Working Paper No. 21-10

The effect of embedding formative assessment on pupil attainment

Jake Anders
University College London

Francesca Foliano
University College London

Matt Bursnall
University of Sheffield

Richard Dorsett
University of Westminster

Nathan Hudson
NatCen Social Research

Johnny Runge
NIESR

Stefan Speckesser
University of Brighton

Evidence suggests that adapting teaching responsively to pupil assessment can be effective in improving students' learning. However, existing studies tend to be small-scale, leaving unanswered the question of how such formative assessment can operate when embedded as standard practice. In this paper, we present the results of a randomised trial conducted in 140 English secondary schools. The intervention uses light-touch training and support, with most of the work done by teacher-led teaching and learning communities within schools. It is therefore well-suited to widespread adoption. In our pre-registered primary analysis, we estimate an effect size of 0.09 on general academic attainment in national, externally assessed examinations. Sensitivity analysis, excluding schools participating in a similar programme at the outset, suggests a larger effect size of 0.11. These results are encouraging for this approach to improving the implementation of formative assessment and, hence, academic attainment. Our findings also suggest that the intervention may help to narrow the gap between high and low prior attainment pupils, although not the gap between those from disadvantaged backgrounds and the rest of the cohort.

VERSION: November 2021. A revised version of this article has been accepted for publication in the Journal of Research on Educational Effectiveness, published by Taylor & Francis.

Suggested citation: Anders, J., Foliano, F., Bursnall, M., Dorsett, R., Hudson, N., Runge, J., & Speckesser, S. (2021). *The effect of embedding formative assessment on pupil attainment* (CEPEO Working Paper No. 21-10). Centre for Education Policy and Equalising Opportunities, UCL. https://EconPapers.repec.org/RePEc:ucl:cepeow:21-10

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- This paper reports the findings of a large-scale trial (140 schools across England) of an intervention that aims to embed high-quality formative assessment in teachers' practice. The sample seems representative (in terms of observable characteristics) of schools in England.
- The trial was rigorously designed and its analysis was pre-registered, which is important to the credibility of the findings.
- The pre-registered primary analysis estimates an effect size of 0.09 on pupils attainment at GCSE (age 16 national examinations), which we view as a medium-sized effect, particularly in the context of a low-cost, scalable programme.
- Sub-group analysis is encouraging for the possibility that this intervention can help to narrow the attainment distribution (more effective among those with lower levels of prior attainment), although not for the possibility of weakening the link between SES and attainment.

Why does this matter?

This trial demonstrates the effectiveness of a light-touch whole school intervention seeking to embed high-quality formative assessment into teachers practice.

The effect of embedding formative assessment on pupil attainment*

Jake Anders[1], Francesca Foliano[2], Matt Bursnall[3], Richard Dorsett[4], Nathan Hudson[5], Johnny Runge[6], and Stefan Speckesser[7]

[1]UCL Centre for Education Policy & Equalising Opportunities
[2]UCL Social Research Institute
[3]University of Sheffield
[4]University of Westminster
[5]NatCen Social Research
[6]National Institute of Economic and Social Research
[7]University of Brighton

A revised version of this article has been accepted for publication in the Journal of Research on Educational Effectiveness, published by Taylor & Francis.

Abstract

Evidence suggests that adapting teaching responsively to pupil assessment can be effective in improving students' learning. However, existing studies tend to be small-scale, leaving unanswered the question of how such formative assessment can operate when embedded as standard practice. In this paper, we present the results of a randomised trial conducted in 140 English secondary schools. The intervention uses light-touch training and support, with most of the work done by teacher-led teaching and learning communities within schools. It is therefore well-suited to widespread adoption. In our pre-registered primary analysis, we estimate an effect size of 0.09 on general academic attainment in national, externally assessed examinations. Sensitivity analysis, excluding schools participating in a similar programme at the outset, suggests a larger effect size of 0.11. These results are encouraging for this approach to improving the implementation of formative assessment and, hence, academic attainment. Our findings also suggest that the intervention may help to narrow the gap between high and low prior attainment pupils, although not the gap between those from disadvantaged backgrounds and the rest of the cohort.

Keywords: Formative assessment; Embedding practice; Professional development; Randomised controlled trial; Pupil attainment.

*Contact details: Jake Anders (jake@jakeanders.uk), UCL Centre for Education Policy & Equalising Opportunities, 20 Bedford Way, LONDON WC1H 0AL

1 Introduction

> "Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited" (Black and Wiliam, 2009)

'Formative assessment' (Bloom, 1968; Bloom et al., 1971) often used interchangeably with the term 'assessment for learning' (Mittler, 1973; Wiliam, 2011), refers to any assessment activities undertaken by teachers–and by students themselves–to obtain evidence which is then used to adapt teaching and/or learning methods to meet student needs and improve learning outcomes (Black and Wiliam, 1998a). Use of the term in this way goes back to at least 1968 (Bloom, 1968), with notable reviews of its use by Natriello (1987) and Crooks (1988). In more recent years the approach was particularly popularised by Black, Wiliam and colleagues including through books aimed at practitioners (Black and Wiliam, 1998b; Black et al., 2003; Wiliam, 2017). A substantial literature theorising (Black and Wiliam, 2009, 2018), developing (Clark, 2015) and critiquing (Bennett, 2011) the approach continues to thrive.

Since high-quality feedback is key part of formative assessment (Sadler, 1998; Nicol and Macfarlane-Dick, 2006), it is heartening that pre-existing reviews of the effectiveness of improving feedback have concluded that it does improve students' learning (Hattie and Timperley, 2007). This includes the Education Endowment Foundation (EEF) toolkit review of the previous evidence, although this notes that "[m]any of the studies included are small scale studies from psychology which demonstrate theoretical principles, but which may be difficult to generalise to educational practice" (Education Endowment Foundation, 2018). However, Kingston and Nash (2011) are more critical of the existing evidence base for formative assessment interventions and conclude the title of their meta-analysis on this topic (which finds an overall weighted mean effect size of 0.20 and a median effect size of the studies reviewed of 0.25) with a call for more high-quality studies.

We would agree that the body of evidence is lacking in some respects. The evidence that exists is largely based on relatively small studies with committed teachers, supported by the close involvement of a team of researchers and recognised experts in the field (e.g. Andersson and Palm, 2017; Havnes et al., 2012). In one such example, particularly relevant to this study, as it was led by one of the co-developers of the intervention evaluated in this paper, Wiliam et al. (2004) found a mean effect size of 0.32 on pupil attainment in participating classes, compared to carefully selected comparator classes within the same schools. Smaller scale studies such as these do offer more scope to explore the psychological underpinnings of how an intervention such as this might engender the substantial change in teachers' practice needed to affect pupil outcomes (Andersson and Palm, 2018).

Nevertheless, many features of these studies may lead us to think that it will be difficult to reproduce effects at a larger scale, particularly of a similar magnitude, especially as the EEF toolkit notes that "larger scale educational studies [of feedback interventions] tend to have lower effects" (Education Endowment Foundation, 2018). Moreover, other studies that have evaluated attempts to roll-out formative assessment in a more 'hands-off' style have found much less encouraging, including negative, results (Smith and Gorard, 2005).

However, that is what the intervention in this study set out to achieve. The 'Embedding Formative Assessment' (EFA) intervention builds on Wiliam and Black's research (Black and Wiliam, 1998a, etc.), and Wiliam and Leahy's experiences with implementing formative assessment programmes (Leahy and Wiliam, 2012). Broadly, the intervention aims to support teachers to embed formative assessment strategies in their teaching practice in order to improve pupil

learning outcomes and attainment. As such, this research differs from previous studies on the effect of formative assessment in that it includes a much larger group of schools (70 treated; 70 control) and delivery that is self-administered by schools with extremely limited day-to-day engagement by experts. In effect, it is not just a test of formative assessment itself but also of this method of embedding the practice in schools. This is important because such approaches will be required for the scalability of any intervention, no matter how effective when delivered in a tightly controlled and supported manner.

Furthermore, we test the effectiveness of EFA using the highly-robust research design of a randomised controlled trial. The approach we follow is carefully chosen to minimise the potential for bias in the treatment effect, including conducting primary analysis on an 'intention to treat' basis, pre-registration of planned analyses to avoid 'p-hacking', and use of administrative outcome data to minimise the potential for selective attrition. Furthermore, the primary outcome chosen is pupils' performance in England's national, high-stakes, externally assessed examinations at age 16 (known as GCSEs; General Certificates of Secondary Education). This increases our confidence that the findings are not driven by choosing a test on which the intervention is particularly able to improve performance, which might not then be replicated in tests (such as GCSEs) shown to affect pupils' subsequent educational transitions (Anders, 2012; DfE, 2013) and later labour market outcomes (McIntosh, 2006). We believe our approach provides the best available evidence from a single study on the effectiveness of this approach to improving pupil attainment.

