JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2022, VOL. 15, NO. 4, 748–779
https://doi.org/10.1080/19345747.2021.2018746

Routledge
Taylor & Francis Group

INTERVENTION, EVALUATION, AND POLICY STUDIES

OPEN ACCESS Check for updates

The Effect of Embedding Formative Assessment on Pupil Attainment

Jake Anders[a], Francesca Foliano[b], Matt Bursnall[c], Richard Dorsett[d], Nathan Hudson[e], Johnny Runge[f] and Stefan Speckesser[g]

[a]UCL Centre for Education Policy & Equalising Opportunities, University College London, London, UK; [b]UCL Social Research Institute, University College London, London, UK; [c]School of Health and Related Research, University of Sheffield, Sheffield, UK; [d]Centre for Employment Research, University of Westminster, London, UK; [e]NatCen Social Research, London, UK; [f]National Institute of Economic and Social Research, London, UK; [g]School of Business and Law, University of Brighton, Brighton, UK

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 study, we present the results of a randomized 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 program at baseline, and complier analysis both suggest 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.

ARTICLE HISTORY
Received 31 May 2020
Revised 15 October 2021
Accepted 28 October 2021

KEYWORDS
Embedding practice; formative assessment; professional development; pupil attainment; randomized controlled trial

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 & Wiliam, 2009).

CONTACT Jake Anders jake@jakeanders.uk UCL Centre for Education Policy & Equalising Opportunities, University College London, Gower Street, WC1E 6BT, London, UK.

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

"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 & 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 popularized by Black, Wiliam and colleagues including through books aimed at practitioners (Black et al., 2003; Black & Wiliam, 1998b; Wiliam, 2017). A substantial literature theorizing (Black & Wiliam, 2009, 2018), developing (Clark, 2015) and critiquing (Bennett, 2011) the approach continues to thrive.

Since high-quality feedback is at the heart of formative assessment (Nicol & Macfarlane-Dick, 2006; Sadler, 1998), it is heartening that preexisting reviews of the effectiveness of improving feedback have concluded that it does improve students' learning (Education Endowment Foundation, 2018; Hattie & Timperley, 2007; Wisniewski et al., 2019). However, this is not without caveats. The Education Endowment Foundation (EEF) review of the evidence on its use notes that "[m]any of the studies included are small scale studies from psychology which demonstrate theoretical principles, but which may be difficult to generalize to educational practice" (Education Endowment Foundation, 2018). Furthermore, Kluger and DeNisi (1996) highlight the risks of negative effects from poorly executed feedback interventions where these have the effect of directing pupil attention to the self rather than to task-motivation and task-learning processes.

In a critical review of the evidence base for formative assessment, which serves as a useful framing for the contributions of the present study, Bennett (2011) raises the concern that, in general, formative assessment is defined too vaguely to describe a consistent set of practices, leading to the potential for differing effects as implementations vary within this definition (noting the similarity to the concerns of Kluger & DeNisi, 1996). Bennett argues that a meaningful definition requires "a theory of action and one or more concrete instantiations" (Bennett, 2011, p. 19), and goes on to consider the Keeping Learning on Track (KLT) program as an example of one such concrete instantiation. The intervention studied in this article (the "Embedding Formative Assessment" program marketed in England by the Schools, Students And Teachers' network (SSAT); details discussed in section 2) shares many features (and a developer) with the KLT program. Based on this, we argue that we are studying a well-defined instantiation and, moreover, a similar (albeit not identical) one to that considered by Bennett—with the strengths and weaknesses that implies.

In particular, Bennett (2011) argues that "rooting formative assessment in pedagogical skills alone is probably insufficient" (Bennett, 2011, p.20), and instead it would be more appropriate to develop domain-specific approaches to embedding formative assessment within particular curricula. A similar point is made by Coffey et al. (2011) in the language of disciplinary substance, framing such domain-general approaches as "strategy-based" and, hence, missing opportunities for engagement with disciplinary substance. Hodgen and Marshall (2005) provide particular examples of differing approaches to formative assessment in the contrasting subjects of math and English—although they also

stress the commonalities. While we see the merits of this argument, we can also see the considerable attraction to schools of a domain-general approach to embedding formative assessment given the possibilities this raises of whole-school professional development rather than attempting to identify and implement multiple approaches on a subject-by-subject basis, that is scalability (Wiliam, 2019). Furthermore, others have argued that domain-general/strategy-based approaches are where the strength of the existing evidence of the effectiveness of formative assessment lies (Shepard et al., 2017; Wiliam, 2018). We think it reasonable to research the effect of this explicitly domain-general approach, while mindful of the arguments of the potential for larger effects through tailored use of domain-specific interventions.

In some ways underlying the above two concerns, Bennett also questioned whether the magnitude of estimated effects of improved formative assessment from existing studies are reasonable. A similar concern is raised by Kingston and Nash (2011), who 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 on the effectiveness of formative assessment is lacking in some respects: it is largely based on relatively small studies with committed teachers, supported by the close involvement of a team of researchers and recognized experts in the field (e.g., Andersson & 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 & Palm, 2018).

Many features of these studies suggest 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 & Gorard, 2005), perhaps as scaling-up increases the risks of teachers providing feedback in ways that Kluger and DeNisi (1996) identify as less effective.

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 & Wiliam, 1998a, etc.), and Wiliam and Leahy's experiences with implementing formative assessment programs (Leahy & 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. 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 randomized controlled trial. The approach we follow is carefully chosen to minimize 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 minimize 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 labor 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 2 years of exposure to the Embedding Formative Assessment program 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 2 years of exposure to the Embedding Formative Assessment program 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 2 years of exposure to the Embedding Formative Assessment program 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.

