < 0.2	0-10%	5				5				
4 	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES < 0.3	11-20%			Adjustment for Balance [0]						
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%				Adjustment for threats to internal validity [0]					
2 	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%					Adjustment for threats to internal validity [0]				
1 	Comparison group with poor or no matching (E.g. volunteer versus others)	MDES < 0.6	41-50%						Adjustment for threats to internal validity [0]			
0 	No comparator	MDES > 0.6	>50%							Adjustment for threats to internal validity [0]		
											Adjustment for threats to internal validity [0]	

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = 5 padlocks
- **Reason for adjustment for balance** (if made): There is a small amount of chance imbalance at the baseline on prior attainment and this is appropriately accounted for in the analysis.
- **Reason for adjustment for threats to validity** (if made): There is no adjustment for threats to validity. Although there was some variation in implementation fidelity, these variations were well documented.
- **Final padlock score:** initial score adjusted for balance and internal validity = 5 padlocks

*Attrition should be measured at the pupil level, even for cluster trials and from the point of randomisation to the point of analysis. Further detail provided here:

https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/2016_Guidance_on_peer_review_for_EEF_trials.pdf

Appendix C: Consent Letter

Dear Parent / Carer

Your child's school has applied to take part in an exciting national research project to improve teaching and learning by evaluating the teaching tool Embedding Formative Assessment (EFA). The aim is to improve your son/daughters attainment by gathering together groups of teachers and exploring the principles and activities of formative assessment.

The school has applied to be one of the 120 secondary schools nationwide who will be involved in the project. Anonymous pupil GCSE outcomes in all 120 schools will be analysed at the end of the two year period to evaluate the impact of the tool. A team from National Institute for Economic and Social Research (NIESR) has been appointed to evaluate the project externally.

For the purpose of research, information provided by your child's school (including your child's name, date of birth, gender, free school meal entitlement, and unique pupil number) will be linked with information about your child from the National Pupil Database (held by the Department for Education) and shared with the evaluators, the Department of Education, EEF's data contractor FFT Education and in an anonymised form to the UK Data Archive. Your child's data will be treated with the strictest confidence and stored securely at all times. We will not use your child's name or the name of the school in any report arising from the research. We may ask your child to take part in a survey or focus group; again all data will be treated with strictest confidence.

If you are happy for information about your child to be used in the evaluation of Embedding Formative Assessment you do not need to do anything. Thank you for your help with this evaluation, your support is much appreciated.

If you would rather your child's school did not share information about your child for use in this evaluation please complete the enclosed form and return it to your child's school by [INSERT DATE].

If you would like further information about the Embedding Formative Assessment evaluation please contact efa@ssatuk.co.uk or 020 7802 2332.

Yours faithfully

Appendix D: School Information Sheet

School Information Sheet

Whole School Embedding Formative Assessment project with the SSAT and the Education Endowment Foundation

Thank you for your interest in the Embedding Formative Assessment project. Please see information below that will outline the project in more detail, including what you need to do for your schools to be considered.

The project

This is an opportunity to be involved in a whole school professional development programme based on formative assessment. Dylan William will introduce schools to the concept at a launch event, and then schools will run a series of structured workshops (Teacher Learning Communities) throughout a two year period supported by Lead Practitioners.

Why you should get involved

Formative assessment involves teachers using evidence of pupils' understanding and learning to make decisions, minute-by-minute and day-by-day, about the next steps in teaching and learning. This evidence could also be used when planning lessons or differentiating activities for individual pupils. When assessing formatively, the feedback from learners and given by teachers moves learners forward. Students are empowered to be owners of their own learning and support each other to progress.

Formative assessment is known to be effective and the difficulty lies in getting teachers to adopt the practice successfully. This trial will evaluate the impact of a cost-effective and scalable route to implementing formative assessment in a large number of schools. Join us in being part of a ground breaking new study.

Evaluation

The project will be evaluated by a team from the National Institute for Social and Economic Research. We are looking to recruit 120 secondary schools from across the country, who will be randomly allocated to receive the pack and implementation support, or form the control group. There will be an evaluation exploring how the materials are being used in schools, and GCSE results from implementation and control schools will be used to estimate the impact of the programme on academic attainment at the end of the second year (summer 2017).

If schools are allocated as a treatment group they will need to:

- Identify a lead person for the project and ensure that this lead person (or a colleague) can attend the launch in London on 11 September 2015.
- Allocate a monthly CPD time for all teachers of at least 75 minutes over two years.

- Engage with Lead Practitioner support through regular contact including face to face meetings over the two years.
- Arrange meetings with the evaluators (including potential interviews and focus groups with teachers and students)

Next steps

If you would like to be a part of this exciting study, please fill in the online form to register your interest before 6th May.

We will be shortlisting from the expressions of interest and then contacting schools for phone interviews between 11 – 22 May.

What you will need to do

If you are shortlisted to be one of the participating schools, you will need to commit to providing anonymised student data and allowing CPD time for all teachers to participate in the programme over the next two years. In order to be considered for the programme, schools will also need to:

- Send out the provided consent and opt out form to all parents.
- Give the evaluators the following details:
 - Your school's Unique Centre Number (UCN).
 - Proportion of pupils who achieved 5A*-C GCSEs for last year.
 - Unique Pupil Numbers (UPNs), including name, date of birth, gender, free school meal entitlements and current year group for all students in the school.
- Sign the MoU and return to SSAT.

This will need to be completed by the 25 June. Schools will need to complete all consent forms and sign the MoU to be eligible to participate in the project. Schools will then be randomly allocated into either the treatment or control groups.

Key dates:

Register your interest before	6 May
Phone interviews	11 – 22 May
School receives documentation to send to parents and the MoU	18 – 29 May
Last date for signed MoUs	25 June
Communication about group allocation	Before 15 July

Appendix E: Memorandum of Understanding with schools

School Agreement to participate in the Whole School Embedding Formative Assessment project with the SSAT and the Education Endowment Foundation

Please complete and sign both copies, retaining one and returning the second copy to Anna Ware, SSAT, 5th floor, 142 Central Street, London EC1V 8AR or by email to efa@ssatuk.co.uk by **25 June 2015**.