The main research questions this study was designed to address are as follows:

1. Primary: What is the effect on children's attainment in GCSE examinations (measured using the aggregate Attainment 8 score) at age 16 of two years of exposure to the Embedding Formative Assessment programme to improve teachers' formative assessment practices through collaborative learning?
2. Secondary: What is the effect on children's attainment in GCSE Mathematics at age 16 of two years of exposure to the Embedding Formative Assessment programme to improve teachers' formative assessment practices through collaborative learning?
3. Secondary: What is the effect on children's attainment in GCSE English at age 16 of two years of exposure to the Embedding Formative Assessment programme to improve teachers' formative assessment practices through collaborative learning?

The paper proceeds as follows. We begin by discussing the intervention and previous evidence on its efficacy in Section 2. Next, in Section 3 we introduce the data that we use as part of this project. The design of the evaluation and the analyses we conduct are reported in Section 4. These analyses include consideration of the balance and representativeness of the sample, as well as the impact estimation itself. The results of these analysis are reported in Section 5. Finally, we conclude in Section 6.

2 The intervention

This research project attempts to estimate the impact of the introduction of the "Embedding Formative Assessment" (EFA) teacher professional development programme (Leahy and Wiliam, 2013) into a school on pupils' attainment. Embedding formative assessment in teachers' practice systematically across a school requires engagement at all levels of a school, making it a kind of whole-school complex intervention of the kind that it has been highlighted are likely to be needed to improve practice (Leithwood et al., 2006; Anders et al., 2017). This also concords with one of the developer's previous work on the need for interventions to be 'tight but loose' (Thompson and Wiliam, 2008) if they are to be successfully scaled up in diverse

contexts.

EFA is designed to be an ongoing activity within a school. For the purposes of its evaluation, it was introduced to participating schools to be carried out for a minimum of two years, during the 2015/2016 and 2016/2017 academic years, with outcomes of interest measured at the end of this period. All classroom teachers in treated schools participated in the intervention and were expected to implement the strategies in lessons to pupils in all year groups across the school. The programme consists of nine monthly Teacher Learning Communities (TLCs) workshops across each academic year and monthly peer observations. Dylan Wiliam and Siobhan Leahy 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 which states that its "professional development and school improvement programmes help leaders and teachers to further outcomes for all young people, and develop leadership at all levels across the system" (SSAT, 2018).

The main element of EFA is the monthly Teacher Learning Community (TLC) workshops, which most participating schools arranged during time they already used for Continuing Professional Development (CPD). 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. It seems important, given the findings of Kennedy (2016) regarding differential effectiveness of professional learning communities, that there is clear guidance on structure and content to engage with as part of these meetings. 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 (Assessment for Learning) 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 of 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. The intervention is primarily a 'strategy-based' approach to use of formative assessment, which has been criticised by Coffey et al. (2011) as missing opportunities for engagement with disciplinary substance. However, others have argued that this is where the strength of the existing evidence of the effectiveness of formative assessment lies (Shepard et al., 2017; Wiliam, 2018).

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 run by one of the pack developers, Dylan Wiliam, and received ongoing support from a designated SSAT Lead Practitioner. Provision of this initial training day was specific to this evaluation and is not routinely provided to schools purchasing the EFA pack; the effect of the single day workshop was not assessed as part of the evaluation design, though the process evaluation showed that lead teachers found it inspirational to meet and work with Dylan William, which potentially led to higher buy-in. Most of these Lead Practitioners were currently school-based in a middle 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 initial 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. 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.

The process evaluation, carried out as part of the project, explored fidelity to the intervention design to provide a better understanding of this variation and reasons for adaptations. This included an end-of-project survey of lead teachers and visits to ten case study schools. Overall, this qualitative work found a high level of variation in how schools implemented the intervention, although some of this is to be expected within the 'tight but loose' framework (Thompson and Wiliam, 2008). 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, particularly ahead of the second year of implementation. In particular, variation was found in relation to the format/structure of TLCs and the use and frequency of peer observations.

3 Data

Both primary and secondary outcomes are derived from pupils' performances in England's national public examinations at age 16 (known as General Certificates of Secondary Education, or GCSEs). As such, our measures of attainment are externally validated and widely recognised. 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.

Data have been obtained for this analysis directly from the National Pupil Database (NPD) held by the UK Department of Education (DfE). As a result, test scores are available for the vast majority of our sample. These scores were obtained for pupils who are in Year 10 (age 15) in participating schools at the start of the intervention, with exception of pupils whose parents contacted their schools to indicate that they did not wish their offspring's data to be processed for this purpose. This exception was made in order to comply with ethical and legal considerations. Provision of information about the trial and the process for objecting to data processing was carried out prior to randomisation, meaning that it is unlikely to be occur differentially between treatment and control groups. While objection could be made at any time during the project, in practice this occurred almost exclusively prior to data collection.

The primary outcome is pupils' GCSE Attainment 8 score (DfE, 2018), which is one of England's main accountability for secondary schools and, as such, is widely recognised and clearly a measure in which schools take a keen interest. The measure provides a summary of pupils' performance across a range of subjects by aggregating pupils' best eight GCSE (General Certificate of Secondary Education; the main examinations taken by pupils at age 16) grades and double-weighting those for English and maths.[1] As such, the aggregated Attainment 8 score can range from 0 to 90. We plot a histogram of the distribution of this variable (separately for treatment and control groups) in Figure 1.

[Figure 1 about here.]

The two pre-registered secondary outcome measures were students' numerical grades for mathematics and English,[2] which again range from 0 to 9. Performance in English and maths are a core part of another of England's accountability measures: the English Baccalaureate. As such, performance in these key subjects is, again, instrumentally important for English schools. More importantly, performance in all these tests has also been found to be important for pupils' future life chances, with evidence that GCSE performance predicts later educational

[1] Specifically, we use the variable KS4_ATT8, as provided in the DfE's National Pupil Database. There is no longer any need to convert between letter grades and numbers (as described in the project protocol) as this cohort received GCSE numerical grades, introduced in 2016/2017, which range from 0 to 9.

[2] Specifically, we use NPD variables KS4_APMAT_PTQ_EE for maths performance and KS4_APENG_PTQ_EE for English performance.

outcomes such as university attendance (DfE, 2013; Anders, 2012) and labour market outcomes (McIntosh, 2006). In particular, there may be a particular benefit from performance in English and maths (Dolton and Vignoles, 2002). In this paper, we also conduct exploratory analysis of pupils' performance in science, humanities and languages[3] to provide additional context to our overall findings.

Prior attainment of the same pupils when aged 11 (measured using national examinations taken at the end of primary schooling)[4] were also obtained. These provide us with information useful to assessing the extent to which treated and control schools are balanced on observable characteristics and, given the predictive power of prior attainment, improving the precision of our treatment estimates. Previous work has estimated strong correlations between these national measures of performance at ages 11 and 16 (Benton and Sutch, 2014), something borne out in the our analysis below. Pupils without prior attainment (less than 2% of the total sample) are excluded from all modelling to ensure consistency of the composition of the sample across models whether or not this is included as a covariate.

We report the trial's CONSORT diagram in Figure 2, demonstrating the flow of schools and pupils through the trial. This highlights the very low levels of attrition from the trial, particularly in terms of follow up, demonstrating the significant benefits of using administrative data as the source of the prior attainment and outcome measures. What loss there is appears primarily to be due to individuals not studying mathematics or English GCSEs, which is in some ways an outcome in itself.

[Figure 2 about here.]

We also worked with the project delivery team to define and capture two binary indicators of compliance with the programme:

- 'Minimal' compliance, simply an indicator of whether the school was still at all engaged with the programme by the end of the two years. 58 of the 70 treatment schools fall into this category.
- 'Maximal compliance', based on survey responses by SSAT Lead Practitioners indicating that the school has fully committed to the project providing wrap around support indicated by the response "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.". Only 14 of the 70 treatment schools fall into this category.

We note that some concerns were raised about this measure of compliance as part of the process evaluation. The compliance measure was compiled by individual SSAT Lead Practitioners who had relatively limited involvement in the schools, meaning they may not have been in the best place to make these judgements. In addition, the process evaluation found that people involved in the intervention had very different interpretations of what constituted high and low engagement and compliance.

4 Design and analysis

4.1 Evaluation design

In this paper, we estimate the effect of a school participating in the 'Embedding Formative Assessment' programme using a randomised controlled trial (RCT). As the intervention is in-

[3] Specifically, we use the NPD variables: KS4_SCIATT_PTQ_EE for science; KS4_EBACHUM_PTZ_EE for humanities; and KS4_EBACLAN_PTQ_EE for languages.

[4] Specifically, we use the NPD variable KS4_VAP2TAAPS_PTQ_EE.

herently whole school in nature, it is not possible to randomise the treatment within schools (for example, to half the teachers in a school). Instead, we selected a two-armed blocked/stratified school-level cluster randomised controlled trial (cRCT).