The Intervention

This research attempts to estimate the impact of the introduction of the "Embedding Formative Assessment" (EFA) teacher professional development program (Leahy & Wiliam, 2013) into a school on pupils' attainment. Dylan Wiliam and Siobhan Leahy designed the intervention and the program materials, and it is delivered by the SSAT, an independent membership organization of schools whose "professional development and school improvement programs help leaders and teachers to further outcomes for all

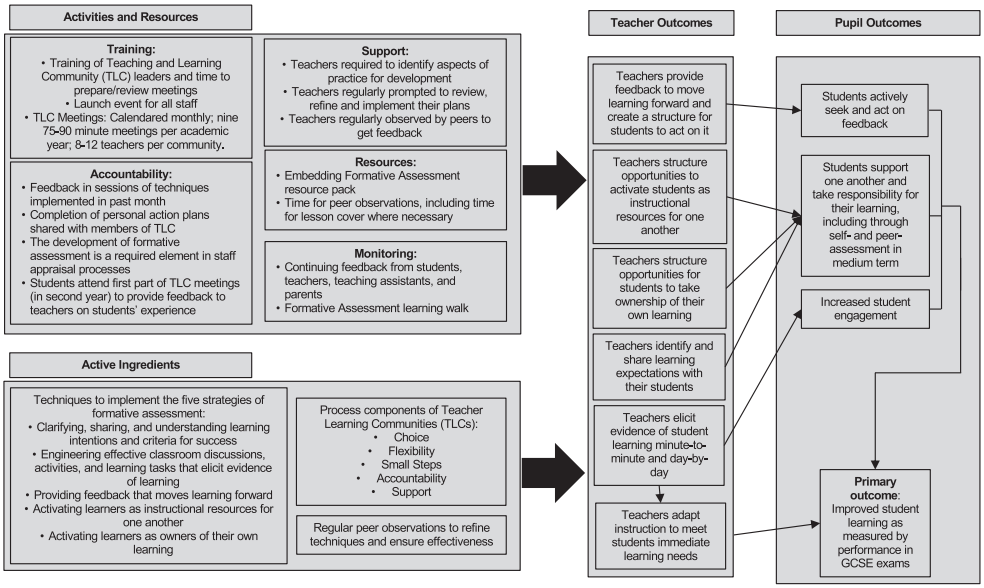


Figure 1. Theory of change. *Notes*. Draws on theories of change for this intervention co-developed with the Education Endowment Foundation (research funder) and SSAT (delivery partner), as well as the KLT theory of change reported by Bennett (2011).

young people, and develop leadership at all levels across the system" (SSAT, 2018). EFA and similar programs (for example the "Keeping Learning on Track" (KLT) marketed in the US[1]) are used in countries across the world, including Scotland (where it has been used by 28 out of 32 local authority districts), England (where the Education Endowment Foundation is funding scale-up work based on the evidence of this trial, and feature EFA prominently in a recent guidance reports, which are heavily promoted to teachers and school leaders, Education Endowment Foundation, 2021b), Australia, and Singapore (Leong & Tan, 2014). Moreover, EFA is a professional development program focused on improving formative assessment—the importance of which to learning is little questioned (Education Endowment Foundation, 2018; Hattie & Timperley, 2007; Wisniewski et al., 2019)—using a model based around the highly popular approach of professional learning communities (Vescio et al., 2008).

EFA is designed to be an ongoing activity within a school. However, for the purposes of this evaluation, it was introduced to participating schools to be carried out for a minimum of 2 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 program consists of nine monthly Teacher Learning Communities (TLCs) workshops across each academic year and monthly peer observations. The intervention's theory of change is reported in Figure 1.

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

[1]Keeping Learning on Track (KLT) was originally marketed in the US by the Educational Testing Service (ETS), and latterly by Northwest Evaluation Association (NWEA).

intervention of the kind that it has been highlighted are likely to be needed to improve practice (Anders et al., 2017; Leithwood et al., 2006). This also concords with one of the developer's previous work on the need for interventions to be "tight but loose" (Thompson & Wiliam, 2008) if they are to be successfully scaled up in diverse contexts. 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). TLCs can be characterized as a form of professional learning community (Thompson et al., 2004), which have become increasingly popular as a model of delivery for teacher professional development (Vescio et al., 2008). Stoll et al. (2006) define a professional learning community as "a group of people sharing and critically interrogating their practice in an ongoing, reflective, collaborative, inclusive, learning-oriented, growth-promoting way […] operating as a collective enterprise" (Stoll et al., 2006, p. 223) and, while the EFA TLCs arguably include additional features, they certainly fit this characterization. In a review of how professional development improves teaching, Kennedy (2016) notes the seeming importance of "collective participation" (also noted in an earlier review by Cordingley et al. (2005)) as a key feature of effective teacher professional development programs, often in the form of professional learning communities.

However, Kennedy (2016) also notes that professional learning communities are no guarantor of success in terms of positive impacts on pupil attainment, arguing that "we need to move past the concept of learning communities per se and begin examining the content such groups discuss and the nature of intellectual work they are engaged in" (Kennedy, 2016, p. 972). Indeed, evaluations of interventions with certain similarities in comparable contexts have not yielded positive impacts on pupil attainment, including the popular Lesson Study approach focused on teacher peer-to-peer observation and feedback (Murphy et al., 2015), and "research learning communities" in which the substantive content varied between participating schools (Rose et al., 2017). It seems important, in light of such findings, that these factors are considered with the EFA TLC models. Specifically, each TLC workshop involves a group of teachers reporting on their use of techniques since the last workshop, sharing new formative assessment ideas to try, and personal action planning for the coming month. 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 min and follows a similar pattern:

- Introduction including the learning intentions for the session (5 min);
- A starter activity (5 min);
- Feedback from all teachers on techniques they have attempted since last session (25 min);
- Formative assessment content (20 min);
- Action planning (15 min); and
- Summary (5 min).

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 min at the start, middle, or end of a lesson. Pairs will then need to find 15 min 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 (the intervention is primarily a "strategy-based" approach to use of formative assessment, as discussed earlier). 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. Thus, implementation within the classroom varied substantially across schools and teachers, but this is by design.