School Name: _____

School Postcode: _____ Head teacher Name: _____

Aims of the Project

The aim for the project is to test the effectiveness of a whole school approach to Embedding Formative Assessment, designed by Professor Dylan William with SSAT. A tool will be provided with the materials that schools need to develop and deliver Teacher Learning Communities (TLCs). Groups of teachers have scheduled meetings and explore the principles and activities for formative assessment with the objective of improving teaching and learning.

Rationale

Formative assessment is known to be effective, however the difficulty lies in schools being able to adopt the practice successfully and in the long term. The Education Endowment Foundation is funding a research project to assess the impact of a whole school approach to Embedding Formative Assessment on pupil attainment. The project will run for two years to ensure that the approach is embedded within the school and will evaluate the impact of a cost-effective and scalable route to implementing formative assessment in a large number of schools.

The Project

After the phone interview, shortlisted schools were asked to inform all parents about the project and collect any Opt Out Forms returned by parents (see separate correspondence sent on Thursday 28 May 2015). Schools were also asked to provide the following school and pupil level information.

- School's Unique Reference Number (URN)
- The school's LAESTAB code.
- Unique Pupil Numbers (UPNs) for all children.

- Child's name, date of birth, gender and free school meals entitlement.
- Pupil's current year group.
- Proportion of pupils eligible for Free School Meals (FSM).
- Proportion of pupil's who achieved 5A*-C GCSEs last year.

The evaluation team will use this information and access the National Pupil Database to collect GCSEs scores for all relevant pupils in order to assess any impact of the project on attainment.

Across the 2 years of the trial the evaluators may survey and/or interview pupils and teachers involved to assess the impact of the tool on teaching and learning. These will be non-invasive and arranged at a time convenient to the school.

The Evaluation

The evaluation is being conducted by National Institute for Economic and Social Research (NIESR) and a Randomised Control Trial (RCT) approach is being applied.

Schools who agree to take part will be randomly allocated to either:

1) **An intervention group** and will receive an EFA pack, attend a launch event with Professor Dylan Wiliam, and will receive a mix of school based and Skype/telephone implementation support by a Lead Practitioner over a two year period.

2) **A control group** and receive no intervention but will be given a small financial incentive.

Random allocation is essential to the evaluation as it is the best way of understanding what effect Embedding Formative Assessment has on children's attainment. It is important that schools understand and consent to this process.

Use of Data

Pupil data will only be used for the purposes of analysis and will be treated with great care to achieve high levels of security. Pupils' questionnaire responses will be treated with the strictest confidence and stored securely, accessible only to restricted users.

At no time will individual or school-level data be disclosed to any third parties. It will only be used for purposes connected with the process including:

- Checking randomisation has worked through analysing the average characteristics of individuals and schools in the treatment and control groups;
- Calculating the estimated impact of the project by comparing the outcomes of individuals in treatment and control schools.

Responsibilities

SSAT will:

- Shortlist successful schools through application and interview process.
- Provide the evaluators with a list of schools to be allocated to the control and trial groups.
- Let schools know which group they have been allocated to, based on the information provided by the evaluators.
- Be the first point of contact for any queries.

Also:

For the control schools;

- Promptly pay the £300 financial incentive after allocation.
- Will share the evaluation findings and final case studies with the school. For the intervention schools;
 - Deliver the pack to the schools.
 - Allocate a Lead Practitioner.
 - Hold a launch event with Dylan Wiliam on Friday 11th September 2015 for school EFA leads
 - Provide on-going implementation support to the school.
 - Liaise with the Lead Practitioners and ensure quality of support.
 - Collate case studies for and produce publications at the end of each year.

- Programme and project management for the duration of the project
- Deliver an event in the final year with Dylan Wiliam.

The EVALUATION TEAM (NIESR) will:

- Conduct the random allocation.
- Collate School and Pupil Level data provided by schools.
- Obtain National Pupil Database data for participants from the DfE.
- Store all data safely and securely.
- Analyse data from the project in order to produce impact estimates.
- Conduct the process evaluation, including analysis and reporting from this.
- Produce end of project evaluation report.
- Disseminate research findings.

In order to participate in the project, both intervention and control SCHOOLS will have:

- Sent all year 7-10 parents the consent and opt out form.
- Provided the evaluators with schools and pupil level data as outlined above.
- Consented to random allocation and committed to the outcome of the result.

If in the **intervention group** schools will:

- Commit to providing CPD time for teachers to hold 18 monthly TLCs over the two year period for a minimum of 75 minutes per meeting.
- Appoint at least one lead teacher who will be the main contact for the project.
- Have a project lead attend the launch on 11th September with Professor Dylan Wiliam.
- Commit to regular and timely communication with SSAT.
- Engage with Lead Practitioner implementation support through regular contact including face to face meetings over the two years.

- Meet with the evaluators (including potential interviews and focus groups with teachers and pupils)
- Write a case study at the end of each year about their experiences on the project for a publication to be circulated nationally.
- Share permissible data with the evaluators, and liaise with them as required, including participation in evaluation activities.
- Inform SSAT if there is a change of lead teacher for the project or Head teacher at the school.
- Allow extra staff time if necessary and within reason to the project.

If allocated to the **control** group:

- Schools may still be asked to provide monitoring data during the course of the trial.