Blocking/stratification was undertaken to minimise the risk of bias at baseline by factors of particular relevance to the study. Proportion of the school eligible for free school meals (FSM) was chosen as a factor because of our intention to carry out a subgroup analysis for pupils meeting this criteria. School attainment was also used as a blocking factor because of the potential for differential impact of EFA by ability. For each characteristic, schools were split into three equally sized quantile groups; blocks were then formed by the nine possible combinations of these three groups. Since the two blocking characteristics are (negatively) correlated, this could have resulted in the blocks for high attainment and high FSM proportion, and for low attainment and low FSM proportion, being small. As such, block combination rules were drawn up with any block with fewer than six schools being 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.

A power calculation was conducted to estimate the sample size required to achieve a Minimum Detectable Effect Size (MDES) of 0.20 for a statistical test at the 0.05 level of significance with 0.8 power. These calculations were based on the following assumptions: an expected average of 100 students in Year 10 at each participating school at the start of trial, a within-school pre-test to post-test correlation of 0.66, a between-schools pre-test to post-test correlation of 0.57, and an 0.20 intra-cluster correlation in the outcome measure. These calculations suggested that recruitment of 120 schools would meet this requirement. Ultimately, the project team were successful in recruiting 140 schools, which contributed to a reduction in the MDES achieved to 0.18.

Within the nine blocks, participating schools were randomly assigned to one of two trial arms in equal proportions. These arms were:

- the treatment group, which received the intervention described in Section 2 above; or
- a control group, which received a one-off payment of £300 at the start of the trial (September 2015), which is equivalent to the cost of purchasing the EFA pack from SSAT.

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 for the duration of the trial.

This random assignment was carried out as follows. Each school was assigned a randomly generated number between 0 and 1 using the Stata 'runiform' function with seed 2387427 to allow for verification. 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 then assigned to the opposite outcome of the previous school. Since this randomisation process was automated using statistical software Stata it was, in this sense, blind.

The evaluation design was published in an evaluation protocol (Anders, 2016) and registered in the ISRCTN registry with registration number ISRCTN10973392 (ISRCTN, 2015). The study was approved through the ethics processes of the National Institute of Economic and Social Research.

4.2 Balance and representativeness

Randomisation of schools to treatment and control groups leads to balance of all observable and unobservable characteristics between these groups, in expectation. However, there always remains the risk of differences emerging by chance or due to post-randomisation selection effects (such as non-random attrition). While our design aims to minimise such possibilities, through use of blocking on key characteristics in randomisation to reduce imbalance and the use of administrative data to avoid missing outcome measurement (only 0.2% of the primary outcome data is missing at the pupil-level), it is important to verify observable differences are minimal.

To do this, we report key school- and pupil-level characteristics in our sample, in the treatment and control groups, and the differences between these two groups. In the case of categorical characteristics, these differences are expressed in terms of percentage point (%pt.) differences; in the case of continuous characteristics, these differences are expressed in terms of both unstandardised median differences and standardised mean differences (Imbens and Rubin, 2015). The standardised difference is calculated as the unstandardised difference between the mean of the characteristic in each group divided by the overall sample standard deviation, as follows:

$$\delta = \frac{\mu_{\text{Treat}} - \mu_{\text{Control}}}{\sigma_{\text{Sample}}} \tag{1}$$

To provide additional context, we also (where possible) report details of the corresponding national average characteristics. This provides important context about the representativeness of our sample of schools relative to those in the country at large.

4.3 Primary analysis

Our primary analysis, as pre-registered in the evaluation protocol (Anders, 2016), estimates the effect of the intervention (captured by a school-level binary variable) on pupils' Attainment 8 GCSE score among the intention to treat (ITT) sample using a linear regression model including a school-level random effect:

$$y_{ij} = \alpha + \beta_1 \text{Treat}_j + \beta_2 \text{KS2}_{ij} + \textbf{Block}_j + \gamma_j + \varepsilon_{ij} \tag{2}$$

where y is the outcome variable of interest for pupil i in school j, Treat is a school-level treatment indicator, KS2 is a pupil-level variable capturing pupils' prior attainment in order to improve the precision of our treatment estimates (KS2 refers to tests taken at the end of the English Education's Key Stage 2 i.e. at age 11), **Block** is a vector of randomisation blocks, γ is a school-level random effect, and ε is a pupil-level idiosyncratic error term. All standard errors are calculated taking into account the potential for school-level clustering effects.

We estimate three further related models as follows:

- M0: simple linear model (i.e. excluding the school-level random effect) including only the treatment dummy variable to demonstrate the result based on raw difference in means;
- M1: linear model including only the treatment dummy and school-level random effect;
- M2: as M1 but adding KS2 prior attainment to increase the precision of the treatment estimate (Bloom et al., 2007);
- M3: as M2 but adding randomisation block dummy variables (i.e. the full specification outlined above) to ensure analysis fully aligns with evaluation design (Rubin, 2008).

To aid comparability with other evaluations of similar interventions, we convert the treatment effect estimate recovered by β_1 into an effect size. We do this by dividing the raw estimate by the unconditional total pooled standard deviation (Cohen, 2013) of the outcome variable as follows:

$$\delta = \frac{\beta_1}{\sigma_{pooled}} \tag{3}$$

where β_1 is the estimate of the treatment effect derived from the primary analysis model in equation 2, and the pooled unconditional total standard deviation σ_{pooled} is estimated as follows:

$$\sigma_{pooled} = \sqrt{\frac{(n_{treat}-1)\sigma^2_{treat} + (n_{control}-1)\sigma^2_{control}}{n_{treat}+n_{control}-2}} \tag{4}$$

in which σ^2_{treat} is an estimate of the unconditional total variance in the treatment group and $\sigma^2_{control}$ is an estimate of the unconditional total variance in the control group both estimated from the intention to treat sample used in the primary analysis.

4.4 Additional analysis and heterogeneity

We conduct a number of additional analyses of three main types:

1. Secondary outcome analysis
2. Sub-group (heterogeneity) analysis
3. Complier analysis

Most of these analyses are pre-registered in the project's protocol and statistical analysis plan, however a small number were not included so should be treated as exploratory. These are clearly identified when they are introduced and in reporting the results so that appropriate caution may be taken in their interpretation.

All the secondary outcome analyses are estimated in exactly the same way as the primary analysis, other than the substitution of the outcome variable. As noted above, the two pre-registered secondary outcome measures were student's numerical grades for mathematics and English. In addition, we consider pupils' performance in science, humanities, and languages as additional exploratory analyses.

All the sub-group analyses are estimated in exactly the same way as the primary analysis, other that the exclusion from the estimation sample of those not fitting the sub-group criterion. Some of these sub-groups are defined on a pupil-level basis and some are defined on a school-level basis. The school-level sub-groups were not identified in the project protocol and, as such, should be treated as exploratory.

The pupil-level sub-groups considered are as follows:

- Free School Meals (FSM) eligible pupils, specifically those who have ever been identified as eligible for Free School Meals in the National Pupil Database;
- Low prior attainers, defined as the bottom tertile of prior attainment defined using Key Stage 2 (age 11) test performance;
- Medium prior attainers, defined as the middle tertile of prior attainment defined using Key Stage 2 (age 11) test performance;

- High prior attainers, defined as the top tertile of prior attainment defined using Key Stage 2 (age 11) test performance.

The school-level sub-groups considered are as follows:

- Low average prior attainment, defined as the bottom tertile of prior average attainment as described for the purposes of randomisation blocking;
- Medium average prior attainment, defined as the middle tertile of prior average attainment as described for the purposes of randomisation blocking;
- High average prior attainment, defined as the top tertile of prior average attainment as described for the purposes of randomisation blocking.
- Non-TEEP schools, defined as schools not subsequently identified by SSAT from their administrative records as also participating in the Teacher Effectiveness Enhancement Programme (TEEP);

Complier analyses are carried out using a two stage least squares instrumental variables technique by estimating a (first stage) model of compliance, as follows:

$$\text{Comply}_j = \alpha + \beta_1 \text{Treat}_j + \beta_2 \text{PreTest}_{ij} + \textbf{Block}_j + \xi_{ij} \tag{5}$$

where Comply is a binary compliance variable (discussed in Section 3), and ξ is an error term. The predicted values of Comply from the first stage are used in the estimation of a (structural) model of our outcome measure y_{ij}. Note that no school-level random effect is included in the instrumental variable modelling. In other respects, the specification remains the same as the primary outcome ITT model. The second stage model is specified as follows:

$$y_{ij} = \alpha + \beta_1 \widehat{\text{Comply}}_j + \beta_2 \text{PreTest}_{ij} + \textbf{Block}_j + \omega_{ij} \tag{6}$$

where $\widehat{\text{Comply}}_j$ are the predicted values of treatment receipt derived from the first stage model, and ω is an error term. Our primary outcome of interest will be β_1, which should recover the effect of the intervention among compliers. Standard errors will be clustered at the school level and adjusted due to the instrumental variables approach.

4.5 Process evaluation

The implementation and process evaluation, carried out as part of the project, included a number of elements. Visits were carried out in ten treatment schools between May 2016 and September 2016. These visits included interviews with lead teachers, focus groups with TLC leads and teachers, observations of TLC sessions and in some cases an interview with the headteacher. Case study schools were selected to include a variety of delivery contexts in accordance to Ofsted rating, proportion of FMS pupils and geographical location, but the sample is not necessarily representative. As such, the qualitative findings provide useful insights about the range and diversity of views and experiences, rather than necessarily the views of the wider population. The data was analysed in NVivo using a framework approach, coding the data into themes and issues.