Within each school, a lead teacher was responsible for implementing the program and appointed the required number of teachers to lead/facilitate each monthly TLC group. The main support mechanism was the resource pack that included all materials for the TLCs, including agendas, leader's agendas and handouts. The lead teacher attended an initial training day run by one of the pack developers, Dylan Wiliam. While there is always an initial training day provided to schools as part of EFA, this particular launch day involving Dylan Wiliam was specific to this evaluation and is not routinely provided to schools purchasing the EFA pack. We cannot disentangle the effect of the single day workshop from the rest of the program, but doubt that a single day could be driving the effects of the intervention.

Finally, lead teachers received ongoing support from a designated SSAT Lead Practitioner. 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 structured sequence of face-to-face meetings with lead teachers at each school at the start of the project and at the end of

the first year. The SSAT Lead Practitioner was also available to be contacted by lead teachers on phone and email throughout the initial 2-year program. Additionally, schools had access to an online forum to share resources. None of this was a deviation from support regularly provided as part of EFA in a non-trial context.

Optimal treatment fidelity was emphasized 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 1 h. In addition, the materials emphasize 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.

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). These measures of attainment are externally validated and widely recognized. 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 of the trial.

Data have been obtained for this analysis directly from the National Pupil Database (NPD) held by the UK Department for Education (DfE). As a result, test scores are available for the vast majority of our sample. These scores were requested for pupils who are in Year 10 (age 15) in participating schools at the start of the intervention, with the 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 randomization, meaning that it is unlikely to 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), measured at the end of the second year of implementation (i.e., end of academic year 2016/17). Attainment 8 is widely recognized and a metric in which schools take a keen interest, since it is one of England's main accountability measures for secondary schools. 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 math.[2] 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 2.

[2]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.

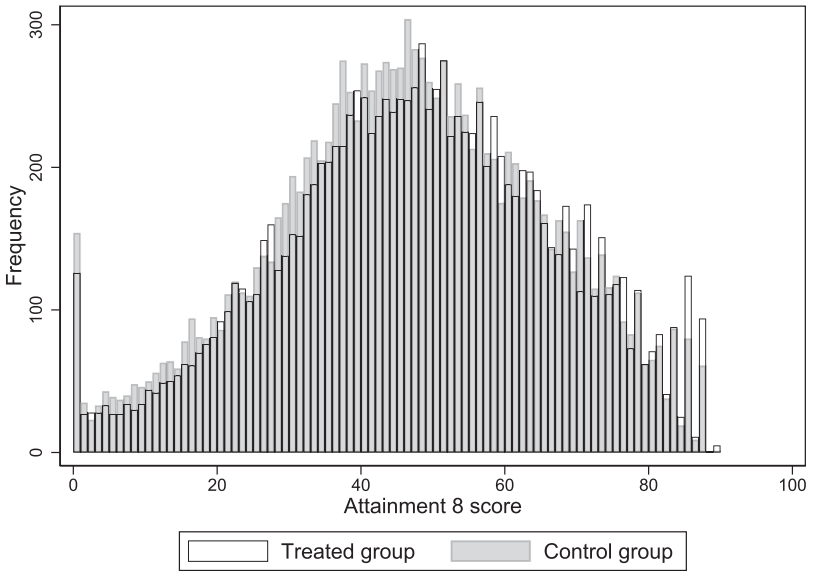


Figure 2. Histogram of primary outcome measure: GCSE Attainment 8 score. *Notes*. Overlapping histograms of the primary outcome measure (GCSE Attainment 8 score) for treatment and control groups. Data from national examinations sat at end of 2016/2017 academic year.

The two pre-registered secondary outcome measures were students' numerical grades for mathematics and English,[3] which again range from 0 to 9. Performance in English and math are a core part of another of England's accountability measures: the English Baccalaureate (often shortened to EBacc). 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 outcomes such as university attendance (Anders, 2012; DfE, 2013) and labor market outcomes (McIntosh, 2006). In particular, there may be a particular benefit from performance in English and math (Dolton & Vignoles, 2002). In this paper, we also conduct exploratory analysis of pupils' performance in science, humanities and languages[4] to provide additional context to our overall findings.

Prior attainment of the same pupils when aged 11 (i.e., at the end of the 2011/12 academic year for the relevant cohort of pupils) were also obtained. This was measured using national examinations taken at the end of primary schooling.[5] These baseline data 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 & Sutch, 2014), something borne out in the analysis below. Pupils

[3]Specifically, we use NPD variables KS4_APMAT_PTQ_EE for math performance and KS4_APENG_PTQ_EE for English performance.

[4]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.