As a school I commit to remaining a part of the Embedding Formative Assessment project as detailed above for the period of September 2015 – July 2017

Head teacher name: _____

School name: _____

Head teacher Signature: _____ Date: _____

Head teacher Email address: _____

School Contact (if different from head teacher): _____

School Contact email address: _____

School Tel no: _____

Number of Teachers (including SLT) at school as at [Insert date] _____

Number of pupils at school as at [Insert date] _____

Thank you for agreeing to take part in this research. Please return this form to:

Anna Ware, SSAT, 5th floor, 142 Central Street, London EC1V 8AR or efa@ssatuk.co.uk

Appendix F: Details of Analysis Code

The model described in the main body of the report was operationalised by restricted maximum likelihood estimation using the below STATA commands.

```
mixed KS4 KS2 TREAT i.[Tertile of school average KS2 in 2014] i.[Tertile of proportion of cohort eligible for FSM] || SCHOOL: reml - for the precise model and:
```

```
mixed KS4 KS2 TREAT || SCHOOL: reml – for the simplest model
```

Appendix G: Multiple Imputation

The first step of the analysis was to assess whether the missing data is missing at random (MAR) as opposed to either 'Missing Completely at Random' (MCAR) or 'Missing Not at Random' (MNAR). If the former case holds the data related to randomness is unobserved and MI is not feasible. If the latter case holds, the only approach would be to adopt a structural modelling approach which we would not adopt because this would deviate from the principles of transparent reporting as findings would be assumption rather than data driven.

To assess whether missing data was MAR, an indicator variable was created for each variable in the impact model specifying whether the data was missing or not and logistic regression was used to test whether the missing status could be predicted from the variables in the precise model plus school average KS2 (continuous variable), eligibility for FSM, gender and ethnicity. An F test for the relationship between these independent variables and the missing-ness dummy was significant at the 5% level and in line with the SAP the below MI analysis was undertaken. In all English and Maths models, there was greater than 5% missing data (see CONSORT diagram and first table below) so we undertook missing data analysis using Multiple Imputation (MI). In all cases the analysis based on imputed data found very similar conclusions to the primary and secondary analysis without imputation.

Estimates from Multiple Imputation

	Effect size		
	n in model (intervention; control)	Hedges g (95% CI)	p-value
Simplest Model, All pupils, English	25,877 (12,600;13,277)	0.021 (-0.063; 0.104)	0.631
Simplest Model, All pupils, Maths	25,877 (12,600;13,277)	0.034 (-0.127; 0.196)	0.679
Simplest Model, FSM, English	7,706 (3,658;4,048)	0.008 (-0.081; 0.097)	0.856
Simplest Model, FSM, Maths	7,707 (3,658;4,048)	-0.012 (-0.088;0.075)	0.885
Simplest Model, Low attainment KS2, English	8,346 (3,976;4,370)	0.001 (-0.089;0.091)	0.97
Simplest Model, Low attainment KS2, Maths	8,347 (3,976;4,370)	0.014 (-0.069;0.097)	0.761
Precise model, All pupils, English	25,877 (12,600;13,277)	0.015 (-0.051;0.081)	0.66
Precise model, All pupils, Maths	25,877 (12,600;13,277)	0.014 (-0.054;0.081)	0.688
Precise model, FSM, English	7,706 (3,658;4,048)	0.006 (-0.075;0.087)	0.873
Precise model, FSM, Maths	7,707 (3,658;4,048)	-0.005 (-0.079;0.069)	0.908
Precise model, Low attainment KS2, English	8,346 (3,976;4,370)	0.002 (-0.074;0.077)	0.956
Precise model, Low attainment KS2, Maths	8,347 (3,976;4,370)	0.016 (-0.055;0.086)	0.682
Simplest Model, All pupils, English	25,877	0.021	0.631

Embedding Formative Assessment

	(12,600;13,277)	(-0.063; 0.104)	
Simplest Model, All pupils, Maths	25,877 (12,600;13,277)	0.034 (-0.127; 0.196)	0.679
Simplest Model, FSM, English	7,706 (3,658;4,048)	0.008 (-0.081; 0.097)	0.856
Simplest Model, FSM, Maths	7,707 (3,658;4,048)	-0.012 (-0.088;0.075)	0.885

Appendix H: ICC and Total Variance

Primary Outcome

Model	ICC (Total SD*)
Simplest model, Full sample, Attainment 8	0.19 (0.925)
Simplest model, FSM, Attainment 8	0.15 (0.904)
Simplest model, Low Attainment KS2, Attainment 8	0.15 (0.870)
Simplest model, High Attainment KS2, Attainment 8	0.15 (0.652)
Precise Model, Full sample, Attainment 8	0.09 (0.880)
Precise Model, FSM sample, Attainment 8	0.10 (0.880)
Precise Model, Low Attainment KS2, Attainment 8	0.15 (0.831)
Precise Model, High Attainment KS2, Attainment 8	0.10 (0.633)

Secondary Outcomes

Model	ICC (Total SD)
Simplest model, Full sample, English	0.11 (0.738)
Simplest model, Full sample, Maths	0.12 (0.672)
Simplest model, Full sample, English	0.11 (0.759)
Simplest model, Full sample, Maths	0.1 (0.678)

* The square root of residual variance plus between group variance

Appendix I: Statistical Analysis Plan

INTERVENTION	
DEVELOPER	Dylan Wiliam and SSAT
EVALUATOR	National Institute of Economic And Social Research (NIESR)
TRIAL REGISTRATION NUMBER	ISRCTN ISRCTN10973392 details available at https://www.isrctn.com/ISRCTN10973392 .
TRIAL STATISTICIAN	Dr Matthew Bursnall
TRIAL CHIEF INVESTIGATOR	Dr Matthew Bursnall
SAP AUTHOR	Dr Matthew Bursnall & Dr Stefan Speckesser
SAP VERSION	Published
SAP VERSION DATE	31 January 2018
EEF DATE OF APPROVAL	31 January 2018
DEVELOPER DATE OF APPROVAL	31 January 2018

Introduction

There is evidence that feedback is effective in improving students' learning. However, existing evidence, as summarised in the EEF toolkit⁵), suggests that teachers find it hard to implement consistently and in ways that respond to students' individual learning barriers and needs. It is possible that feedback is sometimes effective in helping students to overcome specific learning barriers in the short term, but has little formative impact. A team, or whole school approach, appears to be a key component of successful feedback interventions and was included in this project through a workshop model.