In addition, lead teachers in treatment schools and lead applicant contacts in control schools were surveyed at the end of the project between June and July 2017. The treatment survey was completed by 40 schools, equivalent to 57 per cent of all treatment schools, or 69 per cent of schools that finished the programme. The control survey was completed by 39 schools, equivalent to 57 per cent of control schools. Finally, the initial training day in September 2015 and the end-of-project event in September 2017 were observed, and all training content and

project resources were reviewed.

5 Results

5.1 Balance and representativeness

Baseline characteristics by treatment group are reported in Table 1. Given that these groups were randomly assigned we have no reason to expect systematic differences between them in terms of any observable or unobservable characteristics. However, it is nevertheless important to check for these which might indicate problems such as systematic differential attrition.

Reassuringly, there is no evidence of such differences. At the school-level there are similar proportions of 'academy' schools and schools with Ofsted ratings of good or outstanding in the treatment and control groups. In terms of pupil characteristics, a similar proportion of pupils are eligible for free school meals (a proxy for low income) in both groups and prior attainment is also similar both in terms of the median (29.05 vs 28.91) and the mean (27.0 vs. 26.81). There is a slight difference in the number of Year 11 pupils between the two groups, with the control group schools having marginally more pupils in Year 11 than in the intervention schools (medians of 179 vs. 175.5).

Overall, there is little to suggest systematic differences between the groups in terms of these observable characteristics. We believe that this adds to the confidence that our trial has strong internal validity and, thus, that the treatment estimates reported below may be interpreted as causal.

However, we should also consider the external validity of these results and, hence, the extent to which we believe our findings to be generalisable to the wider population of schools in England. To understand this, we compare our sample with nationally reported statistics about these school- and pupil-level characteristics.

The sample of schools included in this project are slightly less likely to be religiously affiliated, more likely to be academies, are slightly larger, have a larger share of pupils eligible for FSM (which was targeted in recruitment), and a larger share of pupils for whom English is an additional language. However, with the possible exception of the proportion that are academies, we do not think our sample is dramatically different from a nationally representative sample, at least in terms of observable characteristics. Unfortunately, we are not able to compare the measure of prior attainment we use, as average point scores are not reported in the national statistics for this year; that said, the other characteristics that we are able to compare are not suggestive of a particularly more advantaged sample of pupils, who we might expect to perform better. Overall, we think this analysis is encouraging for our ability to generalise our findings.

[Table 1 about here.]

5.2 Primary analysis

In this section, we report the outcomes of our pre-registered primary outcome analysis models, contextualised with related models. These are reported in Table 2, with the pre-registered primary analysis model is reported as M3. The treatment effect converted into a Cohen's d effect size[5] is reported at the base of the regression table.

[Table 2 about here.]

[5] Adjustment of our reported Cohen's d effect size into a Hedges' g effect size (Hedges, 2007) makes no difference to the effect sizes in this paper when reported to two decimal places as the correction factor becomes exceedingly small as a trial grows in size.

Although the primary analysis model for this evaluation is pre-specified (as dicussed above), it is helpful to contextualise this model by building it up from the simplest way of estimating the treatment effect in this context, which is simply to compare the treatment and control group means. The coefficient on the treatment variable in M0 recovers exactly this and tells us that the unconditional mean Attainment 8 score of pupils in the treatment group is 1.4 points (an effect size of 0.08) higher than is the case for pupils in the control group.

However, as pupils are nested within schools as part of this evaluation, we add school-level random effects to the model (M1) which may help to improve the efficiency of our estimates by controlling for unobserved school-level differences in pupils' performance, while making some some additional assumptions about the distribution of these unobserved random effects. Conditional on these school-level effects, our estimate of the treatment effect increases to 2.0 points (an effect size of 0.11).

Next, we add a measure of pupils' prior attainment at age 11 to the model (M2). Although we showed in Table 1 that there are only small differences in prior attainment between the treatment group, as would be expected given random allocation, including this characteristic in the model helps to improve the precision of our treatment estimate by effectively comparing differences in performance between the treatment and control groups among those with the same level of prior performance. This increased precision is evident from the increased t-statistic on our treatment estimate resulting from this model change. This is despite a slight reduction in the estimated effect to 1.8 points (an effect size of 0.10).

Finally, we add dummy variables to capture the importance of the school-level blocking that fed into randomisation. It is important to include design features such as this in the analysis model (Rubin, 2008), in particular reflecting the reduced statistical degrees of freedom inherent in using this method of stratified randomisation. However, in our case the inclusion of the blocking variables is also likely to increase the precision of our estimate (as with inclusion of prior attainment) because the blocks were based on stratification by school-average prior attainment and proportion of pupils from low income families, both of which are associated with differences in our outcome of interest.

Having done this, we arrive at our primary estimate of the change in performance associated with a school being allocated to the treatment group in this trial. Pupils in schools allocated to the treatment group have, on average, 1.7 higher Attainment 8 points than pupils in schools allocated to the control group. This translates to a Cohen's d effect size of 0.09 and is roughly equivalent to an improvement of almost two grades across a pupil's best eight subjects. We acknowledge that this effect is not statistically significant at the conventional 5% level, although it is significant at the less-demanding 10% level. That said, this trial was designed to have the statistical power to detect an effect size of 0.20, rather than the 0.09 that we ultimately estimate. This was largely for reasons of cost and practicality, since the trial already involved the systematic delivery of the programme to 140 schools across England.

5.3 Additional analysis and heterogeneity

We explore these findings further in three main ways. First, through consideration of secondary outcome measures. Second, through consideration of differential impacts for sub-groups. Finally, by estimating the effect of school compliance with the intervention.

[Table 3 about here.]

We begin with secondary outcome measures. Alongside the effect on pupils' best 8 GCSEs, we specifically look for an impact on performance in English and mathematics. The estimated effects for both of these are considerably smaller than the effects we estimate for pupils' performance in general.

Based on this, we carried out further exploratory analysis of pupils' performance on other components of the EBacc: Science, Humanities and Languages. Unlike English and mathematics, these subjects are not compulsory and, as such, we first checked if there was evidence of systematic differences in completing these qualifications. We find no evidence of differences in the proportion of pupils taking these subjects between the treatment and control groups. This is unsurprising, given that the intervention would not have started until after pupils' subject choices had already been made. However, this provides some reassurance that differences in the composition of those studying such subjects are unlikely to affect our findings. Turning to the impact estimates themselves, we find larger effects for languages than those evident in English and maths, suggesting that the overall performance improvements were particularly driven by changes in these subjects.

[Table 4 about here.]

[Table 5 about here.]

Next, we consider differential impacts among specific sub-groups. Two of these are pupil-level sub-groups (reported in Table 4), while a further two are defined at the school-level (reported in Table 5). As noted above, the school-level sub-groups were not pre-specified in the evaluation protocol, however, the analysis of non-TEEP schools was specified in the statistical analysis plan, based on the finding in the process evaluation that previous or current involvement in TEEP strongly influenced delivery and experiences of delivery.

Among pupil-level effects, we first estimate the effect among the sample identified as eligible for 'free school meals', which is an imperfect but commonly available administrative proxy for living in a low income household. For EFA to be likely to reduce educational inequality associated with family background, we would need to see a larger effect of the intervention on this group.

Unfortunately, our estimated effect of the intervention for this sub-group is smaller than that for the sample as a whole, being closer to an effect of one improved grade among an individual's best eight (an effect size of 0.07, compared to 0.09 for the sample as a whole). It should be noted that the effects for this sub-group is not statistically significantly different from that for the rest of the sample, however this certainly does not provide evidence of greater effectiveness for this disadvantaged group.

We also stratify our sample by prior attainment into three approximately equally sized groups we refer to as 'low', 'medium' and 'high' attainment[6] to explore the potential for differential effects depending on pupils' prior performance. As with our analysis by FSM-eligibility, larger effects among those with initially low attainment than among those with initially high attainment are suggestive that the intervention helps to narrow educational inequality, and vice versa.

Our analysis shows stronger support for this former possibility, with a considerably larger effect size evident among the 'low' prior attainers, and the smallest effect size among those with 'higher' prior attainment. As with the FSM sub-group analysis, it is not possible to say that the effects among these different sub-groups are statistically significantly different from one another.

We turn next two to school-level sub-groups. First, we again stratify on the basis of prior attainment, but this time on the basis of school-level averages in prior attainment. Perhaps surprisingly, given our findings stratified by pupil-level prior attainment, we find evidence that schools with low attainment intakes see, if anything, negative effects from engaging in the

[6] As stratification for randomisation was done at the school level it is not the case that these groups are quite the same size between the treatment and control groups. We do not expect this substantively to affect our estimates.

programme, while schools with high attainment intakes achieve substantial positive effects. Taken together with our pupil-level prior attainment sub-groups, this suggests the largest effects are likely to be for low attainment pupils in high performing schools. Exploratory analysis of the treatment effects among pupils with different levels of prior attainment in high average prior attainment schools supports this suggestion.