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

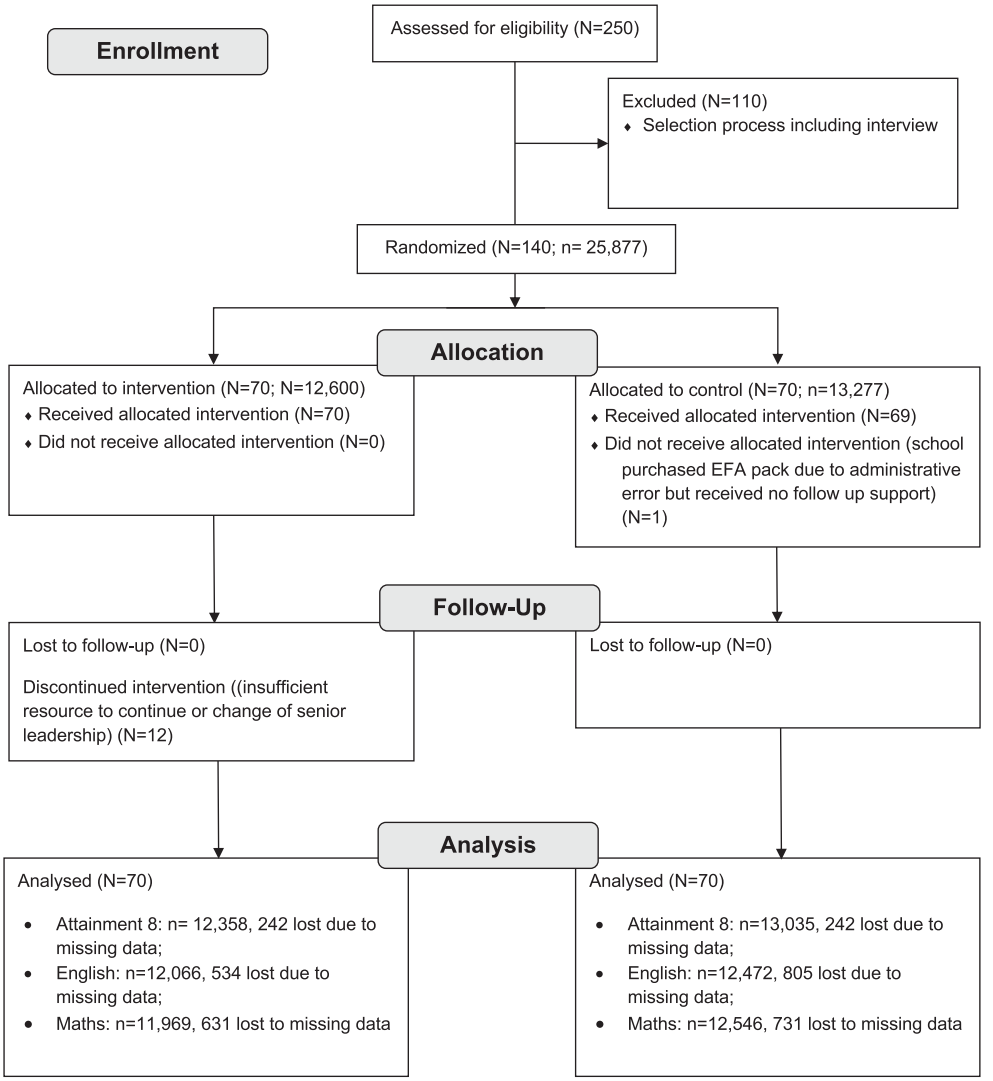


Figure 3. CONSORT diagram.

without prior attainment (less than 2% of the total sample) are excluded from all modeling to ensure consistency of the composition of the sample across models whether or not this is included as a covariate.

Missing pupil data is almost exclusively due to issues in achieving a successful link between our data and administrative datasets. These may be, for example, due to typographic errors in the "Unique Pupil Number" official administrative ID shared with us by schools, or pretest data being missing due to the child not being in a publicly funded school during their primary/elementary education (Anders et al., 2020). Even all taken together, these are an extremely small number: 484 pupils out of 25,877 (under 2%) have to be removed from the primary outcome analysis due to any kind of missing data.

We report the trial's CONSORT diagram in Figure 3, demonstrating the flow of schools and pupils through the trial. 250 schools showed initial interest in response to

recruitment efforts—this was recorded no matter how fleeting this interest was—while 140 were ultimately included in the randomization sample. While there were exclusions due to schools falling outside published eligibility requirements (e.g., outside England; not being a publicly funded school) and a small number were also excluded due to assessment through a telephone interview that they were already using the intervention or there were clear barriers to implementation,[6] the vast majority of drop-out at this point was due to schools disengaging from the recruitment processes. These involved some burdens on the school, notably circulation of ethics documentation to participants and parents and secure transfer of pupil identifiers to the research team, which were deliberately front-loaded in order to minimize burden and, hence, drop-out post-randomization—this represents a tradeoff between internal and external validity. As the vast majority of this pre-randomization attrition was driven by school disengagement, it was difficult to capture systematic data on reasons beyond this. After randomization there were 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.

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

- "Minimal" compliance, which is simply measured by an indicator of whether the school was still at all engaged with the program by the end of the 2 years. It should perhaps be interpreted as a lower bound on compliant implementation. 58 of the 70 treatment schools fall into this category.
- "Maximal" compliance, which is 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.

Furthermore, we note that some concerns were raised about this latter 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, given the light-touch nature of the intervention, 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. Mis-measurement of this type is likely to result in upward biased impact estimates (Jiang & Ding, 2020). Consequently, interpretation of analysis based on this "maximal" compliance measure would be unclear and, therefore we concentrate on the "minimal" compliance measure alone. Ultimately, we feel that not having more granular and informative compliance measures is an

[6]An example of such a barrier would be an imminent change of school leadership which prevent assurances as to the continued support from school leadership for program implementation.

important limitation of this evaluation and one that future work could helpfully address to improve our understanding of how this intervention works.

Design and Analysis

Evaluation Design

In this paper, we estimate the effect of a school participating in the "Embedding Formative Assessment" program using a randomized controlled trial (RCT). As the intervention is inherently whole school in nature, it is not possible to randomize the treatment within schools (for example, to half the teachers in a school). Instead, we selected a two-armed blocked/stratified school-level cluster randomized controlled trial (cRCT).

Blocking/stratification was undertaken to minimize 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 criterion. 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.

A power calculation was conducted to estimate the sample size required to achieve a Minimum Detectable Effect Size (MDES) of 0.20 (chosen in line with research funder policy) for a statistical test at the 0.05 level of significance with 0.8 power. The calculation was also based on the following assumptions:

- An expected average of 100 students in Year 10 at each participating school at the start of trial—this was based on figures obtained from the UK Department for Education as part of a statistical release (DfE, 2015) with a conservative adjustment;
- A within-school pretest to post-test correlation of 0.66 and a between-schools pretest to post-test correlation of 0.57—this was based on our own preliminary analysis of administrative data to estimate these correlations in the English schooling system;
- And an 0.20 intra-cluster correlation in the outcome measure—this was based on prior analysis of administrative data to estimate this parameter for the relevant outcome measure in the English schooling system (Education Endowment Foundation, 2015).