⁵ <https://educationendowmentfoundation.org.uk/resources/teaching-learning-toolkit/feedback/>

The “Embedding Formative Assessment” (EFA) project is a two-year whole school professional development programme on formative assessment, which includes a day’s training and materials to deliver 18 monthly internal workshops (“Teacher Learning Communities”) in schools, in which teachers reflect on the approaches taken to improve the effectiveness of the intervention. A formative assessment is when feedback to students prompts them to do something different as a result and implicitly it will encourage them to reflect on the mistake made so the learning is embedded and they are less likely to make similar mistake or face the same issues in future. The programme operated in the 2015/16 and 2016/17 academic years and will stretch into September 2017 to enable the schools to discuss their public examination results.

The idea is that teachers can guide themselves through a pack of materials, to run a carefully structured series of workshops. The pack was developed by the schools, students and teachers network (SSAT) with Dylan Wiliam⁶

For schools randomly allocated to the intervention group, Dylan Wiliam introduced the concept at a launch event attended by school-nominated “Lead Teachers”. These lead teachers then supported colleagues to run a series of structured workshops (Teacher Learning Communities) throughout a two year period. Support from SSAT was also available.

Aims and Objectives

Project Hypothesis

Use by schools of the "Embedding Formative Assessment", a pack that promotes a systematic approach to developing high quality feedback through continuing professional development, will improve children's performance in academic tests at age 16.

Primary Research Question

How effective are the embedding formative assessment materials compared to usual feedback methods in terms of improving overall GCSE examination performance.

Secondary Research Question

⁶ The schools, students and teacher's Network, and Whole School Embedding Formative Assessment Resource: <https://www.ssatuk.co.uk/cpd/teaching-and-learning/embedding-formative-assessment/>

How effective are the embedding formative assessment materials compared to usual feedback methods in terms of improving examination performance in Maths and English GCSEs, i.e. subjects which are of high importance in terms of progression to employment.

Study design

Eligible population

The trial was a two arm cluster-randomised trial and included secondary schools drawn from across England. Although this is a whole-school intervention where students in all years will be exposed to the EFA methods, in order to provide more timely results this evaluation will focus on students starting Year 10 when the intervention began in 2015/16. Analysis of longer term impacts is beyond the scope of this SAP.

In order to be considered, schools had to agree to

- Provide student data so we can match to extracts from the National Pupil Database (NPD),
- Allow Continuing Professional Development (CPD) time for all teachers to participate in the *Teaching and Learning Communities*, and to
- Cooperate with the project and evaluation teams during the trial as specified in the Memorandum of Understanding with Schools.

SSAT made all secondary schools aware of the opportunity. 250 expressed an interest and 140 were chosen using a selection process that included an interview. There were no exclusion criteria and no targeting of schools by characteristics others than meeting the three criteria above.

Trial design

Randomisation took place at the school level. Schools participating in the trial were randomly assigned to either

- the intervention group, which received the *EFA* pack, one day's training from Dylan Wiliam (the programme developer) at the launch event, and ongoing support by SSAT or
- a control group, which received a one-off payment of £300, the cost of purchasing the *EFA* pack from SSAT.

Sample size

A sample size of 120 schools with equal allocation between treatment and control was recommended by NIESR (see initial sample size calculation section below). A progress report during the second term of the trial (February 2016) suggests that the total number of schools in the trial increased to 140 following initial over-recruiting. The total number of students affected by the intervention is currently not known from the quantitative sources available to NIESR.

Trial arms

School level randomisation in two arms (intervention and control groups):

Intervention group: Purchased *EFA* pack (£300) and received a two-year professional development programme on formative assessment with SSAT support. This includes a day's training for a school-nominated Lead Teacher and a pack of materials. Schools then set up "Teacher Learning Communities" (TLCs) that operated during the 2015/16 and 2016/17 academic years. TLCs met on a monthly basis to discuss and refine how they were using the materials in class. Key elements of the intervention are summarised in the table below at 4 levels, (processes, teachers, students and overall) and some example techniques are provided below the table. This approach to peer learning is not new but the extent to which it has been adopted by schools has varied and the *EFA* programme aimed to build on previous successes with the approach or, begin to embed the approach in schools and teachers within schools, who had not actively engaged with the techniques previously.

Process	<ul style="list-style-type: none"> - Monthly TLCs include: promising colleagues will try a technique, feedback on technique used in class since previous meeting, new FA content introduced and discussed - Schools provide feedback to SSAT including any issues encountered and further advice provided by SSAT
Teachers	<ul style="list-style-type: none"> - Teachers give opportunities for students to take ownership of their learning - Teachers activate students as instructional resources for each other
Students	<ul style="list-style-type: none"> - Students support each other and are more engaged - students take more responsibility for their own learning

Overall	- Responsibility for learning shared between teacher and students, students learn more
---------	--

Example Techniques

- Instead of marking each spelling or grammar mistake put a mark in the margin then students encouraged to find their own mistakes and correct them
- Mark students work in relation to their most recent marks (+ if it is better, = if of equal quality and - if not as good)
- Give anonymous feedback and encourage groups of around 4 students to decide which feedback relates to which piece of work

Control group: receives a one-off payment of £300 at the start of the trial (September 2015/16), the cost of purchasing the *EFA* pack from SSAT.

Number and timing of measurement points

Table 1 below, summarises the data collection schedule in the context of the project milestones.