Finally, we restrict our analysis to those who were not already participating in the Teacher Effectiveness Enhancement Programme (TEEP) at the start of the trial. TEEP and EFA are built on similar collaborative learning principles based on interactive workshops, and the process evaluation found that already using TEEP often changed how Lead Teachers implemented EFA, thus diluting EFA's potential impact. Once we exclude those who were already participating in TEEP, we find a larger effect size of 0.11.

[Table 6 about here.]

We turn finally to analysis of effects among those who complied most fully with the programme. The results are reported in Table 6, with the first column reporting the primary analysis Intention to Treat estimate as a point of comparison. The next two columns focus on the 'minimal compliance', reporting the First Stage estimates on the left and the IV treatment estimate on the right. The first stage demonstrates that treatment status is a strong instrument for this measure of compliance. In the IV model itself, minimal compliance is only found to have a slightly larger effect (d=0.11) than in the ITT analysis, which is understandable given the 82% compliance rate based on this measure.

The final two columns focus on schools that achieve 'maximal compliance', reporting First Stage estimates on the left and the IV treatment estimate on the right. Again, we see evidence that the treatment allocation is a strong instrument for this level of compliance (albeit not as strong as for 'minimal compliance'). Schools found to have achieved high levels of compliance are estimated to achieve substantially larger treatment effects (d=0.22) compared to the intention to treat analysis. However, we should sound a note of caution about this finding because, although this finding is large, there is a substantial reduction in statistical power because we are focusing on the local average treatment effect; as such, this finding is not statistically significantly different from the ITT estimate or from zero, even at the less demanding 10% level of significance.

5.4 Process evaluation

The implementation and process evaluation found that the TLC workshop format was often seen as the key element of the EFA programme. While the exact TLC structure varied by school, participants found that the interactive TLC sessions provided a useful forum for effective sharing and reflection of teaching and learning, leading to improved practices by allowing for valuable dialogue and encouraging experimentation with formative assessment techniques. This led some lead and headteachers to report that the EFA programme had had a positive impact on the school culture, by increasing dialogue between teachers including outside TLC sessions.

The data show that teachers also valued the formative assessment content itself, and found it useful to have a toolbox of different techniques. The formative assessment techniques were generally not seen as revolutionary or ground-breaking, but the monthly TLC sessions and the sustained two-year focus on formative assessment helped refocus staff attention on applying and embedding already-existing good formative assessment practices. As such, the formative assessment content and TLC process was seen to go hand-in-hand. The fact that the programme focused on already-existing formative assessment practices also meant that it was not considered to be an onerous exercise that placed undue additional pressures on

teachers.

The process evaluation identified that teachers thought the real benefits of the programme would be seen in the longer-term. They noted it was a longer process to embed the formative assessment principles into practice, and especially for this to change pupils' approaches to learning and feed into attainment. Teachers, however, reported a number of perceived improvements in non-cognitive outcomes such as behaviour, concentration, confidence and communication. Some teachers reported that younger pupils were more receptive to the techniques, partly because they were less exam-minded. Taken together, these factors may mean that our results, with an older cohort immediately following the two-year intervention, show a minimum effect, and that future studies should explore the effect a number of years after the exposure to the intervention.

6 Discussion

In this paper we have provided high-quality new evidence on the effect of the Embedding Formative Assessment (EFA) intervention, a largely self-administered approach to improving the use of formative assessment in schools. Our findings are from a large scale cluster randomised controlled trial and approach we followed throughout was carefully chosen to minimise the potential for bias in the treatment effect, including conducting primary analysis on an 'intention to treat' basis, pre-registration of planned analyses to avoid 'p-hacking', and use of administrative outcome data to minimise the potential for selective attrition. Furthermore, the primary outcome chosen is pupils' performance in England's national, high-stakes, externally assessed examinations at age 16. As such, we believe our approach provides the best available evidence from a single study on the effectiveness of this approach to improving pupil attainment.

Our results are encouraging for this approach to improving the implementation of formative assessment and, hence, academic attainment, in English secondary schools. In our pre-registered primary analysis, we estimate an effect size of 0.09. We follow Kraft (2020) in viewing this as a medium-sized effect, particularly given the context of this as a low-cost, scaleable, programme analysing the causal effect in a broad sample on a non-proximal outcome, with analysis carried out on an intention to treat basis.

After excluding schools who were found to previously or currently be involved in a similar programme (the 'TEEP' programme), we estimate a larger effect size of 0.11. We acknowledge that this latter analysis was only pre-registered in the project's statistical analysis plan, which was published a few months before analysis but – importantly – still before availability of outcomes data, rather than the evaluation protocol agreed at the design phase. Nevertheless, this sub-group analysis excluding schools already participating in TEEP together with our complier analyses, both suggest that effects may be strongest in the schools whose practice changed the most as a result of implementing Embedding Formative Assessment. We take this as indicating the robustness of our findings, since a clear relationship between dose and response is typically seen as adding weight to the causal interpretation of findings.

The implementation and process evaluation adds further robustness to these findings. It found that schools and teachers valued the programme. In particular, the monthly TLC sessions facilitated valuable dialogue between staff, and the sustained two-year focus on formative assessment helped refocus staff attention on applying and embedding already-existing good formative assessment practices. In addition, these interviews with teachers suggested that they found younger pupils to be more receptive to the intervention than their older and more exam-focused peers, which implies the potential for larger effects in later cohorts than those that we were able to analyse as part of this study.

Our pre-registered pupil-level sub-group analyses are not especially encouraging for the possi-

bility that implementation of EFA will help to particularly improve the performance of those from low income backgrounds (as captured by the administrative eligibility for free school meals indicator). However analysis of treatment effect stratified by pupil-level prior attainment does suggest that effects are stronger among those with lower levels of prior attainment, although the difference in treatment effect estimates between our prior attainment sub-groups are not statistically significantly different from one another. Interestingly, this findings flips if we stratify instead by school-level average prior attainment with larger effects among schools with those higher average test scores on intake: we suspect that it may be easier for schools with such intakes to implement programmes such as EFA, perhaps also reflecting higher prior attainment intakes being correlated with higher average SES intakes.

We did not find much evidence of an effect of the intervention specifically on pupils' performance in English or mathematics. This implies that it is performance in other subjects that improves. We provide some exploratory evidence on this by looking at the effects on performance in science, humanities and languages, finding larger effects in languages, particularly. However, we are unable to comment on why this might be; further research specifically designed to test areas of the curriculum in which formative assessment has more substantial differences would help to explain our findings.

The sample in this study is reasonably representative of the population of state-funded English secondary schools, at least in terms of observable characteristics. That said, we should always remain somewhat cautious about the extent to which a wider roll out of the EFA programme would achieve effects of the magnitude we have observed in this research, although schools choosing to implement EFA in future are likely to be motivated to do so by many of the same factors that led to schools choosing to join our trial, so may be more like the study sample than a random sample of English schools. Nevertheless, we advocate the continued evaluation of EFA in different contexts to continue to build the evidence on the conditions required for it to make the biggest differences to pupil performance.

Word count. 8,309.

Acknowledgements. We gratefully acknowledge the following for their assistance throughout the project, without which it would not have been possible: the development team at SSAT including Fie Rason, Corinne Settle and Anne-Marie Duguid; other members of the implementation and process evaluation team, including Heather Rolfe; the Department for Education (DfE) National Pupil Database (NPD) team, particularly Zoe Davison; the EEF evaluation and projects teams, particularly Elena Rosa Brown, Eleanor Stringer and Guillermo Rodriguez-Guzman. Thanks also to Ruth Dann, Jeremy Hodgen, John Jerrim, and Dominic Wyse for helpful comments and suggestions.

Funding details. This research was funded by the Education Endowment Foundation (EEF). The results of the trial were initially published by the EEF in an evaluation report (Speckesser et al., 2018) with this article building on that initial report.

Disclosure statement. The authors declare no conflicts of interest.

Data availability statement. Data analysed as part of this project has been archived and are available on application to the Education Endowment Foundation Data Archive.

References

Anders, J. (2012). The link between household income, university applications and university attendance. Fiscal Studies, 33(2):185–210. doi:10.1111/j.1475-5890.2012.00158.x.

Anders, J. (2016). Embedding formative assessment: Evaluation protocol. Report, Education

Endowment Foundation. Available from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_EmbeddingFormativeAssessment.pdf.

Anders, J., Brown, C., Ehren, M., Greany, T., Nelson, R., Heal, J., Groot, B., Sanders, M., and Allen, R. (2017). Evaluation of complex whole-school interventions: Methodological and practical considerations. A Report for the Education Endowment Foundation, Education Endowment Foundation, London, UK.

Andersson, C. and Palm, T. (2017). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. Learning and Instruction, 49:92 – 102. doi:10.1016/j.learninstruc.2016.12.006.