These calculations suggested that recruitment of 120 schools would meet this requirement. Ultimately, the project team was successful in recruiting 140 schools, which, along with a larger number of pupils per school, 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 ($459) at the start of the trial (September 2015),[7] which was equivalent to the purchase price of 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. It is also the case that, while the delivery partner prevented control schools from specifically buying the intervention during the delivery period and made efforts to avoid recruitment of schools using the intervention, the process evaluation identified that some treatment and control schools may have accessed the pack prior to the intervention. We discuss the implications of this below.

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 randomized to treatment or control; each subsequent school was then assigned to the opposite outcome of the previous school. Since this randomization 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.

Balance and Representativeness

Randomization 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-randomization selection effects (such as nonrandom attrition). While our design aims to minimize such possibilities, through use of blocking on key characteristics in randomization to reduce imbalance and the use of administrative data to avoid missing outcome measurement (only 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 unstandardized median differences and standardized mean differences (Imbens & Rubin, 2015). The standardized difference is calculated as the unstandardized difference between the mean of the characteristic in each group divided by the overall sample standard deviation, as follows:

[7]Unlike monetary amounts reported later in the report, these figures are not adjusted for inflation and reflect exchange rates from September 2015, rather than at time of writing. This is to report the actual amount paid to schools allocated to the control group.

$$\delta = \frac{\mu_{\text{Treat}} - \mu_{\text{Control}}}{\sigma_{\text{Sample}}} \quad (1)$$

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.

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:

$$y_{ij} = \alpha + \beta_1 \text{Treat}_j + \beta_2 \text{KS2}_{ij} + \mathbf{Block}_j + \gamma_j + \varepsilon_{ij} \quad (2)$$

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 randomization 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 randomization 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:

$$\delta = \frac{\beta_1}{\sigma_{pooled}} \quad (3)$$

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:

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

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.

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 of 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 of 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. Most of these sub-groups are defined on a pupil-level basis, while one is defined on a school-level basis.

The pupil-level sub-group analyses are motivated by an interest in the differential effects of formative assessment, given its original intent as a way of reducing the variation in performance within classrooms (Bloom, 1968; Guskey, 2007). The 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.

We also carry out a school-level sub-group analysis. This was not originally registered in the evaluation protocol—although it was registered in the statistical analysis plan prior to outcome data becoming available—so should be considered exploratory. It is

essentially a robustness check, based on a finding from the process evaluation that previous or current involvement in the Teacher Effectiveness Enhancement Programme (TEEP)[8] strongly influenced delivery and experiences of delivery of the EFA intervention. Specifically, we exclude schools subsequently identified by SSAT from their administrative records as also participating in TEEP.

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

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

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 modeling. In other respects, the specification remains the same as the primary outcome ITT model. The second stage model is specified as follows:

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

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 estimate of interest is β_1, which recovers the effect of the intervention among compliers. Standard errors are clustered at the school level and adjusted due to the instrumental variables approach.

Process Evaluation

A process evaluation was also carried out as part of this research, primarily aiming to explore fidelity to the intervention design to provide a better understanding of this variation and reasons for adaptations. This element is not the primary focus of this article. We aim only to set out basic details that set the scene for reporting relevant insights that contextualize and enrich the findings of the quantitative impact evaluation. The process evaluation included the following elements of data collection:

- 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.
- Ten case study treatment schools were selected to include a variety of delivery contexts (including variation in schools' proportion of low-income pupils, geographical location, and rating from "Ofsted," which is England's schools inspectorate).
- Visits to the case study treatment schools were carried out 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 school's headteacher (principal).

[8]The Teacher Effectiveness Enhancement Programme (TEEP) is another school-wide professional development program offered by SSAT, who deliver the Embedding Formative Assessment program (EFA) in England.

- 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% of all treatment schools, or 69% of schools that finished the program. The control survey was completed by 39 schools, equivalent to 57% of control schools.

Qualitative data gathered from these activities were analyzed in NVivo using a "framework approach" (Spencer et al., 2014a, 2014b),[9] coding the data into themes and issues. We stress that we do not claim that the sample of case study schools is representative. The qualitative findings based on these aim to provide insights about the range and diversity of views and experiences of participants, rather than the views of a wider population.

Integrated into the process evaluation was an expenditure evaluation. This was designed following the Education Endowment Foundation's guidance (Education Endowment Foundation, 2019),[10] which draws on Levin's widely used "ingredients" method (Levin et al., 2018) with some specific additional requirements including, for example, the guidance to separate out financial expenditure and teacher time, without putting a financial value on the latter. Financial expenditure information came primarily from the delivery partners, who provided information about the price of the resource pack, the expenditure associated with arranging the training day and end of project event, as well as for providing support from Lead Practitioners including their expenses for visiting schools. However, this was supplemented by specific expenditure-related questions during visits to treatment schools, including asking Lead Teachers, headteachers, and teaching staff about any additional financial and time commitments associated with the intervention. As expenditure data was collected during the academic year 2016–2017 it has been adjusted using the UK Consumer Price Index to report in 2021 British pounds sterling (1GBP in February 2016 = 1.098 in August 2021), and converted to US dollars using the prevailing exchange rate on 1 August 2021 (1GBP = 1.39USD).

Results

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[11] and schools with Ofsted[12] ratings of "good"

[9]We note the reticence of the authors of the cited text about people referring to the tradition of qualitative data analysis they describe as the 'framework approach', given the diversity of approaches it brings together within a tradition of pragmatism and eclecticism (Ormston et al., 2014). Nevertheless, this description is widely used and seems the most appropriate way to communicate this succinctly.

[10]We cite a more recent version of this guidance since the earlier version is no longer available online.

[11]Similar to charter schools, academies have higher levels of autonomy from their local authority/school district.

[12]Ofsted is a government organization that inspects publicly funded schools in England.