Table 1 – Project milestones and data collection schedule

Date	Milestone / Data	Details
Spring 2015	Milestone	Recruitment of participating schools
July 2015	Data	Names of Participating schools provided to NIESR by SSAT, including school URN
September 2015	Milestone	Start of school year and beginning of trial
December 2015	Data	Tier 2 NPD extract provided to NIESR by DFE. This included GCSE results for the 14/15 cohort (pre-intervention); KS2 scores and eligibility for free school meals. Used to estimate sample size
January 2016	Milestone	Analysis to confirm no systematic bias in the randomisation
September 2015 – July 2017	Data	SSAT monitoring data of which schools dropped out and which schools received the related TEEP ⁷ intervention.
July 2017	Milestone	End of 2 year trial period

⁷ Teacher Effectiveness Enhancement Programme (TEEP) <https://www.ssatuk.co.uk/cpd/teaching-and-learning/teep/>

Expected November 2017	- Data	Tier 2 NPD National Pupil Database administrative data provided to NIESR by DFE, including KS4 achievement and student and school characteristics (for groups taking GCSE's in 2015/16 and 2016/17).
January 2018	Milestone	Draft Evaluation report provided to EEF by NIESR

Randomisation

Unit of randomisation

Secondary schools, which were recruited by SSAT (N=140) were randomly assigned to intervention and control groups using econometric software (Stata) within blocks (see below).

Blocking/stratification

Schools were identified as belonging to blocks based on the proportion of students in each school to achieve 5 A*-C grades in the 2014 GCSE examinations (low, medium, high – where these thresholds are chosen to achieve equal sized groups), and the proportion of students in each school to be eligible for Free School Meals (FSM, low, medium, high) using DfE sources; again thresholds were chosen to achieve equal sized groups⁸.

Within the nine blocks combining the three dimensions of GCSE performance and FSM, schools were randomly allocated to treatment and control groups (half each). This was achieved using a random number generator:

- Each school was assigned a randomly generated number between 0 and 1 using the Stata command 'runiform' with seed 2387427. The randomisation was automated by Stata and in this sense was blind. ;
- Schools were sorted by blocking variable and, within each block, by the random number
- The first school was randomised to treatment or control;
- Each subsequent school was assigned to the opposite outcome of the previous school.

Number randomised to each arm

As mentioned above, 70 schools were randomised to the treatment group and 70 to the control group.

⁸ Because of correlation between FSM and GCSE performance, a block with fewer than 6 schools would have been combined with the block with the same level of students achieving 5 A*-C at GCSE, but a higher proportion of FSM students (unless it is the high FSM block, in which case it would be combined with the medium block instead). However, this was not implemented in practice as all blocks were sufficiently populated.

Timing of randomisation relative to baseline testing

Randomisation was implemented ahead of the two-year development programme; no base line data collected other than GCSE results from the pre-intervention period, eligibility for free school meals and KS2 scores, from an NPD Tier 2 request; see below.

Calculation of sample size

In line with the standard approach in the Randomised Control Trial literature, the sample size was chosen in line with an expected effect size. Since outcome variables – i.e. the GCSE Capped 8 attainment score and attainment in English and Mathematics as individual variables – have different distributions, the sample was chosen in relation to an effect size of a Standardized Mean Difference (of 0.20 standard deviations) which equates to an improvement of approximately one third of a GCSE grade which was considered by the SSAT and EEF advisory panels to be an acceptable level of improvement from a policy perspective to roll the intervention out more widely.

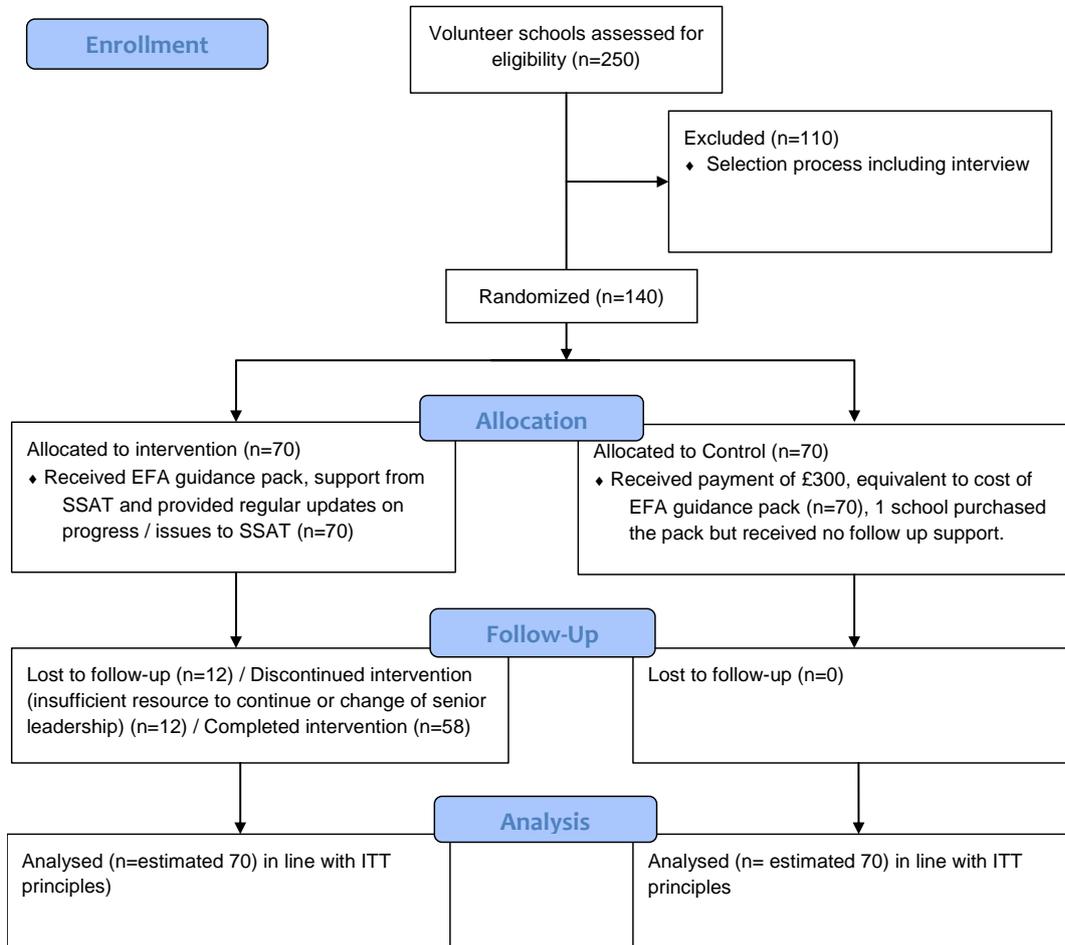
The advised sample sizes were based on an expectation that 120 schools would be allocated randomly: 60 into the treatment group and 60 into the control group, with an average of 100 students participating in each school. The total expected sample size was therefore 12,000 students (100 students per cluster for 0.05 significance level, 0.8 power, 0.20 intra-cluster correlation⁹, The calculation is no longer available because the data set used for the sample size calculation has been deleted in line with the NPD data sharing protocol. As mentioned above, 140 schools were recruited for the trial.