Andersson, C. and Palm, T. (2018). Reasons for teachers' successful development of a formative assessment practice through professional development – a motivation perspective. Assessment in Education: Principles, Policy & Practice, 25(6):576–597. doi:10.1080/0969594X.2018.1430685.

Bennett, R. E. (2011). Formative assessment: a critical review. Assessment in Education: Principles, Policy & Practice, 18(1):5–25. doi:10.1080/0969594X.2010.513678.

Benton, T. and Sutch, T. (2014). Analysis of use of key stage 2 data in GCSE predictions. Report to Ofqual Ofqual/14/5471, Cambridge Assessment, Cambridge, UK.

Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D. (2003). Assessment for Learning : Putting It into Practice. McGraw-Hill Education, New York, NY.

Black, P. and Wiliam, D. (1998a). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1):7–74. doi:10.1080/0969595980050102.

Black, P. and Wiliam, D. (1998b). Inside the Black Box: Raising Standards Through Classroom Assessment. King's College London.

Black, P. and Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1):5–31. doi:10.1007/s11092-008-9068-5.

Black, P. and Wiliam, D. (2018). Classroom assessment and pedagogy. Assessment in Education: Principles, Policy & Practice, 25(6):551–575. doi:10.1080/0969594X.2018.1441807.

Bloom, B. S. (1968). Learning for mastery. Evaluation Comment, 1(2):1–12.

Bloom, B. S., Hastings, J. T., and Madaus, G. F., editors (1971). Handbook on the Formative and Summative Evaluation of Student Learning. McGraw-Hill, New York, NY.

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educational Evaluation and Policy Analysis, 29(1):30–59. doi:10.3102/0162373707299550.

Clark, I. (2015). Formative assessment: translating high-level curriculum principles into classroom practice. The Curriculum Journal, 26(1):91–114. doi:10.1080/09585176.2014.990911.

Coffey, J. E., Hammer, D., Levin, D. M., and Grant, T. (2011). The missing disciplinary substance of formative assessment. Journal of Research in Science Teaching, 48:1109–1136. doi:10.1002/tea.20440.

Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences. Taylor & Francis, Abingdon.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, 58:438–481. doi:10.3102/00346543058004438.

DfE (2013). A comparison of gcse results and as level results as a predictor of getting a 2:1 or above at university. DfE Research Report DFE-00060-2013, Department for Education.

DfE (2018). Secondary accountability measures. DfE Guide for maintained secondary schools, academies and free schools DFE-00278-2017, Department for Education.

Dolton, P. J. and Vignoles, A. (2002). The return on post-compulsory school mathematics study. Economica, 69(273):113–142. doi:10.1111/1468-0335.00273.

Education Endowment Foundation (2018). Feedback. Teaching & learning toolkit, Education Endowment Foundation.

Hattie, J. and Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1):81–112. doi:10.3102/003465430298487.

Havnes, A., Smith, K., Dysthe, O., and Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. Studies in Educational Evaluation, 38(1):21 – 27. doi:https://doi.org/10.1016/j.stueduc.2012.04.001.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. Journal of Educational and Behavioral Statistics, 32(4):341–370. doi:10.3102/1076998606298043.

Imbens, G. M. and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, New York, NY.

ISRCTN (2015). Embedding formative assessment. Trial registration, International Standard Randomized Controlled Trial Number Registry. https://doi.org/10.1186/ISRCTN10973392.

Kennedy, M. M. (2016). How does professional development improve teaching? Review of Educational Research, 86(4):945–980. doi:10.3102/0034654315626800.

Kingston, N. and Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. Educational Measurement: Issues and Practice, 30(4):28–37. doi:10.1111/j.1745-3992.2011.00220.x.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher. doi:10.3102/0013189X20912798.

Leahy, S. and Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In Gardner, J., editor, Assessment and Learning, chapter 4. SAGE, London, UK.

Leahy, S. and Wiliam, D. (2013). Embedding Formative Assessment. Specialist Schools and Academies Trust, London, UK.

Leithwood, K., Day, C., Sammons, P., Harris, A., and Hopkins, D. (2006). Seven strong claims about successful school leadership. Research report, National College for School Leadership.

McIntosh, S. (2006). Further analysis of the returns to academic and vocational qualifications. Oxford Bulletin of Economics and Statistics, 68(2):225–251. doi:10.1111/j.1468-0084.2006.00160.x.

Mittler, P. J. (1973). Purposes and principles of assessment. In Mittler, P. J., editor, Assessment for learning in the mentally handicapped. Churchill Livingstone, London.

Natriello, G. (1987). The impact of evaluation processes on students. Educational Psychologist, 22:155–175. doi:10.1207/s15326985ep2202_4.

Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies in Higher Education, 31(2):199–218. doi:10.1080/03075070600572090.

Rubin, D. (2008). Comment: The design and analysis of gold standard randomized experiments. Journal of the American Statistical Association, 103(484):1350–1353. doi:10.1198/016214508000001011.

Sadler, D. R. (1998). Formative assessment: revisiting the territory. Assessment in Education: Principles, Policy & Practice, 5(1):77–84. doi:10.1080/0969595980050104.

Shepard, L. A., Penuel, W. R., and Davidson, K. L. (2017). Design principles for new systems of assessment. Phi Kappan Kelta, 98(6):47–52. doi:10.1177/0031721717696478.

Smith, E. and Gorard, S. (2005). 'They don't give us our marks': The role of formative feedback in student progress. Assessment in Education: Principles, Policy & Practice, 12(1):21–38. doi:10.1080/0969594042000333896.

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., and Anders, J. (2018). Embedding formative assessment: Evaluation report and executive summary. Report, Education Endowment Foundation. Available from `https://educationendowmentfoundation.org.uk/public/files/EFA_evaluation_report.pdf`.

SSAT (2018). About SSAT. Retrieved from `https://www.ssatuk.co.uk/about/` on 22/10/2018.

Thompson, M. and Wiliam, D. (2008). Tight but loose: A conceptual framework for scaling up school reforms. In Wylie, E. C., editor, Tight but Loose: Scaling Up Teacher Professional Development in Diverse Contexts, chapter 1, pages 1–44. Educational Testing Service, Princeton, NJ.

Wiliam, D. (2011). What is assessment for learning? Studies in Educational Evaluation, 37:3–14. doi:10.1016/j.stueduc.2011.03.001.

Wiliam, D. (2017). Embedded Formative Assessment: Strategies for Classroom Assessment That Drives Student Engagement and Learning. Solution Tree, Bloomington, IN.

Wiliam, D. (2018). Assessment for learning: meeting the challenge of implementation. Assessment in Education: Principles, Policy & Practice, 25(6):682–685. doi:10.1080/0969594X.2017.1401526.

Wiliam, D., Lee, C., Harrison, C., and Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. Assessment in Education: Principles, Policy & Practice, 11(1):49–65. doi:10.1080/0969594042000208994.

Table 1: Balance of observable baseline characteristics between treatment and control groups and comparison to national characteristics