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

Variable School-level (categorical)	Intervention group *n/N*(missing)	Percentage	Control group *n/N*(missing)	Percentage	Difference %pt.	Total Percentage	England 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)
% Special Educational Needs with support	69 (1)	11.98 (6.49)	70 (0)	11.84 (5.94)	0.02	11.91 (6.19)	11.00 (N/A)
% Special Educational Needs 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 as an 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
% Special Educational Needs with support	69/70 (1)	11.33	70/70 (0)	11.34	−0.01	11.33	N/A
% Special Educational Needs with statement	69/70 (1)	1.26	70/70 (0)	1.41	−0.15	1.38	N/A
% English as an 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	11,390/12600 (1210)	74	11,908/13277 (1369)	73	1	73	N/A
Mathematics points at KS2	11,534/12600 (1066)	74	12,084/13277 (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 (standardized differences) for means, and unstandardized differences for medians. Administrative data collected by the UK Department for Education in the academic year before intervention delivery began (i.e., 2014–2015), except for pupil-level prior attainment which is from 2011 to 2012 when the participant cohort sat relevant tests. School-level missing data results from challenges in linkage to administrative datasets, e.g., changes to school identifiers. 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; prior attainment variable not released in these published statistics.

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 generalizable 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 generalize our findings.

Primary Analysis

In this section, we report the outcomes of our pre-registered primary outcome analysis models, contextualized with related models. These are reported in Table 2, with the pre-registered primary analysis model reported as M3. The treatment effect converted into a Cohen's d effect size[13] is reported at the base of the regression table.

Although the primary analysis model for this evaluation is pre-specified (as discussed above), it is helpful to contextualize 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

[13]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.

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% CI	−0.06, 0.21	−0.05, 0.26	−0.04, 0.24	−0.01, 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 Attainment 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 3–6: *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) national curriculum tests. Blocks indicates a vector of school-level stratification dummy variables used in the process of randomization. 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.

of our estimates by controlling for unobserved school-level differences in pupils' performance, while making 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 randomization. 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 randomization. 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

Table 3. Secondary outcome results.

	Attainment 8	English	Math	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% CI	−0.01, 0.18	−0.05, 0.14	−0.05, 0.14	−0.05, 0.15	−0.02, 0.15	−0.01, 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 recognized high-stakes examinations taken at age 16. "English" is specifically numerical grade on the GCSE English high-stakes examinations. "Math" is specifically numerical grade on the GCSE mathematics high-stakes examination. "Science," "Humanities" and "Languages" are three components of the DfE's English Baccalaureate (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.

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 program to 140 schools across England.

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.

We begin with secondary outcome measures (Table 3). 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 DfE's "English Baccalaureate" performance measure:[14] Science, Humanities and Languages. Unlike English and mathematics, these subjects are not compulsory and, so, 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

[14]The "English Baccalaureate" or "EBacc" is a set of subjects that the DfE (2019) argues "keeps young people's options open for further study and future careers"

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% CI	−0.01, 0.18	−0.03, 0.17	−0.03, 0.19	−0.00, 0.13	−0.06, 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 Attainment 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. See notes to Table 2 for further details on reporting.

Table 5. Sub-group analysis results—school-level sub-groups.

	Full	Non-TEEP
Treated	1.658	2.135
	(1.76)*	(2.09)**
Prior attainment	0.893	0.896
	(17.03)***	(16.28)***
Blocks	Yes	Yes
Cohen's d	0.09	0.11
95% CI	−0.01, 0.18	0.01, 0.22
R^2	0.23	0.23
R^2_w	0.14	0.14
R^2_b	0.61	0.62
ρ	0.11	0.12
N_i	25,393	22,709
N_j	140	125

Notes. All models have GCSE Attainment 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); "Non-TEEP" are schools not identified as participating in a related program also offered by the developers: the Teacher Effectiveness Enhancement Programme. See notes to Table 2 for further details on reporting.

than those evident in English and math, suggesting that the overall performance improvements were particularly driven by changes in these subjects.

Next, we consider differential impacts among specific sub-groups. Most of these are pupil-level sub-groups (reported in Table 4), while one is defined at the school-level (reported in Table 5). As noted above, the school-level sub-group was not pre-specified in the evaluation protocol and should be considered exploratory; however, it was specified in the statistical analysis plan, based on findings from the process evaluation that previous or current involvement in TEEP strongly influenced schools' experiences of delivery.

We first estimate the effect among the sample of pupils 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

Table 6. Complier analysis results.

	ITT	First Stage	IV
Treated	1.658	0.636	
	(1.76)*	(10.93)***	
Prior attainment	0.893	−0.00163	0.892
	(17.03)***	(−2.02)**	(15.51)***
Minimal Compliance measure			2.165
			(1.66)*
Blocks	Yes	Yes	Yes
Cohen's d	0.09		0.11
95% CI	−0.01, 0.18		−0.02, 0.25
N_i	25,393	25,393	25,393

Notes. Models are as follows: 'ITT' is the full analysis sample (replication of M3 in Table 2; "First Stage" and "IV" report 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). ITT and structural models have GCSE Attainment 8 score as their dependent variable. See notes to Table 2 for further details on reporting.

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 effect 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[15] 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 to exploratory analysis of a school-level sub-group, specifically restricting our analysis to those participating schools 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 potentially diluting EFA's impact. Once we exclude those who were already participating in TEEP, we find a larger effect size of EFA of 0.11.

We turn finally to analysis of effects among those who were deemed to have shown at least "minimal" compliance with the program. The results are reported in Table 6,

[15]As stratification for randomization 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.

with the first column reporting the primary analysis Intention to Treat estimate as a point of comparison. The next column reports the First Stage estimates and the final column the IV treatment estimate. 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. Given the low bar for schools to be considered "minimally compliant" (as noted above) this should be interpreted as a lower bound of the effect of the program when implemented as planned.

Process Evaluation

The process evaluation provided several insights that are important to the interpretation of the impact evaluation findings. Overall, the case studies found a high level of variation in how schools implemented the intervention, which some would see as consistent with some of the concerns about the definition of formative assessment expressed by Bennett (2011), while others would argue that this is to be expected within the "tight but loose" framework (Thompson & Wiliam, 2008). While schools generally achieved the broader aim of facilitating dialogue and reflection, sharing of practices, and trialing of formative assessment techniques through the use of monthly workshops, implementation of the program varied significantly.