Follow-up

The below CONSORT 2010 chart outlines what is known to date about follow up.

⁹ Intra-cluster correlation coefficients, EEF, 2015
https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf

CONSORT 2010 Flow Diagram



Outcome measures

Primary outcome

The primary outcome will be a student’s GCSE Attainment 8 score¹⁰. This uses the new GCSE numerical grades introduced in 2016/17 which ranges between 0 and 9, using the NPD variable KS4_ATT8.

Secondary outcomes

¹⁰ Progress 8 and Attainment 8: Guide for maintained secondary schools, academies and free schools https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/583857/Progress_8_school_performance_measure_Jan_17.pdf

Two secondary outcome measures will be individual's numerical grade for Maths and English using NPD variables KS4_APMAT_PTQ_EE and KS4_APENG_PTQ_EE which again range between 0 and 9.

Baseline imbalances

This was assessed using the NPD data received in December 2015 (see table 1), including 11,000 students in the 140 schools in the study, based on the proportion eligible for FSM and grand means for Key Stage 2 results (achieved in 2011/12 (i.e. the cohort who will sit KS4 exams in 2016/17). This was based on the NPD variables EVERFSM_6 and KS4_VAP2TAAPS_PTQ_EE respectively. The previous lead analysts concluded that there were no baseline imbalances but the data has now been deleted so the effect size cannot be reported in the SAP. However effect sizes will be reported for both variables in the final report.

Analysis

Primary intention-to-treat (ITT) analysis

The two types of schools included in the trial are:

- a) intervention schools that deliver EFA
- b) control schools

The estimated impact will be based on the difference in KS4 scores between a) and b) for all schools where data is available, regardless of drop out, but only those schools and pupils who consented to be included. This is in order to estimate the "intention to treat" (ITT) effect. Analysis will be conducted in Stata.

In line with the latest EEF guidance, two models will be fitted:

- 'Simplest model': including prior attainment and allocation dummy as fixed covariates, and school as a random effect.
- 'Precise model': including prior attainment, the allocation dummy and indicator variables specifying membership of the randomisation blocks (all fixed effects) and schools as a random effect

Grand and group means will also be reported as exploratory data analysis and the impact of KS2 scores will be reported as an effect size. In the analysis below, KS4 score will be standardised using the approach outlined under effect size on page 14. The centring will be around the (treatment) group means but not around individual school means because we are interested in the intercept for the treatment but not the separate intercepts for the schools.

Model equation:

$$y = X\beta + Z\mu + \epsilon$$

Where:

$$y = \beta X + Z\mu + \epsilon$$

y = vector of outcome scores [KS4]

X = covariate matrix [KS2 scores in 'simplest' model and this plus dummies for stratification groups in the 'precise' model]

Z = design matrix identifying which school (or cluster) an individual attended.

μ = vector of school random effects

β, = fixed effect parameters

ε_{ij} = residual error term for j-th member of cluster (school) i

with the covariance structure given by Σ, where:

$$\Sigma = (\sigma_a^2 + \sigma_e^2) \begin{bmatrix} I & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & I \end{bmatrix}$$

Where σ_a^2 is a measure of school level variation; σ_e^2 is a measure of student level variation and I is given by:

$$I = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{bmatrix}$$

And ρ is the intra-school correlation coefficient:

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

The fixed effect parameters and variance components will then be estimated by restricted maximum likelihood estimation using the STATA command:

mixed KS4 KS2 TREAT i.[Tertile of school average KS2 in 2014] i.[Tertile of proportion of cohort eligible for FSM] || SCHOOL: reml - for the precise model and:

mixed KS4 KS2 TREAT || SCHOOL: reml – for the simplest model

Key model outputs will be a point estimate for the coefficient of the bivariate treatment variable [TREAT] and a 95% confidence interval for this estimate (standard Stata outputs).

If the effect size associated with the point estimate of TREAT is found to be significantly different from zero and exceeds 0.2 (the equivalent of one third of a GCSE grade as outlined in the sample size section above) the intervention will be considered to have been fully successful. If the effect size associated with the point estimate of TREAT is found to be significantly different from zero and of between 0.1 and 0.2, the intervention will be considered as a partial success. As readers will be provided with a standard 95% confidence (based on co-efficient standard errors) interval and goodness of fit

estimates (standard Stata outputs) they will be able to make their own interpretation about the weight of evidence for or against the intervention if findings are other than the two situations specified.

Interim analyses

No interim analysis was undertaken

Imbalance at baseline for analysed groups

School and pupil characteristics and measures of prior attainment will be summarised descriptively by randomised group both as randomised and as analysed in the primary analysis (for identified pupils). Continuous measures will be reported as a mean, standard deviation (SD), minimum and maximum, while categorical data will be reported as a count and percentage.