Variable	Intervention group		Control group		Difference	Total	England
School-level (categorical)	n/N(missing)	Percentage	n/N(missing)	Percentage	%pt.	Percentage	Percentage
Religiously affiliated	11 /70 (0)	15.71	12 /70 (0)	17.14	-1.43	16.43	18.73
Academy	53 /70 (0)	75.71	48 /70 (0)	68.57	7.14	72.14	64.83
Community School	10 /70 (0)	14.29	16 /70 (0)	22.86	-8.57	18.57	17.47
Voluntary or Foundation school	5 /70 (0)	7.14	3 /70 (0)	4.29	2.85	5.71	9.41
Voluntary aided school	2 /70 (0)	2.86	3 /70 (0)	4.29	-1.43	3.57	8.29
Ofsted: Outstanding	12 /57 (13)	21.05	12 /56 (14)	21.43	-0.38	21.24	N/A
Ofsted: Good	30 /57 (13)	52.63	32 /56 (14)	57.14	-4.51	54.87	N/A
Ofsted: Satisfactory	13 /57 (13)	22.81	12 /56 (14)	21.43	1.38	22.12	N/A
Ofsted: Inadequate	2 /57 (13)	3.51	0 /56 (14)	0.00	3.51	1.77	N/A
Non-TEEP	59 /70 (0)	84.29	66 /70 (0)	94.29	-10.00	89.29	N/A
School-level (continuous)	n(missing)	Mean (SD)	n(missing)	Mean (SD)	Std. Diff.	Mean (SD)	Mean (SD)
Number of pupils	69 (1)	1080.38 (362.10)	70 (0)	1123.67 (390.29)	-0.12	1102.18 (375.83)	938.96 (419.72)
% of Free School Meal	69 (1)	13.34 (9.64)	70 (0)	14.49 (8.73)	-0.13	13.92 (9.18)	11.24 (8.44)
% SEN with support	69 (1)	11.98 (6.49)	70 (0)	11.84 (5.94)	0.02	11.91 (6.19)	11.00 (N/A)
% SEN with statement	69 (1)	1.62 (1.40)	70 (0)	1.71 (1.22)	-0.07	1.66 (1.31)	1.7 (N/A)
% English Additional Language	69 (1)	19.25 (22.11)	70 (0)	18.84 (20.57)	0.02	19.04 (21.27)	15.24 (19.76)
School-level (continuous)	n(missing)	Median	n(missing)	Median	Diff.	Median	Median
Number of pupils	69 /70 (1)	1021	70 /70 (0)	1072	-51.00	1057	925
% of Free School Meal	69 /70 (1)	11.60	70 /70 (0)	14.35	-2.75	13.00	9
% SEN with support	69 /70 (1)	11.33	70 /70 (0)	11.34	-0.01	11.33	N/A
% SEN with statement	69 /70 (1)	1.26	70 /70 (0)	1.41	-0.15	1.38	N/A
% English Additional Language	69 /70 (1)	10.73	70 /70 (0)	9.89	0.84	10.00	6.5
Pupil-level (categorical)	n/N(missing)	Percentage	n/N(missing)	Percentage	%pt	Percentage	Percentage
Female	6596 /12600 (0)	52.35	6783 /13277 (0)	51.09	1.26	51.70	49.10
Ever FSM	3658 /12600 (0)	29.03	4048 /13277 (0)	30.49	-1.46	29.78	11.24
Ethnicity: Asian	1292 /12600 (0)	10.25	1463 /13277 (0)	11.02	-0.77	10.65	10.14
Ethnicity: Black	912 /12600 (0)	7.24	921 /13277 (0)	6.94	0.30	7.08	5.42
Ethnicity: Mixed	633 /12600 (0)	5.02	684 /13277 (0)	5.15	-0.13	5.09	4.59
Ethnicity: White	9373 /12600 (0)	74.39	9814 /13277 (0)	73.92	0.47	74.15	76.40
Pupil-level (continuous)	n(missing)	Mean (SD)	n(missing)	Mean (SD)	Std. Diff.	Mean (SD)	Mean (SD)
English points at KS2	11390 (1210)	73.45 (14.69)	11908 (1369)	72.91 (14.56)	0.04	73.17 (14.63)	N/A
Mathematics points at KS2	11534 (1066)	70.56 (19.82)	12084 (1193)	69.81 (19.99)	0.04	70.18 (19.91)	N/A
Pupil-level (continuous)	n(missing)	Median	n(missing)	Median	Diff.	Median	Median
English points at KS2	11390 /12600 (1210)	74	11908 /13277 (1369)	73	1	73	N/A
Mathematics points at KS2	11534 /12600 (1066)	74	12084 /13277 (1193)	73	1	73	N/A

Notes. Reporting sample sizes (n indicates sub-group size, N indicates overall sample size), missing values, means, standard deviations (SD), and medians. Difference column reports %pt. (percentage point) differences for categorical values, std. diff (standardised differences) for means, and unstandardised differences for medians. Final column reports national average characteristics based on published summary statistics from the UK Department for Education “Schools, pupils and their characteristics” Statistical First Release, where available and as applicable.

Table 2: Primary outcome analysis

	M0	M1	M2	M3
Treated	1.477	2.054	1.881	1.658
	(1.12)	(1.37)	(1.41)	(1.76)*
Prior attainment			0.897	0.893
			(17.20)***	(17.03)***
Blocks	No	No	No	Yes
Cohen's d	0.08	0.11	0.10	0.09
95%	-0.06	-0.05	-0.04	-0.01
CI	0.21	0.26	0.24	0.18
R^2	0.00	0.00	0.16	0.23
R^2_w		0.00	0.14	0.14
R^2_b		0.01	0.27	0.61
ρ		0.21	0.18	0.11
N_i	25,393	25,393	25,393	25,393
N_j		140	140	140

Notes. All models have GCSE Atainment 8 score as their dependent variable. M0 is an Ordinary Least Squares model, M1-M3 are hierarchical linear models incorporating school level random effects. Some or all of the following notes also apply to Tables 3, 4, 5, and 6: t statistics (calculated taking into account school-level clustering) in parentheses; stars indicate statistical significance as follows: * $p < 0.10$, ** $p < 0.05$, ***. $p < 0.01$. Prior attainment variable is average performance across English, mathematics and science in UK's Key Stage 2 (age 11) SATs national tests. Blocks indicates a vector of school-level stratification dummy variables used in the process of randomisation. Cohen's d effect size followed lower and upper 95% confidence intervals. R^2 reports overall variance explained by model; R^2_w reports within school variance explained by model; R^2_b reports between school variance explained by model. ρ reports intra-cluster correlation conditional on model covariates. N_i reports number of pupils in model; N_j reports number of schools in model.

Table 3: Secondary outcome results

	Attainment 8	English	Maths	Science	Humanities	Languages
Treated	1.658	0.0831	0.0985	0.114	0.182	0.248
	(1.76)*	(0.97)	(1.01)	(1.04)	(1.56)	(1.81)*
Prior attainment	0.893	0.0835	0.0927	0.0937	0.0910	-0.0126
	(17.03)***	(18.34)***	(15.51)***	(16.34)***	(14.11)***	(-2.37)**
Blocks	Yes	Yes	Yes	Yes	Yes	Yes
Cohen's d	0.09	0.05	0.05	0.05	0.07	0.09
95%	-0.01	-0.05	-0.05	-0.05	-0.02	-0.01
CI	0.18	0.14	0.14	0.15	0.15	0.18
R^2	0.23	0.20	0.19	0.22	0.19	0.05
R^2_w	0.14	0.12	0.13	0.13	0.10	0.00
R^2_b	0.61	0.54	0.54	0.59	0.58	0.24
ρ	0.11	0.10	0.10	0.12	0.12	0.15
N_i	25,393	24,538	24,515	24,689	19,657	12,497
N_j	140	140	140	140	140	140

Notes. Outcome measures indicated at top of table are as follows: "Attainment 8" is pupils' GCSE Attainment 8 score, calculated from the best 8 nationally recognised high-stakes examinations taken at age 16. "English" is specifically numerical grade on the GCSE English high-stakes examinations. "Maths" is specifically numerical grade on the GCSE mathematics high-stakes examination. "Science", "Humanities" and "Languages" are three components of the EBacc set of subjects. All models are hierarchical linear models incorporating school level random effects. See notes to Table 2 for further details on reporting.

Table 4: Sub-group analysis results - Pupil-level sub-groups

	Full	FSM	Low Attain.	Med. Attain.	High Attain.
Treated	1.658	1.309	1.532	1.243	0.256
	(1.76)*	(1.32)	(1.47)	(1.83)*	(0.38)
Prior attainment	0.893	1.029	-0.213	3.485	4.993
	(17.03)***	(17.00)***	(-6.30)***	(26.65)***	(39.27)***
Blocks	Yes	Yes	Yes	Yes	Yes
Cohen's d	0.09	0.07	0.08	0.07	0.01
95%	-0.01	-0.03	-0.03	-0.00	-0.06
CI	0.18	0.17	0.19	0.13	0.08
R^2	0.23	0.19	0.06	0.11	0.26
R^2_w	0.14	0.15	0.01	0.07	0.19
R^2_b	0.61	0.53	0.53	0.51	0.55
ρ	0.11	0.12	0.09	0.08	0.10
N_i	25,393	7,470	8,471	8,470	8,452
N_j	140	140	139	139	140

Notes. All models have GCSE Atainment 8 score as their dependent variable. Sub-groups indicated at top of table are as follows: "Full" is the full analysis sample (replication of M3 in Table 2; "FSM" is the sub-sample of pupils who have ever been eligible for Free School Meals, an administrative indicator of low income; "Low/Med./High Attain." are the bottom, middle and top tertiles of pupil-level prior attainment defined using Key Stage 2 (age 11) test performance; "Non-TEEP" are schools not identified as participating in a related programme also offered by the developers: the Teacher Effectiveness Enhancement Programme. See notes to Table 2 for further details on reporting.

Table 5: Sub-group analysis results - School-level sub-groups

	Full	Low Attain.	Med. Attain.	High Attain.	Non-TEEP
Treated	1.658	-0.379	0.995	4.457	2.135
	(1.76)*	(-0.29)	(0.93)	(1.94)*	(2.09)**
Prior attainment	0.893	0.802	1.102	0.753	0.896
	(17.03)***	(8.68)***	(13.05)***	(8.67)***	(16.28)***
Blocks	Yes	Yes	Yes	Yes	Yes
Cohen's d	0.09	-0.02	0.05	0.23	0.11
95%	-0.01	-0.16	-0.06	-0.00	0.01
CI	0.18	0.12	0.16	0.47	0.22
R^2	0.23	0.12	0.19	0.17	0.23
R^2_w	0.14	0.12	0.19	0.11	0.14
R^2_b	0.61	0.12	0.31	0.38	0.62
ρ	0.11	0.06	0.05	0.19	0.12
N_i	25,393	7,437	9,757	8,199	22,709
N_j	140	47	48	45	125

Notes. All models have GCSE Atainment 8 score as their dependent variable. Sub-groups indicated at top of table are as follows: "Full" is the full analysis sample (replication of M3 in Table 2; "FSM" is the sub-sample of pupils who have ever been eligible for Free School Meals, an administrative indicator of low income; "Sch. Low/Med./High Attain." are the bottom, middle and top tertiles of school-level average prior attainment defined using Key Stage 4 (age 16) test performance; "Non-TEEP" are schools not identified as participating in a related programme also offered by the developers: the Teacher Effectiveness Enhancement Programme. See notes to Table 2 for further details on reporting.

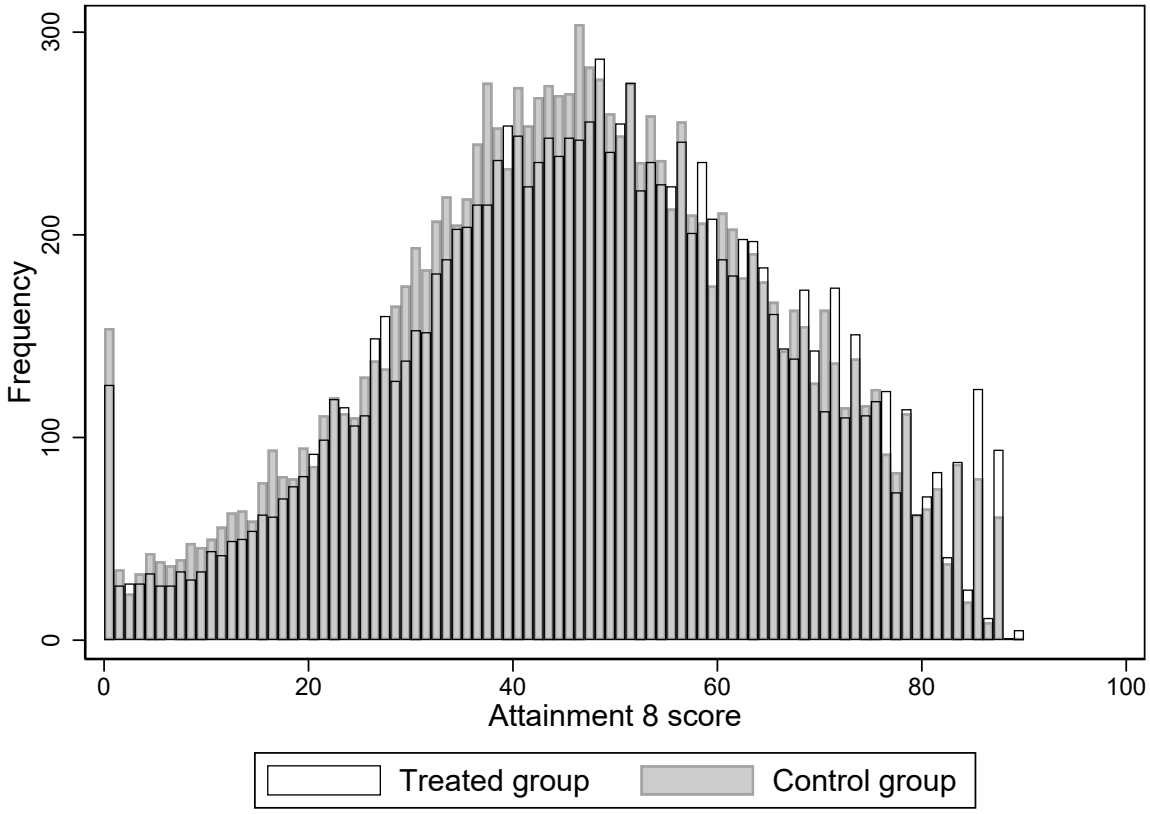
Table 6: Complier analysis results

	ITT	First Stage	IV	First Stage	IV
Treated	1.658	0.636		0.329	
	(1.76)*	(10.93)***		(5.65)***	
Prior attainment	0.893	-0.00163	0.892	0.000237	0.887
	(17.03)***	(-2.02)**	(15.51)***	(0.29)	(14.95)***
Minimal Compliance measure			2.165		
			(1.66)*		
Maximal Compliance measure					4.186
					(1.61)
Blocks	Yes	Yes	Yes	Yes	Yes
Cohen's d	0.09		0.11		0.22
95%	-0.01		-0.02		-0.05
CI	0.18		0.25		0.49
N_i	25,393	25,393	25,393	25,393	25,393

Notes. Models are as follows: 'ITT' is the full analysis sample (replication of M3 in Table 2; "Minimal Compliance" reports first stage and structural models for a two stage least squares estimation where treatment status instruments the binary minimal compliance measure (discussed in Section 2); "Maximal Compliance" reports first stage and structural models for a two stage least squares estimation where treatment status instruments the binary maximal compliance measure (discussed in Section 2). ITT and all structural models have GCSE Atainment 8 score as their dependent variable. See notes to Table 2 for further details on reporting.

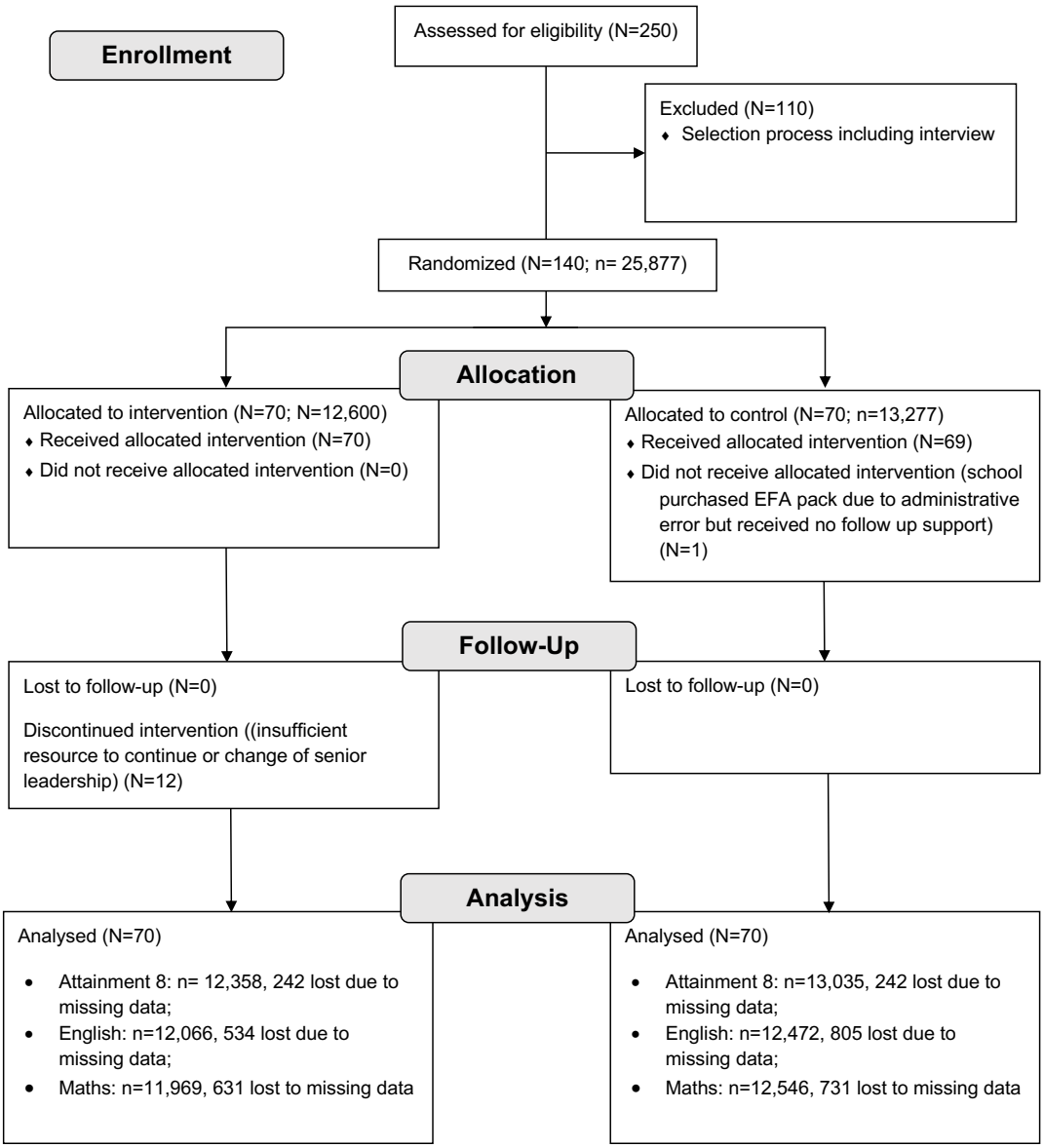
List of Figures

Figure 1: Histogram of GCSE Attainment 8 score



Notes. Overlapping histograms of GCSE Attainment 8 score for treatment and control groups.

Figure 2: CONSORT diagram



 @cepeo_ucl

ucl.ac.uk/ioe/cepeo