Most case study schools had made adaptations to the program, 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. Schools that were considered to achieve high implementation fidelity typically organized the TLCs according to the guidelines, using the resource pack materials. They also managed to facilitate the trialing of formative assessment techniques, as well as effective dialogue and sharing of these practices. Those schools that were considered to achieve low fidelity had adapted the TLC structures and content substantially; crucially, these adaptations, sometimes over time, changed the focus of the program, sometimes away from trialing, reflecting on, and sharing of formative assessment practices, to a specific school focus such as a certain marking policy. After the intervention was completed, SSAT acknowledged that it should have made it more explicit to schools exactly what changes and adaptations were permitted as part of the program; they provided a 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.

It was found that the TLC workshop format was often seen as the key element of the EFA program. 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

interviewees to report that the EFA program had had a positive impact on the school culture, by increasing dialogue between teachers including outside TLC sessions.

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 2-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 program 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. On the one hand, this may mean that formative assessment practices were not sufficiently different compared to existing practices or those of control schools, which may explain the smaller than expected effect size. On the other hand, the help to refocus staff attention on formative assessment practices often led to substantial changes in practices and dialogue. Lead teachers found it inspirational to meet and work with Dylan William, which potentially led to higher buy-in.

Teachers generally felt that the real benefits of the program 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 behavior, 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 2-year intervention, show a minimum effect, and that future studies should explore the effect a number of years after the exposure to the intervention.

The expenditure required to deliver this whole-school intervention as in this trial is estimated at around £4,277 ($5,945) over the 2 years of the intervention, or around £2,141 ($2,976) per year per school over the 2 years. Based on an average of 1,086 pupils per school in the treatment arm, this is equivalent to an estimated average expenditure of £1.97 ($2.74) per pupil per year. This covers several components, including expenditure on the SSAT resource package (£324; $450), expenses associated with attendance at training days (such as travel) (£384; $534) and support from SSAT Lead Practitioners during the 2 years (£3,569; $4,961). Schools also need to consider the staff time required to engage in the intervention: teaching staff were required to commit around 2 h each month to attend TLCs, and to carry out peer observations and feedback. However, the interview findings suggested that teachers did not spend any significant additional time on the intervention. They often emphasized that the EFA intervention did not place undue additional pressures on them, because it enhanced, rather than added to, their usual professional development plans and formative assessment practices.

The end of study survey carried out in control schools identified an issue of contamination of the control group. These schools were asked whether they had used any materials/resources, or participated in any interventions, aimed at improving assessment or feedback during the period of the project; 5 out of 39 schools that responded to the

control group survey (approx. 13%) reported having used Dylan Wiliam's Embedding Formative Assessment resources.

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 randomized controlled trial and the approach that we followed throughout was carefully chosen to minimize 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 minimize 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. 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, scalable program analyzing 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 program ("TEEP"), 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 analysis, both suggest that effects may be stronger 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 process evaluation carried out as part of this research project adds important context to our findings. It found that schools and teachers valued the program, most particularly highlighting that the monthly TLC sessions facilitated beneficial dialogue between staff, and the sustained 2-year focus on formative assessment helped refocus staff attention on applying and embedding already-existing good formative assessment practices. In addition, interviews with teachers suggested that they found younger pupils to be more receptive to changes in practice resulting from 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 analyze as part of this study.

The survey of control schools identified an issue with contamination of this group, with 13% of those who responded to this survey reporting having used Embedding

Formative Assessment resources. This does not mean that these schools were fully implementing the intervention in the way that schools allocated to treatment were. Furthermore, this kind of contamination (control schools accessing treatment) would be expected to attenuate our impact estimate and, so, we do not view it as undermining our core finding. If anything, the risk is that our estimate of the impact of embedding formative assessment into teacher practice is conservative.

Our pre-registered pupil-level sub-group analyses are not especially encouraging for the possibility 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, as would be consistent with the original intent of formative assessment to narrow the attainment distribution in a classroom (Bloom, 1968), although the difference in estimates between our prior attainment sub-groups are not statistically significant.

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 program 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—and welcome the subsequent "scale up" evaluation commissioned by the Education Endowment Foundation (2021a)—to continue to build the evidence on the conditions required for it to make the biggest differences to pupil performance.

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 provide specific evidence on why this might be. We do, however, note the possibility that these differences between subjects reflect differential importance of domain-specific (curricular) vs domain-general (strategy-based) formative assessment practices, perhaps suggesting a more complex picture in terms of understanding the interplay of these (Bennett, 2011; Wiliam, 2019). Further research specifically designed to test such differences would help to unpack our findings and, more importantly, provide important insights on how formative assessment can best be improved across curricula and at scale.

Acknowledgments

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.

Declaration of interest statement

No potential conflict of interest was reported by the author(s).

Funding

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.

ORCID

Jake Anders http://orcid.org/0000-0003-0930-2884
Francesca Foliano http://orcid.org/0000-0003-0145-3434
Richard Dorsett http://orcid.org/0000-0002-4180-8685
Stefan Speckesser http://orcid.org/0000-0002-2442-7194

Data availability statement

Data analyzed as part of this project have 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. https://doi.org/10.1111/j.1475-5890.2012.00158.x

Anders, J. (2016). *Embedding formative assessment: Evaluation protocol*. Report, Education Endowment Foundation. 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., & Allen, R. (2017). *Evaluation of complex whole-school interventions: Methodological and practical considerations*. A Report for the Education Endowment Foundation, Education Endowment Foundation.

Anders, J., Green, F., Henderson, M., & Henseke, G. (2020). Determinants of private school participation: All about the money? *British Educational Research Journal*, *46*(5), 967–992. https://doi.org/10.1002/berj.3608

Andersson, C., & 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. https://doi.org/10.1016/j.learninstruc.2016.12.006

Andersson, C., & 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. https://doi.org/10.1080/0969594X.2018.1430685

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

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

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. McGraw-Hill Education.

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

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. King's College London.

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

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

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

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (Eds.). (1971). *Handbook on the formative and summative evaluation of student learning*. McGraw-Hill.

Bloom, H. S., Richburg-Hayes, L., & 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. https://doi.org/10.3102/0162373707299550

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

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, *48*(10), 1109–1136. https://doi.org/10.1002/tea.20440

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Taylor & Francis.

Cordingley, P., Bell, M., Thomason, S., & Frith, A. (2005). *The impact of collaborative continuing professional development (CPD) on classroom teaching and learning*. Research Evidence in Education Library, EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, London, UK. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=136 on 17/12/2020.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4), 438–481. https://doi.org/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 (2015). *Schools, pupils and their characteristics: January 2015. Statistical First Release SFR 16/2015*. Department for Education.

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

DfE (2019). *English Baccalaureate (EBacc)*. https://www.gov.uk/government/publications/english-baccalaureate-ebacc/english-baccalaureate-ebacc.

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

Education Endowment Foundation (2015). *Intracluster correlations*. Technical document, Education Endowment Foundation. Retrieved from https://web.archive.org/web/20171025020731/https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf.

Education Endowment Foundation (2018). *Feedback*. Teaching & learning toolkit, Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/feedback/.

Education Endowment Foundation (2019). *Cost evaluation guidance for eef evaluations. Guidance document, Education Endowment Foundation. Retrieved from* https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/Cost_Evaluation_Guidance_2019.12.11.pdf.

Education Endowment Foundation (2021a). *Embedding formative assessment (re-grant)*. Project summary, Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.

org.uk/pdf/generate/?u=https://educationendowmentfoundation.org.uk/pdf/project/?id=3110&t=EEF%20Projects&e=3110.

Education Endowment Foundation (2021b). *Teacher feedback to improve pupil learning*. Guidance report, Education Endowment Foundation. Citing ahead of publication. Link can be provided from 11 June 2021.

Guskey, T. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "Learning for Mastery. *Journal of Advanced Academics*, *19*(1), 8–31. https://doi.org/10.4219/jaa-2007-704

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

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation*, *38*(1), 21– 27. 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. https://doi.org/10.3102/1076998606298043

Hodgen, J., & Marshall, B. (2005). Assessment for learning in England and mathematics: A comparison. *The Curriculum Journal*, *16*(2), 153–176. https://doi.org/10.1080/09585170500135954

Imbens, G. M., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

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

Jiang, Z., & Ding, P. (2020). Measurement errors in the binary instrumental variable model. *Biometrika*, *107*(1), 238–245. https://doi.org/10.1093/biomet/asz060

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

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

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Leahy, S., & Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In J. Gardner (Ed.), *Assessment and learning, chapter 4*. SAGE.

Leahy, S., & Wiliam, D. (2013). *Embedding formative assessment*. Specialist Schools and Academies Trust.

Leithwood, K., Day, C., Sammons, P., Harris, A., & Hopkins, D. (2006). *Seven strong claims about successful school leadership. Research report*. NCSL.

Leong, W. S., & Tan, K. (2014). What (more) can, and should, assessment do for learning? observations from 'successful learning context' in Singapore. *The Curriculum Journal*, *25*(4), 593–619. https://doi.org/10.1080/09585176.2014.970207

Levin, H., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*. SAGE. https://doi.org/10.4135/9781483396514

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

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

Murphy, R., Weinhardt, F., & Wyness, G. (2015). *Who teachers the teachers? A RCT of peer-to-peer observation and feedback in 181 schools. IZA Discussion Paper 11731*. IZA Institute of Labor Economics.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, *22*(2), 155–175. https://doi.org/10.1207/s15326985ep2202_4

Nicol, D. J., & 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. https://doi.org/10.1080/03075070600572090

Ormston, R., Spencer, L., Barnard, M., & Snape, D. (2014). 1: The foundations of qualitative research. In J. Ritchie, J. Lewis, & C. McNaughton Nicholls (Eds.), *Qualitative research practice* (2nd ed.). SAGE Publications.

Rose, J., Thomas, S., Zhang, L., Edwards, A., Anwandter, A., & Roney, P. (2017). *Evaluation of research learning communities*. Evaluation report, Educational Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Research_Learning_Communities.pdf.

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

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

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

Smith, E., & 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. https://doi.org/10.1080/0969594042000333896

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

Spencer, L., Ritchie, J., O'Connor, W., Morrell, G., & Ormston, R. (2014a). 10: Analysis in practice. In J. Ritchie, J. Lewis, C. McNaughton Nicholls, & R. Ormston (Eds.), *Qualitative research practice* (2nd ed.). SAGE Publications.

Spencer, L., Ritchie, J., Ormston, R., O'Connor, W., & Barnard, M. (2014b). 9: Analysis: Principles and processes. In J. Ritchie, J. Lewis, C. McNaughton Nicholls, & R. Ormston (Eds.), *Qualitative research practice* (2nd ed.). SAGE Publications.

SSAT (2018). *About SSAT*. https://www.ssatuk.co.uk/about/

Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change*, *7*(4), 221–258. pages https://doi.org/10.1007/s10833-006-0001-8

Thompson, M., & Wiliam, D. (2008). Tight but loose: A conceptual framework for scaling up school reforms. In E. C. Wylie (Ed.), *Tight but loose: Scaling up teacher professional development in diverse contexts* (chapter 1, pp. 1–44). Educational Testing Service.

Thompson, S., Gregg, L., & Niska, J. (2004). Professional learning communities, leadership and student learning. *RMLE Online*, *28*(1), 1–15. https://doi.org/10.1080/19404476.2004.11658173

Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education*, *24*(1), 80–91. https://doi.org/10.1016/j.tate.2007.01.004

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/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.

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

Wiliam, D. (2019). Conclusion: Why formative assessment is always both domain-general and domain-specific and what matters is the balance between the two. In H. Andrade, R. Bennett, & G. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (chapter 10, pp. 243–264). Routledge.

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

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087