Missing data

We will describe and summarise the extent of missing data in the primary and secondary outcomes, and in the model associated with the analysis incorporating the fidelity data collected by SSAT (see below) (and also for all control variables), Reasons for missing data will also be described. For all models we will trigger a full multiple imputation strategy if more than 5% of data in the model is missing. We will also trigger imputation if more than 10% of data for a single variable or a single school is missing. The below approach will be followed separately for each instance of model and outcome for which the threshold is exceeded. The first step will be to assess whether the missing data is missing at random (MAR). We will use the standard approach where we create an indicator variable for each variable in the impact model specifying whether the data is missing or not and use logistic regression to test whether the missing status can be predicted from the following variables: all variables in the precise model plus school average KS2 and eligibility for FSM (continuous variable as opposed to tertiles), gender and ethnicity. Where predictability is confirmed we will proceed with MI. Where the missingness cannot be predicted, we will assume the data is either 'Missing Completely at Random' (MCAR) or 'Missing Not at Random' (MNAR). In the first case we are unable to observe data related to randomness and MI is not feasible. In the second case the only approach would be to adopt a structural modelling approach which we would not adopt because this would deviate from the principles of transparent reporting as findings would be assumption rather than data driven.

For the models which meet the thresholds above and for which the MAR assumption holds we will use all variables in the precise model plus those mentioned above and adopt an MI strategy using a fully conditional specification, implemented using STATA MI to create 20 imputed data sets. We will re-estimate the treatment effect using each dataset and take the average and estimate standard error using Rubin combination rules.

We will base confirmation of the effectiveness of the treatment on complete data points only but assess the sensitivity of the estimate to missingness using the imputed estimates.

If the complete data only model confirms effectiveness but the imputed estimate does not we must assume that the missing data is missing not as a random to such an extent as to invalidate our conclusion of effectiveness.

Treatment effects in the presence of non-compliance

For each school involved in the project a lead liaison was employed by SSAT to work with the schools in implementing the intervention. The SSAT employed leads completed a survey on the extent to which there was high fidelity to the planned intervention. Four questions from the survey will be used to create binary indicators of 'baseline' and 'gold standard' compliance, as outlined in table x, for use in the CACE analysis at the school level using the standard formula: $\alpha_{CL} = \alpha / P_{CL}$

Where α is the effect size based on ITT analysis for all schools who responded to the compliance survey and for whom we have NPD data, P_{CL} is the proportion of these determined to be compliers and α_{CL} is the effect size for compliers only.

Table 2

Binary measure	Question	Answer	
		Yes	No
	Q2. TLCs are meeting approximately once per month (mostly every 3-5 weeks) over the course of the year?	51	6
	Q3. TLC's held are approx. 75 minutes	53	4
	Q4. The school is on target to complete all 18 TLC's over the two years	51	6
Baseline	Answered Yes to Q2, Q3 and Q4	48	9
Gold Standard	Q8. The school has fully committed to the project providing wrap around support, compliant if answer is 3 " <i>Staff are supported beyond TLC meetings, with support/time to complete peer observations. The project is high profile with staff and students. There is regular input e.g. briefings, newsletters, celebration events etc.</i> " [Other answer options below ¹¹]	17	40

Secondary outcome analyses

¹¹ The other options are:

0 = No;

1= Minimal Support in place which has not been maintained in over time or infrequent

2 = Staff have been given regular support in between meetings for peer observations or regular briefings

Model structure for secondary outcomes will be identical to that for the primary outcome variable above. Secondary outcomes being:

- GCSE English grade (KS4_APENG_PTQ_EE)
- GCSE Maths grade (KS4_APMAT_PTQ_EE)

Additional analyses

Impact of related trials

We will look at the sub-set of schools which did not undertake the related TEEP intervention (identified in the SSAT monitoring data outlined in table 1) and again compare the parameter estimates with those from the full sample. 11 treated and 3 control schools received TEEP prior to the start of the evaluation. One treated school started TEEP and dropped out of EFA in 15/16 and one control school started TEEP in 16/17. We would expect the parameter estimates of impact to be lower for this sub-group analysis because exposure to TEEP is un-balanced (greater in the treatment group) and evidence¹² on formative assessment suggests a long lead time between embedding FA techniques and them realising their full benefits.

Subgroup analyses

Estimates will look into differential effects of three subgroups – based on the appropriate sub-samples of the data rather than additional variables, defined by student characteristics:

- Students who have ever received free school meals [NPD variable EVERFSM_6];
- Student with low attainment scores in Key Stage 2 attainment tests (bottom third as used for the randomisation) [NPD variable KS4_VAP2TAAPS] ;
- Student with high attainment scores in Key Stage 2 attainment tests (top third as used for the randomisation) [NPD variable KS4_VAP2TAAPS] .

In addition, we will undertake two separate subgroup analyses incorporating the measures of compliance described in the non-compliance with intervention section above.

¹² The schools, students and teacher's Network, and Whole School Embedding Formative Assessment Resource: <https://www.ssatuk.co.uk/cpd/teaching-and-learning/embedding-formative-assessment/>

Effect size calculation

In line with EEF best practice guidance, irrespective of whether the difference between treatment and control groups is significant, the key output will be the effect size based on the model estimate of the difference between groups divided by the pooled standard deviation (Hedges g):

$$\frac{\widehat{\beta}_{\delta}}{\sqrt{\widehat{\sigma}_e^2 + \widehat{\sigma}_a^2}}$$

Where:

$\widehat{\beta}_{\delta}$ is the model estimate of the difference between groups, and

, $\widehat{\sigma}_e^2$ and $\widehat{\sigma}_a^2$ are the estimated error components for individuals and classes,

All taken from the standard Stata outputs

.

List of variables included in the NPD request

A Tier 2 NPD request of approximately 11,100 student records from 140 schools to cover all KS4 leavers of the schools included in the trial (following opt-out) for the 2015/16 and 2016/17 academic years. All variables KS4 Tier 4 variables need to be included. Matched at individual level to prior attainment in Key Stage 2 (2010/11 and 2011/12) and further variables from School Census for 2015/16 and 2016/17. Table 3 on the following page provides a summary of the main variables used in the analysis.

