Guidance for EEF pilot evaluations

October 2023

This document outlines the EEF's approach to selecting programmes for pilot evaluations and our key principles on designing and reporting pilot evaluations. It aims to provide high-level guidance to support evaluators who are planning or conducting EEF-funded pilot evaluations. Some of the information may also be relevant for the programme delivery teams when setting up the pilots.

The guidance has been developed in association with the EEF's Evaluation Advisory Group and reviewed by evaluators that have recently conducted EEF pilot evaluations. We are grateful for all the feedback we have received. This is a working document that we will continue to review and update to take account of evaluators' experiences and feedback.

Contents

Introduction

Definition and purpose of EEF pilot evaluations

Most evaluations that the Education Endowment Foundation (EEF) has funded to date have been impact evaluations, in particular, randomised controlled trials (RCTs). However, many of these trials have not found positive impact results on pupil outcomes. Large-scale RCTs are resource intensive and expensive to conduct (Speich et al., 2019). Running them in the education sector is especially challenging as large numbers of schools need to be recruited for the trial to be appropriately powered to detect an effect. The literature suggests two key reasons for why RCTs have null results or barriers to interpreting their results in education: (i) insufficient evidence underpinning the programme theory —with many programmes' causal links based on 'insights', not data and (ii) implementation failure or issues with implementation meaning that limited information about the programme theory can be drawn from the findings (Styles and Torgeson, 2018; Dawson, Yeomans and Brown, 2018; Lortie-Forgues and Inglis, 2019).

Pilots are small-scale preliminary studies conducted to inform the preparation of a more comprehensive investigation (Cadete, 2017). They are valuable in addressing some of these challenges before an impact evaluation of a programme is commissioned by understanding whether and how the programme can be implemented with fidelity, ensuring that the programme theory is supported by the literature, and that evidence gaps are identified early. Pilot studies test whether a programme is feasible to be implemented as intended—or at all—and identify the barriers and facilitators that affect implementation so that appropriate adjustments can be made to the programme before it is tested more rigorously. As such, the EEF is increasing its capacity to commission pilot evaluations to enable programmes to be in a stronger position to move through the evidence pipeline and, potentially, to scale.

Pilot or feasibility studies typically either focus on investigating an intervention or testing an intended trial design, or both simultaneously (Eldridge et al., 2016; Pearson et al., 2020). These types of studies are useful in supporting the development and testing of effective implementation strategies, addressing the uncertainties around evaluation design and methods, and identifying potential causal mechanisms of an intervention (Pearson et al., 2020). At the EEF, we do not make a distinction between pilot and feasibility studies. For our purposes, we refer to a 'pilot study' or 'pilot evaluation' as a stand-alone, independent evaluation of an intervention that investigates whether the programme (i) is feasible to implement or deliver; (ii) has evidence of promise to support the theory of change (ToC), and (iii) is ready to be evaluated in a trial before committing to funding an impact evaluation. Generally, we commission pilot evaluations that take the form of a mixed-methods implementation and process evaluation (IPE).

Pilot evaluations may include a small-scale pilot RCT that tests the feasibility of different evaluation designs or explores the acceptability of evaluation requirements to external stakeholders. However, these pilot RCTs are typically run in a small number of settings, and we rarely collect outcome data or publish findings related to the 'impact' of the programme as the estimates are not likely to be reliable. We occasionally fund pre-trial pilot work that tests or refines a *specific element of a programme or evaluation design*, for example, an online versus face-to-face training approach or paper versus online outcome data collection. Pilot evaluations are distinct from such shorter, pre-trial pilot work in that they serve a broader purpose and act as a 'trial run' of a *fully codified programme* to ensure that it has a good chance of being implemented and evaluated successfully in a trial.

A mixed-methods design is typically used to evaluate pilot programmes, focusing on implementation and processes without assessing programme impact. Therefore, pilot evaluations are not set up to tell us whether a programme worked or not; rather, they help identify the necessary or ideal conditions under which the programme could produce an effect, and support refinement of specific programme approaches before committing the resources for an efficacy trial. Findings from the pilot are used to assess whether the intervention is ready for progressing to an efficacy trial, the first stage in our pipeline at which we assess impact on pupil outcomes.

EEF's evidence pipeline

For context, it is useful to understand the EEF's evidence pipeline, which has five stages: (1) early-stage programme development, (2) pilot evaluation, (3) efficacy trial, (4) effectiveness trial, and (5) scale-up. We assess programme applicants and determine where they should enter our pipeline based on several factors (more detail provided in the following section). Each stage of the pipeline is designed to address specific purposes.

Early-stage programme development projects sit at the front of the pipeline and are intended to address an evidence or practice gap. These projects may go through several phases of adaptation, often through user-testing or formative feedback. The intention is that some of these projects will pass to the next stage of the pipeline to undergo pilot evaluations.

Pilot programmes are selected by the EEF based on the criteria outlined below. These are applied consistently but not rigidly, meaning that all criteria must be met—but to varying degrees. Some criteria might be considered more relevant or important for a particular round of commissioning. Programmes that are not ready for pilot may be considered for early-stage programme development work.

This guidance does not cover the following:

- the evaluation of early-stage programme development projects;
- projects that aim to evaluate commissioning processes and strategies—for example, EEFective Kent Pilot, Accelerator Fund evaluation;
- testing adaptations to specific delivery models (for example, moving from face to face training to an online webinar); such adaptations are usually tested formatively by the developers or as part of pre-trial pilot work; or
- piloting of impact assessments or data collection tools to establish reliability and validity of testing recruitment strategies or data collection approaches to inform trial design.[1]

Criteria for pilot programmes selection

When are programmes suitable for pilot evaluations?

The following criteria are considered when deciding whether a programme might qualify as a pilot evaluation.

[1] The EEF typically does not fund stand-alone studies to test data collection, recruitment approaches, or psychometric properties of assessments. This work may be included as part of a pilot evaluation or pre-trial pilot. This document does not provide specific guidance on how these should be conducted.

1. **Fit with portfolio**. The programme aligns with our strategic priorities and addresses a gap in the evidence, but few programmes in this area have been rigorously evaluated and no other similar programmes are ready for trial.
2. **Scale of delivery**. The programme has been delivered in a small number of schools or within a group of schools (such as one MAT). The programme has the potential to be delivered at scale as an efficacy or effectiveness trial (involving roughly 50 or more schools).
3. **Capacity and experience**. The delivery team currently has the capacity and experience to deliver to a small number of schools—around 15 to 20. The team is prepared to build capacity to scale the programme, if it were to progress to trial.
4. **Level of development**. Programme activities and materials are codified but may not be fully developed and have not been adequately tested.
5. **Feasibility of implementation**. There is some uncertainty around feasibility (which includes acceptability), but the concerns are not severe[2].
6. **Evidence for the ToC**. There is some evidence supporting the ToC but further evidence is needed to understand whether the causal assumptions are likely to hold.
7. **Programme differentiation**. The programme activities are sufficiently distinct from existing practice.

When are programmes not suitable for pilot evaluations?

A programme is considered unsuitable as a pilot evaluation when:

- it does not address a clear gap in the evidence or the EEF's strategic aims;
- it has not previously been delivered in any schools;
- there is inadequate organisational capacity or experience to scale and deliver the programme to a group of schools;
- no codified programme activities or materials exist;
- there are severe concerns about the cost or acceptability of the programme;
- there is weak or no evidence supporting the programme's theory or principles; or
- its activities are insufficiently distinct from usual practice.

When is a programme ready for an efficacy trial?

To be considered for regranting as an efficacy trial from the pilot stage, the pilot evaluation would need to demonstrate positive results in three areas. The same criteria are used for deciding whether a programme that has been piloted outside of EEF funding is suitable for an efficacy trial. It should demonstrate:

[2] *Feasibility* is defined as the extent to which a programme can be successfully used or carried out within a given setting (Karsh, 2004; Proctor et al., 2011): it covers the practical and logistical issues of delivery such as time commitment and resources; *acceptability* refers to the perception among implementation stakeholders that a given intervention, approach, or training is agreeable, palatable, or satisfactory (adapted from Proctor et al., 2011).

1. **feasibility of implementation**—that it is not overly burdensome to implement and can be accommodated by schools, is considered affordable to schools, and its content and implementation approach are acceptable to teachers or practitioners;[3]
2. **evidence of promise**—evidence in support of the programme's ToC, for example, the pilot evaluation has captured changes in teaching practice or pupil engagement; and
3. **readiness for trial**—evidence to suggest the programme can be scaled for delivery to around 50 schools: the key indicators are usually whether there is a manualised version of the intervention and whether the delivery team have sufficient capacity to deliver the programme with fidelity.

Given that one of the main purposes of commissioning pilot evaluations is to support programmes to scale and progress through the evidence pipeline, it is crucial to design pilot evaluations so that a clear judgement can be made about the above three areas.

Programmes that have been tested or piloted in a different context

A programme that has been successfully tested or piloted in another context—for example, region or country, age or year group, or setting type—can be funded as an efficacy trial without requiring a pilot evaluation in the new context if there is enough evidence to suggest that it will be feasible to implement, that the causal assumptions will hold, and that the programme can be scaled in the new context. Otherwise, re-piloting in the new context may be required ahead of an efficacy trial. If the programme is piloted in the new context, the same criteria apply to determine whether it is ready for an efficacy trial.

Below are examples from two projects where different decisions were made based on the following factors:

- the extent to which contextual factors were expected to influence the feasibility to implement the programme in the new context;
- the extent to which the causal assumptions underlying the logic model were expected to hold in the new context; and
- the cost and resources required to deliver the programme.

The Tips by Text programme is a text message curriculum developed by academics in the U.S.A. and adapted to the English context by the Behavioural Insights Team (BIT). The content of the text messages was edited to be aligned with the Early Years Foundation Stage Profile and was piloted in a number of schools in parts of England. The decision to test this programme in England as an efficacy trial was based on, first, robust evidence from RCTs demonstrating the impact of targeted and personalised texts in eliciting behavioural change in a variety of contexts and, second, the fact that the burden to participants and the cost of delivery were both very low. The developers tested the adapted messages with a few parents to ensure the language was understandable and appropriate ahead of the trial.

The Early Years Toolbox (EYT) is an app-based assessment providing early years practitioners with a low-cost and robust way to measure children's abilities and to help inform practice. The EYT was developed and tested in Australia and is being used across the world, with no existing

[3] The criteria of what is deemed 'feasible' or 'acceptable' should have been agreed between delivery team and evaluators at the set-up stage (see key considerations for pilot evaluation design).

evaluation conducted in England. We funded this programme as a pilot to explore the feasibility of implementation and the acceptability of the programme to early years practitioners and to understand whether the intervention could be scaled in a larger number of early years settings in England. As the programme involves training early years practitioners and the background and experience of early years practitioners was expected to differ across context, it was important to ensure its feasibility and acceptability and identify barriers to delivery before committing to funding a trial to test its impact on pupil outcomes.

The following section outlines the key considerations for conducting pilot studies across the planning and design and reporting stages.

Key Principles for Designing Pilot Evaluations

1. Tailor research focus to the needs of the programme

EEF pilot evaluations are typically set up to focus on testing a programme's evidence of promise, the feasibility of implementation, and its readiness for trial as these are the broad criteria for programmes to move through the evidence pipeline. However, the focus within these categories needs to be tailored to the specific needs of a programme. A reflection from our earlier pilot evaluations is that studies tended to ask generic questions on these three aspects that were not adequately specific to the intervention. The pilot objectives should be based on gaps in evidence from the existing ToC and on the information needed to make evidence-informed decisions in preparation for a trial. Other research questions may be valuable in some pilot evaluations. For example, some evaluations may want to consider understanding the level of spillover—evidence of the intervention affecting pupils who are not receiving the programme in the same class or setting—which can inform the design of a subsequent efficacy trial.

Below are some examples of research questions under the three pilot categories.

Feasibility of implementation

For feasibility of implementation, research questions about programme implementation can draw on a number of implementation outcomes including feasibility, acceptability, and fidelity. Dimensions including dosage and quality of delivery—and participant responsiveness also falls under 'fidelity' (Proctor et al., 2011). Relevant question may be:

- Does the programme seem implementable and easy to use?
- Are schools willing to adopt or adapt the programme?
- Are most participants adhering to and completing the programme? If not, why not? Is the dosage acceptable to participants?
- What contextual factors support, or act as barriers to, take-up?
- Is the implementation support system effective? What changes might need to be made?

Evidence of promise

For evidence of promise, the focus can be about indicative, actual, or perceived programme impact on a short-term (proximal) outcome. It can also explore aspects of the ToC including testing relationships between different components within the causal chain or the extent to which the programme differs from usual practice. Programmes selected for EEF pilots should already have a well codified programme TOC but the causal links between inputs, outputs, and outcomes may require further testing. Research questions that test some of the underlying causal or contextual assumptions can be helpful for informing a future trial. For example:

- Is there a change in teachers’ confidence and knowledge?
- Is there a relationship between teachers’ confidence or knowledge and pupils’ levels of engagement?
- Did pupils engage with and enjoy the programme activities?
- Is the level of support provided by the programme sufficient to enable practitioners to change their practice?
- How much does this programme differ from usual practice?

Readiness for trial

In relation to readiness for trial, the focus can be about the scalability of the programme or its evaluation elements. Questions can be asked about different scaling or delivery models, cost, or other factors or questions that could affect the trial design (for example, spillover effects, eligibility criteria, and recruitment barriers). The pilot evaluation should also comment on the extent of development or adaptation work that is required and make appropriate recommendations on next steps based on the pilot findings. This information will form a key success indicator and support the EEF’s regranting decision.

Below are some example questions for readiness for trial:

- Is the programme scalable in its current form? What level of programme modification is required for scaling to more schools, for a trial and beyond?
- Is the delivery approach optimal and cost-effective? Are there any alternatives?
- Is there any indication of contamination between parents or teachers within the school?
- Are the eligibility criteria acceptable and reasonable?
- What are the recruitment rates and retention levels? What is a reasonable recruitment timeline?

Some pilots might not cover all three aspects if there is already good evidence supporting some of these. For example, a programme that has strong evidence to support its ToC may only need an evaluation that focuses on testing its feasibility, acceptability, and readiness for trial. Evaluations could consider drawing information from various methods and sources to obtain a comprehensive view of each research question. They should aim to consider the variation in the intervention specificity and maturity (and system readiness), ensure that we address questions that are most relevant to the specific programme and its implementation, and catch potential problems, preventing them from escalating before an impact evaluation occurs. The pilot design can also include data collection to inform the next steps in terms of intervention (re)design, implementation, and impact evaluation: for example, collecting feedback on improving stakeholders' engagement strategies to help avoid attrition at a trial stage. The methodology should follow our IPE guidance, be pre-specified, and mapped out by research questions in the pilot evaluation plan (see below Design Pre-Specification section for more detail).

2. Pre-specify success indicators

The EEF uses the findings from the pilot evaluation to make informed decisions on regranting and therefore it is important that the pilot evaluation includes clear success indicators in the three areas of pilot focus. We expect multiple indicators to support the decision under each of the three areas and each success indicator is likely to be informed by various measures, quantitative or qualitative. These should be agreed between the evaluation team and the delivery team, and with the EEF, at the set-up phase of the study. A pre-specified set of success indicators can minimise (although not fully mitigate) risk of bias and ensure full transparency. If the pilot programme is to be regranted (either as an efficacy trial or a second pilot), this will be subject to the usual EEF process for appointing independent evaluators.

When considering success indicators, the evaluators' role is to use information from the ToC, the logic model, and discussions from the set-up meetings to identify the key variables that are expected to contribute to the success of the programme and specify appropriate measures for assessing each of these (see Appendix A for an example). The success indicators should be well aligned with the research questions in that no additional information would need to be captured beyond that already collected for addressing the research questions. For pupil-level quantitative measures, project teams can discuss whether there is justification to set a minimum threshold considered sufficient for success (for example, monitoring data from the delivery team indicates that more than 70% of pupils attended all the sessions). However, where there is a high level of missing data, the findings should be cautiously interpreted due to the potential for bias. We anticipate that for teacher- and school-level quantitative measures, it would be inappropriate to set thresholds of success due to the small sample sizes (of 15 to 20). This makes interpretation of results difficult as small differences in responses can lead to large differences in percentages for small samples. Evaluators could instead consider pre-specifying some broader categorical criteria to aid interpretation for success, for example, 'more than half' or 'nearly all'. Setting thresholds for success may support clearer shared expectations between delivery and evaluation teams and support interpretation of the findings. However, decisions on whether the programme progresses to trial will be made by the EEF on the basis of all the evidence collected against the success indicators and not solely on whether individual quantitative success indicators are met.

In some cases, it may be helpful for the evaluators to support in iterative testing of programme materials or processes, or to share feedback from teachers and observations during the evaluation. This would be acceptable given that programmes are codified but may not be fully manualised at pilot stage. However, any support from the evaluator on programme design should ideally be agreed during set up, focus on supporting specific adaptations (rather than sharing of general feedback), and be clearly documented in the report (including what information was shared and what adaptations to the programme were made as a result) to ensure full transparency to aid accurate interpretation of the programme's success.

3. Strong justification needed for including a control group

We expect most EEF pilot evaluations to be implementation and process evaluations (IPEs) focusing on schools and pupils taking part in the intervention. In exceptional cases where evaluators have a strong justification to include a comparison group—for example, when using an observational or case-control study, a quasi-experimental design, or an RCT—this will need to be agreed with the EEF and the delivery team.

For example, the SHINE in Secondaries pilot tested the feasibility of a QED, specifically regression discontinuity, as the developer had a strong preference not to carry out random allocation as it would mean the evaluation was not testing SHINE's usual model. The pilot evaluation was, therefore, an opportunity to test the feasibility of the evaluation design and establish thresholds and estimates to inform a larger scale evaluation.

If evaluators conduct exploratory quantitative statistics (such as chi-square or t-tests) as part of the pilot evaluation, the findings should be interpreted with caution and no causal inferences should be drawn. In some instances, this level of evidence can be valuable if it contributes to evidence in support of the ToC and requires minimal additional burden to teachers and pupils.

4. Strong justification needed to include pupil assessments

Pilot designs that include pupil assessments should have a clear rationale for their inclusion. For example, there may be a clear gap in the ToC that this could support or the assessment may be part of the programme delivery (such as an end of module test). If there is currently limited information on suitable pupil assessment measures for the programme to be evaluated in a potential efficacy trial, evaluators could consider conducting desk-based research in the first instance. If further information is needed about the recommended measure(s), it may be justified to collect pupil assessment data to assess the suitability of this measure to be used in a trial context. The rationale for testing the measure should be informed by what information is missing about the measure. For example, if it is a validated one-to-one measure that is adapted for use as a whole-class measure, the focus of the testing should be on the feasibility of implementation. If the measure has not been validated in previous studies, the testing could include assessment of any ceiling or floor effects or the psychometric properties of the measure. The analysis and interpretation of the assessment data should be aligned with the rationale and focus of the testing. The assessment of the psychometric properties of a measure could be carried out by either the evaluation team or the delivery team—and should in all cases be carried out by researchers with the necessary expertise. If the pilot evaluation aims to test the feasibility of implementing the assessment measure or use the measure to inform the theory of change, the data should be collected and analysed by the independent evaluator. For pilot evaluations, we encourage teams to consider minimising the burden on teachers and schools as far as possible.

5. Focus on short-term (proximal) outcomes to demonstrate evidence of promise

Not all pilot evaluations need to include quantitative measures to capture changes in outcomes. Programmes with a well-evidenced ToC may focus only on perception or gather feedback to understand the feasibility of implementation or the scalability of the programme. We encourage evaluators to use a mixed-methods approach and avoid drawing conclusions about feasibility or scalability based only on qualitative data from a selective sample. Programme monitoring data is crucial to assessing fidelity and should be collected and analysed, where possible. This may include website analytics, records of attendance, and programme documentation for individual pupils.

If the pilot evaluation aims to demonstrate evidence of promise then it would be appropriate to focus on short-term outcomes that can measure observable change relatively quickly, such as perception, behaviour, attitude, or skills. These can be measured through surveys, observations, or monitoring data.

The outcomes measured in a pilot can be at school, teacher, or pupil level, depending on the ToC and the level of existing evidence supporting parts of the causal chain. For example, if there is evidence from correlational studies showing an association between teachers modelling meta-cognitive strategies and pupils' use of meta-cognition, then the pilot evaluation may want to focus only on whether there is an indication of the programme changing teacher-level behaviour, without the need to measure pupil outcomes.

6. Appropriate sample size and sampling approach

While sample size calculation is not appropriate for pilot evaluations given the absence of impact analysis, we expect evaluators to justify the number of schools and pupils to be included for the pilot evaluation in the study plan. Based on previous EEF pilots, we think that between ten and 20 schools is often appropriate for the type of qualitative and quantitative data needed to test the feasibility and acceptability of a programme. However, this should be tailored to the research questions and focus. We expect evaluators to follow the same principles outlined in the EEF's IPE guidance on sampling for process evaluations, and to ensure that there is a good range and representation of characteristics in schools and participants involved in the evaluation.

7. Explore core components and compliance thresholds

Programmes that are selected for pilot evaluations typically are already well codified but the causal links in the ToC may not be supported by much or any evidence. The pilot evaluation is therefore a good opportunity to provide additional evidence around the causal pathway and clarification on which programme activities are perceived to be key to improving outcomes (that is, core) and which are 'nice-to-haves'. All pilot evaluations should aim to use the results from the pilot to refine the programme ToC as an output.

If the aim of the pilot is to assess the core components of the programme, the evaluation should aim to capture the implementation of each core component and the compliance with it. For example, a programme that has a teacher training workshop as a core activity should capture data on attendance and engagement with the training as well as exploring the fidelity of this component and its association with perceived outcomes.

This data could also inform the indicator and threshold of compliance that should be adopted in a trial. For example, if the pilot results suggest that most pupils only completed five out of ten sessions but there were indications of the expected behaviour change, this may provide evidence to support a minimum threshold of five sessions as the compliance measure for the trial.

8. Explore 'feasibility of implementation' and causal assumptions for disadvantaged pupils

Pilot evaluations provide an excellent opportunity to start exploring whether a programme may lead to differential impact for pupils who are eligible for free school meals (FSM). We encourage evaluators to work with delivery teams to articulate any potential barriers to implementation for FSM-eligible pupils and mechanisms that could lead to differential impact. This may be explored through reviewing the causal and contextual assumptions and understanding whether they might be different for FSM-eligible pupils. These factors can then be incorporated into the evaluation design to be explored in more depth—for example, to assess whether different implementation strategies or approaches are needed for the intervention to reach FSM-eligible pupils or whether the intervention is equally acceptable to them. A success indicator for a programme could then include something around 'no indications of implementation barriers for disadvantaged pupils' (see Appendix A).

Setting up pilots

Agreeing the evaluation design

The process of setting up pilot evaluations is similar to that for EEF trials. The EEF will hold a ToC workshop with the delivery team before commissioning the evaluation. Once an evaluation team has been appointed, an IDEA workshop followed by two set-up meetings will take place where the research questions, evaluation design, and success indicators are discussed and agreed between the delivery team and evaluation team and with the EEF. The agreed evaluation design and corresponding project budget will then be scrutinised and approved by the EEF's Grants Committee.

Duration of the evaluation

Testing the full duration of the programme may not be required for all pilot evaluations. We recommend tailoring the timeline of the pilot to reflect the demands of the research questions. If the programme's ToC is already supported by evidence and the prime purpose of the pilot is to test the feasibility of delivery and scalability of the delivery model, it may be sufficient to test the delivery of the programme over just one academic term. For example, a pilot may want to specifically test the feasibility of delivery over the summer term when schools have other priorities like exams. In other cases, it may be beneficial to test the entire length of the programme.

Design pre-specification

Pilot evaluation study plans will be peer reviewed by one peer reviewer alongside the EEF's technical review before sharing with the delivery team and publishing online. We expect the evaluation study plans to follow the EEF pilot study plan template. Where needed, the template can be adapted in consultation with the EEF evaluation manager.

Incentives

Any incentives provided to settings or schools for taking part in pilot evaluations should be considered on a project-by-project basis, with the following three principles in mind.

1. Burden of evaluation activities

We anticipate that the burden of evaluation activities for pilots is relatively low and therefore, in most cases, school-level incentive payments would be unnecessary. Across past evaluations, if schools are asked to administer or support external administration of pupil assessments, it has been common to provide an incentive payment of £150 to £250 per setting, per period of assessment. However, if school staff are invited to participate in interviews, focus groups, or surveys, a small incentive paid to the school or the individual could be considered to increase participation. This may vary depending on the length of the evaluation activity. In rare circumstances where it is necessary to video-record lessons, evaluators may want to consider providing additional incentives for this.

2. Importance of the evaluation activity

If the evaluation activity contributes to the success indicator, we may want to consider incentivising this more heavily to maximise data returns. For example, if survey results are the only way of obtaining an indicator that supports a key decision in the programme's success then we may want to consider increasing the incentive amount for this activity.

3. Incentives should not detract from assessing feasibility and acceptability

If incentives are considered necessary for pilot evaluations, it should be clear to settings and schools that these are given as a thank-you for participating in the evaluation activities, not for delivering the programme. To avoid any detraction from assessing feasibility or acceptability of the programme, evaluators should consider the messaging and timing of delivering the incentives carefully. It may be advisable for the evaluator, rather than the delivery team, to distribute the incentive payments for case studies, observations, interviews, and focus groups and as a lump sum once all evaluation activities have been completed.

Data protection and archiving

We expect all pilot projects to obtain ethical approval and adhere to all data protection regulations. All relevant procedures for ensuring data quality, anonymity, and confidentiality should be specified in the evaluation plan and report and be included in the recruitment materials. There is no expectation that pilot IPE data requires archiving, nor that the evaluation plan needs to be registered. If the pilot evaluation collects pupil assessment data, we would expect this data to be archived for future analyses or meta-analyses.

Reporting on pilot findings

All EEF pilot evaluations are published on the EEF's website. They follow a similar quality assurance process as trial reports in that each pilot report is reviewed by the EEF evaluation manager and by at least one external peer reviewer from the EEF's peer review panel with relevant expertise and skills. The pilot report template can be found on the EEF's website. Where needed, the template can be adapted in consultation with the EEF evaluation manager.

Below are key considerations when writing pilot reports:

- Follow recommendations from this guidance and the EEF's IPE guidance.
- Provide clear justification for including each research question, linking this to the ToC and clarifying how the research questions address evidence gaps.
- The discussion should include:
    - whether the findings observed in the study are likely to be transferable to other schools, teachers, or pupils with a range of characteristics;
    - implications for programme design, highlighting areas where adaptation may be needed based on pilot findings;
    - implications for the programme's ToC, contextual assumptions, or causal chain; and
    - implications for the design of the implementation and evaluation plan of a future trial of the programme, if it is taken to trial.

- The conclusion section should consider the extent to which the success indicators set out in the study plan are met, clarifying how the results support the conclusion on evidence of promise, feasibility of implementation, and readiness for trial, and provide recommendations for next steps.

Presentation timeline and format

Ideally, the decision to take a programme from pilot to efficacy stage can be made with sufficient time for the evaluation to be set up in the following academic year. As such, we would recommend evaluators and the EEF to agree on the reporting timeline and format at project set-up so that timely decisions can be made.

All pilot evaluations should include a PowerPoint presentation of the preliminary key findings before the full draft report is completed. The presentation is attended by the programme and evaluation managers at the EEF and the programme delivery team. The structure and focus of the presentation should be agreed with the EEF in advance. We anticipate that the results relating to the success indicators would be presented alongside the methodology and limitations (see Appendix B for an example agenda for the presentation). The evaluation team should present preliminary key findings from the pilot evaluation as early as possible so that EEF can make timely programme regranting decisions and recommendations and seek our Grants Committee's approval. In some (rare) cases where delivery is only over a term or two, a findings presentation may not be needed if the evaluation team can submit a draft report by mid-August.

References

Cadet, L. (2017) 'What is a Pilot Study?': https://s4be.cochrane.org/blog/2017/07/31/pilot-studies

Dawson, A., Yeomans, E. and Rosa Brown, E. (2018) 'Methodological Challenges in Education RCTs: Reflections from England's Education Endowment Foundation', *Educational Research*, 60 (3): 292-310. doi:10.1080/00131881.2018.1500079

Eldridge, S. M., Chan, C. L., Campbell, M. J., Bond, C. M., Hopewell, S., Thabane, L., et al. (2016) 'CONSORT 2010 Statement: Extension to Randomised Pilot and Feasibility Trials', BMJ. 2016;355:i5239. doi: 10.1136/bmj.i5239

Karsh, B. T. (2004) 'Beyond Usability: Designing Effective Technology Implementation Systems to Promote Patient Safety', *Quality and Safety in Health Care*, 13, pp. 388–394. doi: 10.1136/qshc.2004.010322

Lortie-Forgues, H. and Inglis, M. (2019) 'Rigorous Large-Scale Educational RCTs are Often Uninformative: Should We Be Concerned?', *Educational Researcher*, 48 (3), pp. 158–166. https://doi.org/10.3102/0013189X19832850

O'Cathain, A., Hoddinott, P., Lewin, S., Thomas, K. J., Young, B., Adamson, J., et al. (2015) 'Maximising the Impact of Qualitative Research in Feasibility Studies for Randomised Controlled Trials: Guidance for Researchers', *Pilot Feasibility Studies*, 1, article 32. https://doi.org/10.1186/s40814-015-0026-y

Pearson, N., Naylor, P. J., Ashe, M. C., Fernandez, M., Yoong, S. L. and Wolfenden, L. (2020) 'Guidance for Conducting Feasibility and Pilot Studies for Implementation Trials', *Pilot and Feasibility Studies*, 6, article 167. https://doi.org/10.1186/s40814-020-00634-w

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R. and Hensley, M. (2011) 'Outcomes for Implementation Research: Conceptual Distinctions, Measurement Challenges, and Research Agenda', *Administration and Policy in Mental*

Health and Mental Health Services Research, 38 (2), pp. 65–76. doi: 10.1007/s10488-010-0319-7

Speich, B., Schur, N., Gryaznov, D., von Niederhäusern, B., Hemkens, L. G., Schandelmaier, S., et al. (2019) ‘Resource Use, Costs, and Approval Times for Planning and Preparing a Randomized Clinical Trial Before and After the Implementation of the New Swiss Human Research Legislation’, *PLoS ONE*, 14 (1): e0210669. https://doi.org/10.1371/journal.pone.0210669

Styles, B. and Torgerson, C. (2018) ‘Randomised Controlled Trials (RCTs) in Education Research: Methodological Debates, Questions and Challenges’, *Educational Research*, 60 (3), pp. 255–264. doi: 10.1080/00131881.2018.1500194

Appendix A: Examples of success indicators

Success indicators and assessment methods, taken from the SPACE pilot evaluation plan.

Pilot criteria	Success indicators	How to assess this?
Feasibility of implementation	F1. Teaching staff consider the intervention implementable (with minor amendments)	Teaching staff surveys, teaching staff interviews, observations
	F2. Teaching staff consider the intervention acceptable (with minor amendments)	Teaching staff surveys, teaching staff interviews, observations
	F3. Schools are able to deliver the intended intervention dosage within the defined period *	Programme monitoring data, teaching staff interviews
	F4. Schools are able to deliver the intervention with medium to high fidelity **	Teaching staff surveys, teaching staff interviews, programme monitoring data, observations
	F5. There are no indications of specific implementation barriers for pupils from disadvantaged backgrounds accessing the programme	Teaching staff and SLT interviews
Evidence of promise	P1. Findings indicate that SPACE has a positive influence on teacher knowledge, understanding and/or confidence in spatial skills	Teaching staff surveys, teaching staff interviews
	P2. Indication of SPACE leading to improvements in children's spatial reasoning and mathematics skills	Teaching staff surveys, teaching staff interviews, child outcome measures
Readiness for scale	S1. There are viable strategies to collect sufficient data to monitor compliance and fidelity	Programme monitoring data, delivery team focus groups
	S2. SPACE can be scaled for an efficacy trial (with minor amendments)	Observations, teaching staff and SLT interviews
	S3. There is a viable vision for delivering SPACE at scale	Delivery team focus groups, SLT interviews

* The programme is designed to be delivered as 12 30-minute sessions delivered over six weeks with a maximum of two sessions per week. To account for school term times, staff illness, and other unforeseeable interruptions, 'intended dosage' allows for delivery of the 12 sessions over eight weeks with a maximum of three sessions per week.

** Fidelity will be established using a composite score which incorporates key elements of programme fidelity including dosage, training attendance, and delivery quality.

Appendix B: Suggested agenda and structure for the pilot findings presentation

Example Agenda:

1. Introduction: EEF (~10 mins)
2. Presentation of interim findings: evaluation team (~40 mins)
3. Q&A: EEF, delivery team, evaluation team (~20 mins)
4. Next steps: EEF (~5 mins)
5. AOB: EEF (~5 mins)

Suggested structure for the findings presentation:

1. Introduction
    a. Evaluation aims
    b. Interim reporting focus
2. Description of the research questions included in this presentation
3. Methods and data sources
4. Key findings—from each data source, linked to research questions
5. Success indicators (see Appendix A)
    a. Description of the pilot criteria
    b. Assessment based on data collected thus far
6. Limitations of the evaluation and lessons learned
7. Key recommendations for the delivery team (if there is time)
8. Next steps