Table 3: List of main variables used in the analysis:

NPD Alias	Description	Values	Tier of Variable	Use in analysis
KS4_ATT8	Capped GCSE and equivalents new style point score.		4	Primary outcome
KS4_APENG_PTQ_EE) (English attainment point Score.	0, 15, 21, 27, 33, 39, 45, 51	4	Secondary outcome/pre-programme and control
KS4_APMAT_PTQ_EE	Maths attainment point Score.	0, 15, 21, 27, 33, 39, 45, 51	4	Secondary outcome/pre-programme and control

EVERFSM_6	The student has either been eligible for free school meals at some point in the last 6 years	1 = True 0 = False	2	Control variable Level 1

Report tables

We will follow the EEF trial report template¹³ when reporting the findings from this study.

¹³ <https://educationendowmentfoundation.org.uk/evaluation/resources-centre/writing-a-research-report/>

Appendix J: Execution of Randomisation

GCSE Groups	FSM Groups	Random	Treatment
Low	Low	0.339764	0
Low	Low	0.545918	1
Low	Low	0.60219	0
Low	Low	0.660772	1
Low	Low	0.664113	0
Low	Low	0.844138	1
Low	Medium	0.024139	0
Low	Medium	0.061181	1
Low	Medium	0.154683	0
Low	Medium	0.194312	1
Low	Medium	0.203141	0
Low	Medium	0.284661	1
Low	Medium	0.325339	0
Low	Medium	0.443238	1
Low	Medium	0.548645	0
Low	Medium	0.567201	1
Low	Medium	0.623351	0
Low	Medium	0.635405	1
Low	Medium	0.63633	0
Low	Medium	0.728592	1
Low	Medium	0.777078	0
Low	Medium	0.874682	1

Low	Medium	0.933977	0
Low	High	0.016379	1
Low	High	0.083628	0
Low	High	0.092919	1
Low	High	0.09958	0
Low	High	0.117754	1
Low	High	0.159908	0
Low	High	0.198559	1
Low	High	0.30266	0
Low	High	0.347773	1
Low	High	0.373902	0
Low	High	0.424115	1
Low	High	0.476795	0
Low	High	0.619433	1
Low	High	0.627163	0
Low	High	0.659476	1
Low	High	0.716136	0
Low	High	0.792332	1
Low	High	0.806864	0
Low	High	0.822467	1
Low	High	0.876486	0
Low	High	0.883167	1
Low	High	0.88645	0
Low	High	0.917486	1
Low	High	0.934123	0
Medium	Low	0.104696	1

Medium	Low	0.106123	0
Medium	Low	0.178199	1
Medium	Low	0.203417	0
Medium	Low	0.204192	1
Medium	Low	0.331243	0
Medium	Low	0.405296	1
Medium	Low	0.490068	0
Medium	Low	0.536793	1
Medium	Low	0.555714	0
Medium	Low	0.67021	1
Medium	Low	0.766088	0
Medium	Low	0.907329	1
Medium	Low	0.923646	0
Medium	Medium	0.016362	1
Medium	Medium	0.016556	0
Medium	Medium	0.022165	1
Medium	Medium	0.10078	0
Medium	Medium	0.105873	1
Medium	Medium	0.134078	0
Medium	Medium	0.198035	1
Medium	Medium	0.218404	0
Medium	Medium	0.289876	1
Medium	Medium	0.308618	0
Medium	Medium	0.323247	1
Medium	Medium	0.343417	0
Medium	Medium	0.551702	1

Medium	Medium	0.58125	0
Medium	Medium	0.719124	1
Medium	Medium	0.753656	0
Medium	Medium	0.756596	1
Medium	Medium	0.76147	0
Medium	Medium	0.77949	1
Medium	Medium	0.881122	0
Medium	Medium	0.942159	1
Medium	Medium	0.960769	0
Medium	High	0.120685	1
Medium	High	0.190498	0
Medium	High	0.241563	1
Medium	High	0.298655	0
Medium	High	0.436302	1
Medium	High	0.534896	0
Medium	High	0.566909	1
Medium	High	0.628157	0
Medium	High	0.922626	1
Medium	High	0.974107	0
Medium	High	0.979073	1
Medium	High	0.993783	0
High	Low	0.073556	1
High	Low	0.100591	0
High	Low	0.116429	1
High	Low	0.126376	0
High	Low	0.133357	1

High	Low	0.134974	0
High	Low	0.146316	1
High	Low	0.170725	0
High	Low	0.221749	1
High	Low	0.235312	0
High	Low	0.262304	1
High	Low	0.276561	0
High	Low	0.391598	1
High	Low	0.393974	0
High	Low	0.427302	1
High	Low	0.494035	0
High	Low	0.561993	1
High	Low	0.640278	0
High	Low	0.677029	1
High	Low	0.824244	0
High	Low	0.857075	1
High	Low	0.869653	0
High	Low	0.929951	1
High	Low	0.941365	0
High	Low	0.945608	1
High	Low	0.952712	0
High	Low	0.96723	1
High	Medium	0.057991	0
High	Medium	0.131542	1
High	Medium	0.176834	0
High	Medium	0.221546	1

High	Medium	0.312878	0
High	Medium	0.355107	1
High	Medium	0.565709	0
High	Medium	0.86588	1
High	Medium	0.900495	0
High	High	0.148119	1
High	High	0.200955	0
High	High	0.245862	1
High	High	0.262619	0
High	High	0.299029	1
High	High	0.640084	0
High	High	0.765076	1
High	High	0.820427	0
High	High	0.85614	1

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

OGL This information is licensed under the Open Government Licence v3.0. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/>

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at www.educationendowmentfoundation.org.uk



Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21-24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk