Evaluation of Complex Whole-School Interventions: Methodological and Practical Considerations

A Report for the Education Endowment Foundation

October 2017

Jake Anders, Chris Brown, Melanie Ehren, Toby Greany, Rebecca Nelson
UCL Institute of Education
20 Bedford Way
London
WC1H 0AL
Tel: 020 7331 5149
Email: jake.anders@ucl.ac.uk

Jessica Heal, Bibi Groot, Michael Sanders
Behavioural Insights Team
4 Matthew Parker Street
London
SW1H 9NP

Rebecca Allen
Education Datalab
11 Tufton Street
London
SW1P 3QB

Contents

1 Introduction

Evaluating the impact of complex whole-school interventions (CWSIs) is challenging. However, what evidence there is suggests that school leadership and other elements of whole-school contexts are important for pupils' attainment (Leithwood *et al*., 2006), suggesting that interventions aimed at changing these have significant potential to improve pupil outcomes. Furthermore, strong leadership is likely important for the effective implementation of many interventions funded by the EEF since even class-level or targeted programmes are more likely to work best within supportive and effective settings.

We therefore welcome the EEF's commitment to exploring the issues inherent in evaluating CWSIs. Developing design and practice for evaluations of this type of intervention, focusing on the issues of complexity and managing change across a whole school, increases the scope of projects of which the EEF may confidently fund evaluations.

In this document, we provide key messages for EEF evaluators on how to get the most out of evaluations of CWSIs, including considerations for both design and implementation. As far as possible, our suggestions aim to be practical steps that evaluators can implement immediately. A number of issues, and points 13 and 14 below in particular, require either further investigation or decisions from the EEF.

Defining Complex Whole-School Interventions (CWSIs)

For the purposes of this document, we define a CWSI as an intervention that combines multiple components that interact with one another within a context and aims to produce change (Moore et al., 2015). CWSIs may have many potential 'active ingredients' (Oakley et al., 2006). Other elements of complexity may include measurement of a range of outcomes, or targeting different levels of the organisation.

Rogers (2008) suggests that interventions designed to influence improvement in the whole school may be 'complicated', with significant factors working in combination at different levels with variations dependent on context, or 'complex', which may involve 'recursive causality (with reinforcing loops), disproportionate relationships (where at critical levels, a small change can make a big difference—a "tipping point") and emergent outcomes' (p.29). It is worth noting that CWSIs are also likely to be complicated.

Rogers (2008) suggests that evaluations of complex interventions must address these issues by collecting early evidence of small changes that can be tracked throughout implementation and allow the identification of 'tipping points' where a small change can have a disproportionately large effect (p. 38), and by employing a flexible logic model that can be used to guide planning and implementation but also revised as plans change (p. 40).

'School improvement by design' in the U.S.A. (Rowan *et al*., 2009), illustrated with examples of models funded through the Comprehensive School Reform (CSR) federal initiative from 1998 to 2007, took a similar approach to defining relevant interventions. Suitable interventions must be applicable across many schools, and might consist of curricular or pedagogical changes but, 'equally important, these designs also frequently include blueprints for organizational practices that allow the core instructional parts of the design to be implemented faithfully and used effectively in schools' (p. 11). In addition, such programmes involve the application of an approach which has been designed by an external developer and with implementation and evaluation support.

1.1 Using this document

We suggest that all evaluators planning the evaluation of a CWSI read **Section 2** (on the overall design) and **Section 5** (on implementation and process evaluation). The most

useful parts beyond this will depend on the nature of the intervention, its expected logic model, and the design selected after consultation with the EEF and the project team. This document should be read alongside previous EEF guidance. For example, **Section 5** does not replace the EEF's guidance on implementation and process evaluation (Humphrey *et al*., 2016a) but rather highlights particularly important considerations in the context of evaluating CWSIs.

- **Section 2 discusses the overall design** of the evaluation of CWSIs, including:
  - when to consider alternatives to randomised controlled trials (RCTs); and
  - the use of a multi-stage design approach.
- **Section 3 discusses designs of RCTs**, including:
  - which are best suited for CWSIs,
  - highlighting other elements of RCT design with particular considerations when evaluating CWSIs.
- **Section 4 discusses the use of quasi-experimental designs** (QEDs) for impact evaluation, focusing primarily on the potential for matched difference in differences evaluations, including:
  - planning a QED impact evaluation;
  - selecting QED approach; and
  - the analysis and reporting of a QED impact evaluation.
- **Section 5 discusses the use of iterative mixed-methods implementation and process evaluation** (IPE) designs, in which RCTs and QEDs should be embedded to understand fully the impact of CWSIs, including:
  - the overall approach to IPE design for this kind of evaluation; and
  - particular considerations when developing a logic model and hypothesising and seeking to measure core components, moderators, and mediators.
- **Section 6 discusses measures to capture likely mediators to attainment relevant in CWSIs**, including school leadership, for the purposes of evaluations of this type, providing examples of instruments and models that might be adapted for collecting standardised data.
- **Section 7 presents the potential for 'data availability-led evaluations'** as an alternative way of establishing an evaluation which reduces the costs and burden of testing while providing outcomes data from across a school's age-range.

At the beginning of each of the sections, we provide a more detailed breakdown of the content to guide the reader. Many of the sections refer to tools or further detail in appendices. In addition, a glossary of terms is included in Appendix 1.

Key Messages to consider for the evaluation of complex whole-school interventions (CWSIs)

Designing evaluations of CWSIs

1. The EEF should commission RCTs of CWSIs, where this is possible. In instances where this is not possible, it should consider QEDs instead. Where neither is possible, non-experimental evaluations may be useful to develop a better understanding of an early-stage programme to allow future RCTs or QEDs.
2. One solution to the problem of not knowing the mechanisms and channels through which a CWSI might work in advance is to use multi-stage evaluation protocols, described in Section 2.2.
3. Although they are not without their challenges, a parallel school-level cluster randomised controlled trial (cRCT) is the most robust and practical way of evaluating the effect of CWSIs.

Using randomised controlled trials to evaluate CWSIs

4. To capture the full effect of an intervention, evaluations of CWSIs are likely to be longer than most of the evaluations currently funded by the EEF. However, they must not be allowed to become longer than necessary.
5. Longer trials are likely to be associated with higher rates of attrition and non-compliance. The length of the evaluation should be taken into consideration when estimating statistical power.
6. Measures discussed in Section 3.2.2, such as building in real-time monitoring of process data to allow targeted support of at-risk schools, should be considered to increase retention of schools.
7. Implementation and Process Evaluation (IPE) (discussed further in Section 5) should explore why pupils/teachers/schools drop out and whether this was related to the intervention.
8. Project and evaluation teams should agree and implement a combined recruitment, communication, and engagement strategy during recruitment and into the period of implementation. This should include the elements discussed in Section 3.2.4.
9. Use the checklist in Appendix 4 to ensure school readiness to participate in the intervention and evaluation.

Using quasi-experimental designs to evaluate CWSIs

10. Where it pursues a quasi-experimental approach to evaluation, the EEF should concentrate on planned evaluations, rather than ‘natural’ experiments. This will guard against their use in settings where they are unlikely to produce credible estimates of impact.
11. Analysis of QEDs should be carried out on an intention-to-treat (ITT) basis to maximise comparability with this approach for RCTs.
12. To ensure there is clear identification of the analysis sample, the evaluation and project teams should agree a ‘quasi-randomisation’ date at which the sample of intention-to-treat schools is finalised.
13. We recommend that the EEF considers the use of a matched ‘difference in differences’ design for future projects where RCTs are not feasible. It will not be suitable for all projects, nor is it the only possibility: the decision should be taken in conjunction with project and evaluation teams.
14. We do not advocate one specific approach to matching but rather transparency about the approach used and ample robustness-checking of that approach (see below).
15. Evaluators should test the robustness of their findings to making different decisions about the matching approach. We suggest that at least five robustness checks are pre-specified at the design stage and the results reported in the evaluation report. We should not have much confidence in findings that are fragile and strongly dependent on the specific method used.
16. Use the principles identified in Box 4.1, such as being guided by theory and previous evidence, to select variables for inclusion in the matching model. It is especially important to ensure that matching variables could not have been affected by the treatment itself.
17. Failure of common support undermines the comparability between treatment and matched comparison samples. It should be explored and, if necessary, imposed as part of analysis.
18. Matching should be carried out blind to outcomes. To ensure this, best practice would be to carry out matching at a time roughly analogous to when randomisation would be carried out in an RCT.
19. The matching process should be made as transparent as possible to ensure the credibility of findings and allow for replication. Recommended reporting requirements are outlined in Section 4.3.
20. Statistical inference in difference in differences analysis should be conducted in line with the principles highlighted by Brewer, Crossley and Joyce (2013).
21. Balancing checks on observables, as currently reported for RCTs, should also be reported in this

setting. This should be done for, at a minimum, all characteristics on which schools have been matched.

22. When using difference in differences, analysis of pre-treatment common trends in the outcome variable between the treated group and the matched sample should be reported. This should include both placebo tests of pre-treatment impact and graphical plots of pre-treatment trends in outcome variables among treatment and comparison groups.

Implementation and process evaluation for CWSIs

23. Use a multi-phase mixed methods IPE design to ensure findings are maximising the efficiency and accuracy of the evaluation. The order of the quantitative and qualitative data collection and analysis should be fully rationalised in the evaluation protocol.
24. In some cases, there may be value to commissioning a follow-up IPE where there are unexpected results from the impact evaluation to help understand such results.
25. The development of a model of the logical relationships between the core components, moderators, mediators, and outcomes of a programme is particularly important in the context of CWSIs. In line with Key Message 2, the logic model should not be a static product, rather it should be updated as part of a multi-stage evaluation protocol approach.
26. Create a preferred set of potential moderators that the EEF requires evaluators to collect, potentially via a standardised method of data collection, from all schools participating in evaluations of CWSIs.
27. Isolate moderators that influence implementation to a greater or lesser extent by creating 'scales', hierarchies, or definitions. An example of a moderator 'dashboard' that could be developed further as part of individual evaluations or for a standardised EEF approach is provided in Appendix 2.
28. It is important to monitor mediators identified as important by the logic model throughout the evaluation. These should include elements such as teacher attitudes that go beyond tick-box measures of fidelity. Further suggestions on measurement are provided in Key Messages 30 and 31.
29. In the context of CWSIs, the IPE should pay particular attention to the importance of Reach, which should consider all persons potentially influenced by, or who could be influencing, the intervention.

Measures to capture likely mediators relevant to CWSIs

30. Evaluators should consider adaptation and re-use of the measures of mediators summarised in Appendix 8 in the IPE of CWSIs.
31. Where possible, measures and data collection regarding mediating mechanisms should attempt to complement the types of self-assessment most schools conduct as a matter of course. Consultation with schools during recruitment may help to support this.

Alternative approaches to attainment measurement

32. At efficacy trial stage, a lower value is generally placed on external validity compared to costs (since there is generally less evidence of promise at this stage). During design, implementation teams should consider whether it is possible to make use of tests that are already administered in schools, carefully documenting all the drawbacks of doing so.
33. Evaluators should consider testing a random sample of pupils from across the school to collect outcome measures. Samples of at least 30 pupils per school are unlikely to reduce statistical power much relative to whole-school testing. However, this restricts outcome measures to those suitable for use across the whole age range. Furthermore, the benefits in terms of reducing costs and burden of testing may not be as large as the reduced number of pupils tested would suggest.

2 Designing evaluations of complex whole-school interventions

The evaluation of CWSIs will often require adaptation of one or more elements of the EEF's usual approach to evaluation. This section lays out suggested changes to the underlying design of evaluations. This covers two main elements:

- consideration of whether the most appropriate design is a randomised controlled trial, or whether the use of quasi-experimental and, possibly, non-experimental approaches to evaluation would be more appropriate; and
- advocating use of a multi-stage design approach, typified by use of a multi-stage evaluation protocol, rather than the traditional single evaluation protocol.

At the outset, it is worth highlighting that we consider RCTs the best way to estimate the unbiased causal impact of CWSIs, where they are feasible (which will be in many cases). However, inevitably, we discuss the alternatives rather more than usual EEF practice because the latter is covered by existing guidance.

2.1 Experimental, quasi-experimental, and non-experimental evaluations

Key Message 1: Use RCTs where possible but consider alternatives
The EEF should commission RCTs of CWSIs where this is possible. Where not possible, it should consider QEDs instead. Where neither are possible, non-experimental evaluations may be useful to develop a better understanding of an early-stage programme to allow future RCTs or QEDs.

We argue that the range of situations where an RCT is likely to be technically impossible is likely to be relatively small, as the technique is robust to a multitude of challenges. Nevertheless, when evaluating CWSIs there will be situations where RCTs are not desirable (in terms of not being the best way to answer the question at hand given budgetary constraints) or are simply not feasible.

In general, it is likely to be the case that the more complex an environment is, the more complex an intervention is; and the longer the period over which the outcome is to be measured, the more challenging an RCT will be to run. This is not a hard and fast rule: we are aware of RCTs with long follow-up periods from the medical literature, however these seldom have both the complexity and whole organisation nature of the interventions we are considering here.

Features of an intervention that suggest it is important to consider the feasibility of RCTs

- Intervention is such that it is likely to be difficult to recruit schools to be randomly allocated because it would require dramatic organisational changes at short notice (such as changes in staff deployment) depending on the outcome of the randomisation.
- Programmes requiring groups of schools to work together meaning that:
  - the sample required for an RCT to be sufficiently powered makes it unfeasibly expensive and difficult to manage;
  - it may not be feasible to expect schools formed into these groups before randomisation not to work together if allocated to the control group, preventing this from being a true 'business as usual' control.
- Intervention requires a lengthy period of implementation over which it will be difficult to preserve the integrity of a control group.

In these cases, RCT designs, which function well with well-defined interventions, outcome measures, and timescales, are unlikely to sensibly serve the purpose of providing good

quality causal estimates of the effectiveness (or otherwise) of programmes. Such circumstances should be considered an exception to the rule in the work that the EEF funds, but will exist.

In Figure 2.1, we outline:

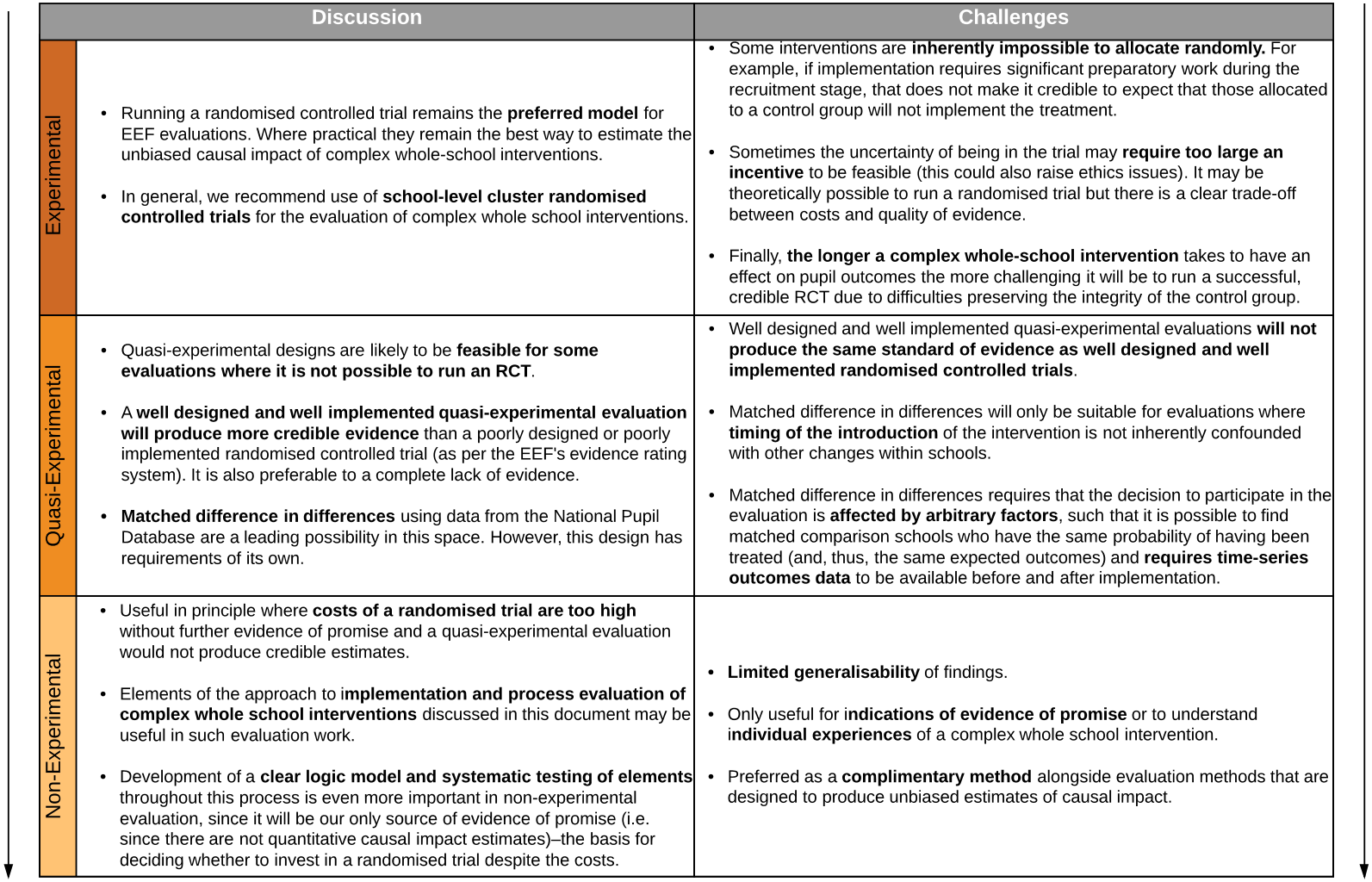
- key reasons why RCTs may not be feasible and when a QED alternative (discussed in detail in Section 3) should be considered; and
- settings in which QEDs are also likely to face serious challenges in producing a credible impact estimate.

The process of choosing the appropriate design for any given evaluation is not, and cannot be, a formulaic process. Although there may be a hierarchy in terms of 'quality of design' in the EEF's classification of the security of evaluation findings, there is not a hierarchy in terms of what will be possible: there may be interventions for which an RCT would be possible, but a QED would not. For example, a support programme for new school leaders could be randomly allocated but would be difficult to evaluate with a difference in differences QED due to being confounded with the change of leadership itself. As such, we do not provide hard and fast triggers for switching from an RCT to, say, a QED. Instead, we present the main challenges to successful implementation of each design to allow an assessment of the best design for each evaluation.

Reasons for the infeasibility of an RCT or a QED for evaluations of certain CWSIs may well emerge from the commissioning process. However, this will not always be the case as issues may not emerge until the appointed evaluator is able to discuss the key elements of the intervention with the developer. If this is the case, it may be necessary for the evaluator and developer to work together on a desk-based or small-scale feasibility study in order to map out how the evaluation would proceed under either model, illuminating inherent problems that would otherwise not become evident until the launch of the evaluation.

Where conducting an RCT or a QED presents serious challenges, it may still be productive for the EEF to commission non-experimental evaluation work. We recommend that this employs the principles for IPE outlined by Humphrey *et al*. (2016a) and developed for CWSIs in Section 4. In some ways, this is not new: the EEF has commissioned pilot studies since its inception. A key aim of such work should be to define the core components and active ingredients (see Section 4) that make the intervention work, vital to allow evaluation using RCTs or QEDs in future research.

Figure 2.1: Considering alternatives to randomised controlled trials



	Discussion	Challenges
Experimental	• Running a randomised controlled trial remains the **preferred model** for EEF evaluations. Where practical they remain the best way to estimate the unbiased causal impact of complex whole-school interventions. • In general, we recommend use of **school-level cluster randomised controlled trials** for the evaluation of complex whole school interventions.	• Some interventions are **inherently impossible to allocate randomly.** For example, if implementation requires significant preparatory work during the recruitment stage, that does not make it credible to expect that those allocated to a control group will not implement the treatment. • Sometimes the uncertainty of being in the trial may **require too large an incentive** to be feasible (this could also raise ethics issues). It may be theoretically possible to run a randomised trial but there is a clear trade-off between costs and quality of evidence. • Finally, **the longer a complex whole-school intervention** takes to have an effect on pupil outcomes the more challenging it will be to run a successful, credible RCT due to difficulties preserving the integrity of the control group.
Quasi-Experimental	• Quasi-experimental designs are likely to be **feasible for some evaluations where it is not possible to run an RCT**. • A **well designed and well implemented quasi-experimental evaluation will produce more credible evidence** than a poorly designed or poorly implemented randomised controlled trial (as per the EEF's evidence rating system). It is also preferable to a complete lack of evidence. • **Matched difference in differences** using data from the National Pupil Database are a leading possibility in this space. However, this design has requirements of its own.	• Well designed and well implemented quasi-experimental evaluations **will not produce the same standard of evidence as well designed and well implemented randomised controlled trials**. • Matched difference in differences will only be suitable for evaluations where **timing of the introduction** of the intervention is not inherently confounded with other changes within schools. • Matched difference in differences requires that the decision to participate in the evaluation is **affected by arbitrary factors**, such that it is possible to find matched comparison schools who have the same probability of having been treated (and, thus, the same expected outcomes) and **requires time-series outcomes data** to be available before and after implementation.
Non-Experimental	• Useful in principle where **costs of a randomised trial are too high** without further evidence of promise and a quasi-experimental evaluation would not produce credible estimates. • Elements of the approach to i**mplementation and process evaluation of complex whole school interventions** discussed in this document may be useful in such evaluation work. • Development of a **clear logic model and systematic testing of elements** throughout this process is even more important in non-experimental evaluation, since it will be our only source of evidence of promise (i.e. since there are not quantitative causal impact estimates)–the basis for deciding whether to invest in a randomised trial despite the costs.	• **Limited generalisability** of findings. • Only useful for i**ndications of evidence of promise** or to understand i**ndividual experiences** of a complex whole school intervention. • Preferred as a **complimentary method** alongside evaluation methods that are designed to produce unbiased estimates of causal impact.

2.2 Evaluation protocols for complex whole-school interventions: multi-stage protocols

Key Message 2: Multi-stage evaluation protocols
One solution to the problem of not knowing the mechanisms and channels through which a CWSI might work in advance is to use multi-stage evaluation protocols.

2.1.1 The problem

The use of pre-registered protocols for evaluations in education is, as in other fields, an increasingly recognised part of conducting rigorous research. Although analysis not specified in a protocol is not strictly prohibited by this process, it is treated as exploratory only. Importantly, findings emerging from exploratory analysis should not be treated as though they were the intended goal of an evaluation, and should ideally be verified using a separate evaluation in which these analyses are pre-specified and confirmatory. As such, it is generally recommended to capture and confirm the subgroups and theories in advance of commencing delivery and even randomisation (Humphrey *et al*., 2016a, p. 6).

The purpose and practice of protocols
Protocols document in some detail the process of an evaluation—from recruitment, through to the implementation of an intervention, to its monitoring processes, data collection, and ultimately its analysis.

Pre-registering an analytical strategy at the beginning of an evaluation, as well as having analysis and evaluation conducted by an independent third party, allows researchers to avoid the 'garden of forking paths', whereby analytical decisions are taken after data is collected, and so can be chosen based on which strategy confirms the researcher's hypothesis.

The EEF requires evaluators to publish a protocol before starting an evaluation, and a Statistical Analysis Plan within three months of the evaluation starting. These can be updated as the evaluation progresses, and any updates are published on the project page of the EEF website.

However, in the case of CWSIs, this purist method is unlikely to be practical. Conducting a second, perhaps five-year-long, study to confirm a finding from exploratory analysis would mean waiting perhaps 12 years from a study being initially commissioned to being able to talk confidently about its findings. As noted in our definition of CWSIs, emergence is an important aspect of complex interventions, which, Rogers (2008) argues, implies that 'specific outcomes, and the means to achieve them, emerge during implementation of an intervention' (p. 39). In this setting, this implies that it will be difficult to predict in advance which (if any) sub-groups, or outcomes, are likely to be particularly influenced by the intervention. Although for some interventions we might expect the effects to be particularly pronounced for maths or for English results, or for students from less affluent families, the interactions between a CWSI and its environment are likely to be more nuanced.

Example of problems with a single-stage evaluation protocol for a CWSI
Leadership interventions, for example, are likely to be affected by the personal characteristics of the headteacher involved, and that headteacher's relationships with particular teachers. Where the headteacher is a maths teacher, for example, one might imagine an intervention having a more positive effect on maths grades than English grades, and vice-versa if an English teacher. Although this is both plausible and testable (by conducting a subgroup analysis that focuses on students' grades in the head's own subject), it may be difficult to predict this relationship beforehand.

At the very least, the number of hypotheses for the mechanism through which a CWSI works which are plausible *ex ante* is likely to be large. Pre-specifying many outcome

measures or subgroups might be less problematic than deciding on them after the fact, but it is far from ideal. It would require potentially drastic corrections for multiple comparisons, such as those suggested by List *et al*. (2016), reducing the study's power and, thus, requiring far larger sample sizes to detect a given effect.

2.2.1 Multi-stage evaluation protocols

We suggest the use of multi-stage evaluation protocols to help alleviate this problem in the context of CWSIs while, importantly, preserving the principle of making analytical decisions before interacting with the data. These principles may be applied to RCTs (at both efficacy and effectiveness stages) and QED evaluations. They do not solve all the flexibility and design issues commonly raised by critics of such impact evaluation models, but can go some way to address concerns while preserving features of the design vital to its credibility.

Under this model, the following process would be adopted:

- An evaluation protocol—including description of the primary, full sample analysis—is written and published as usual in advance of the evaluation launch. In line with good evaluation practice, the primary analysis should not be altered once it has been agreed. As such, this protocol remains the main document relating to the evaluation. However, it should state at its conclusion that a second-stage evaluation protocol will be published later, ideally giving a date.
- At the end of the experimental period, the implementation and process evaluation (IPE; further discussion in Section 5) should be conducted and analysed – certainly before impact analysis is conducted and, ideally, before the quantitative data for this analysis are even available to evaluators. If designed and conducted well, the IPE should provide insight into the mechanisms and channels through which the intervention is working in a way that was not possible before the evaluation.
- The findings from the IPE are then used to form hypotheses testable using the quantitative data for impact analysis. This should be written up and published as the second-stage protocol. This can also reflect any changes to the evaluation that occurred due to its duration or complexity.

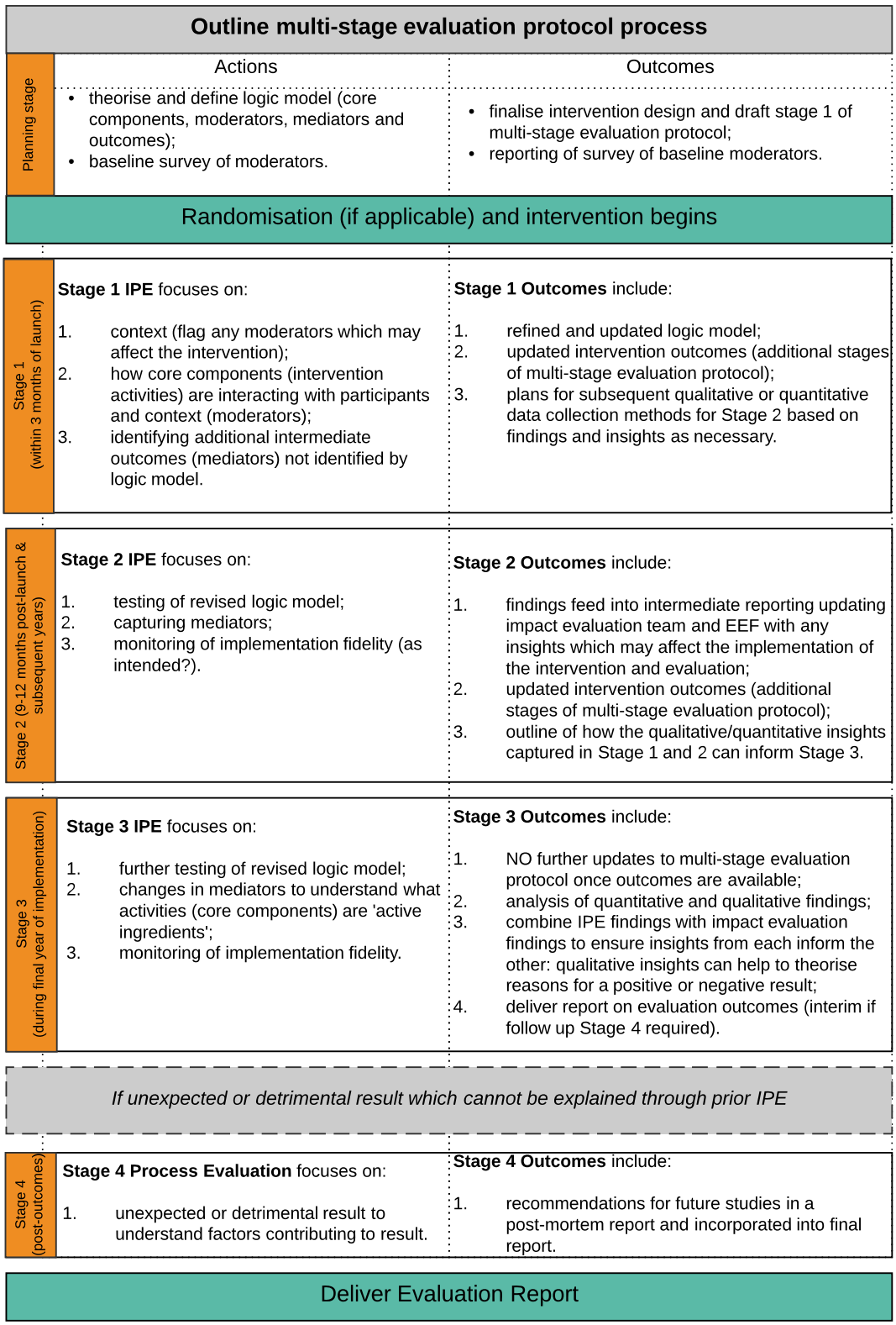
Figure 4.4 outlines the ongoing reflection and evaluation required in preparation for writing the second protocol, which should:

- start with the logic model[1] outlined at the evaluation design stage;
- draw on emerging findings from the IPE to provide evidence for, or contradict, the logic model outlined at the design stage to refine the approach; and
- work chronologically through the logic model, assessing—
    - the core components and active ingredients,
    - the moderators, and
    - the mediators.

Some elements, such as how well the intervention is implemented, will be assessed throughout the process.

[1] To be consistent with the approach taken by Humphrey *et al*. (2013) we use the terms logic model and theory of change interchangeably, although we are aware of some differences in the literature, but generally use the term logic model for consistency.

Figure 2.2: Model for developing multi-stage evaluation protocol-based evaluation of a complex whole-school intervention

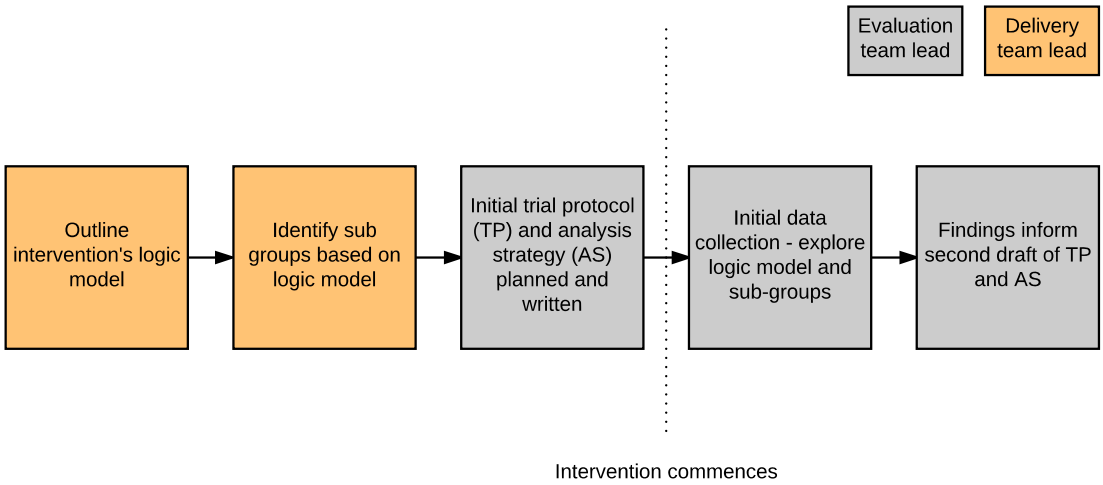


Follow-up IPE (see Section 5.1.2)
As shown in Stage 4 of Figure 2.2, this ongoing reflection and refinement process can provide the basis for commissioning a follow-up IPE in the event of an unexpected result from the impact evaluation, although this is less likely to occur if multi-stage evaluation is adopted. This is discussed further in Section 5.1.2.

We suggest organising reflection sessions at points throughout implementation (for example every 6 months) where members of the evaluation team meet and reflect on (using the logical flow outlined in Figure 2.3):

- how theory (core components, mediators and moderators) is reflected in practice, and where it is different; and
- how the logic model was envisaged at design stage, and if and how it should be updated to reflect changes observed as the evaluation proceeds.

Figure 2.3: Logical flow in a multi-stage protocol evaluation



The methods and approach outlined require a significant amount of work which will need to be resourced appropriately. It is also essential to outline the importance of these methods to schools in order that they understand the data 'ask' from the evaluation team.

We include an example outline of the first stage of a multi-stage evaluation protocol in Appendix 2.

3 Using randomised controlled trials to evaluate complex whole-school interventions

In this section:

- we assess the challenges of using RCTs for evaluating CWSIs and, in light of this, consider ways in which a standard RCT could or should be modified to suit CWSIs;
- we describe the situations in which a standard, parallel RCT—either individually or cluster randomised—might be appropriate for evaluating a CWSI; and
- we describe challenges to the validity of RCTs that are raised by using them to evaluate CWSIs, and describe mitigating actions.

3.1 RCT design

Key Message 3: Use parallel school-level cluster RCTs
Although they are not without their challenges, a parallel school-level cluster randomised controlled trial (cRCT) is the most robust and practical way of evaluating the effect of CWSIs.

We outline the advantages and disadvantages of different designs in Table 3.1. Individually randomised trials are fairly rare in education and are only suited to very simple interventions where spillover between individual participants (also known as 'interference' or violation of the Stable Unit Treatment Value Assumption, SUTVA) is either impossible or highly unlikely. In the case of CWSIs, it is impossible to conduct randomisation at this level: it is not practically feasible to randomise at a level lower than that at which the intervention is to be delivered, and so a CWSI needs to be randomly assigned to entire schools at a time.

Where individual randomisation is not possible, the most straightforward design is a parallel cluster randomised controlled trial (cRCT). This has the advantage of preventing spillovers between treated and control participants in the same school, and can be easier to implement. There are, though, substantial consequences for statistical power: compared to individually randomised trials: trials with a higher level of randomisation require many more individual participants in order to achieve a robust estimate.

Table 3.1: Advantages and disadvantages of RCT designs applied to the evaluation of CWSIs

Design	Advantages	Disadvantages
Individual randomisation	• Can be well-powered with small numbers of participants. • Analysis straightforward and transparent.	• Only suitable for simple interventions delivered to individuals where spillover between these individuals is impossible. Therefore, unsuited to evaluation of CWSIs.
Class-level cluster randomisation	• Requires fewer participants for equivalent power than school RCTs. • School-level covariates are balanced between treatment and control	• Only suitable for interventions delivered to classes with low risk of spillover between classes and therefore unsuited to CWSIs.

	• Lower risk of school attrition as all schools receive treatment.	• Requires larger numbers of schools to achieve the same level of statistical power than is the case with individual randomisation designs.
School-level cluster randomisation	• Suited to an intervention where the delivery unit is the school. • Low risk of spillover between treatment and control schools.	• Large number of schools (and therefore pupils) required for sufficient power. • High risk of attrition of control schools.
Stepped-wedge design	• Discourages attrition as all schools are treated.	• Analysis is more complex • Could only be rolled out in steps of several years (depending upon the length of time expected to see outcomes) due to the whole-school nature of implementation. • In some applications there is a lack of concealment in that schools or units that know they will be denied access to the intervention for a period of time adjust their behaviour in the intervening period in anticipation of programme entry, thus biasing their status with regard to control conditions.
Cross-over design	• Discourages attrition as all schools are treated. • Suited for short, discrete interventions.	• Unsuited to the evaluation of CWSIs because these cannot generally be ‘stopped’.

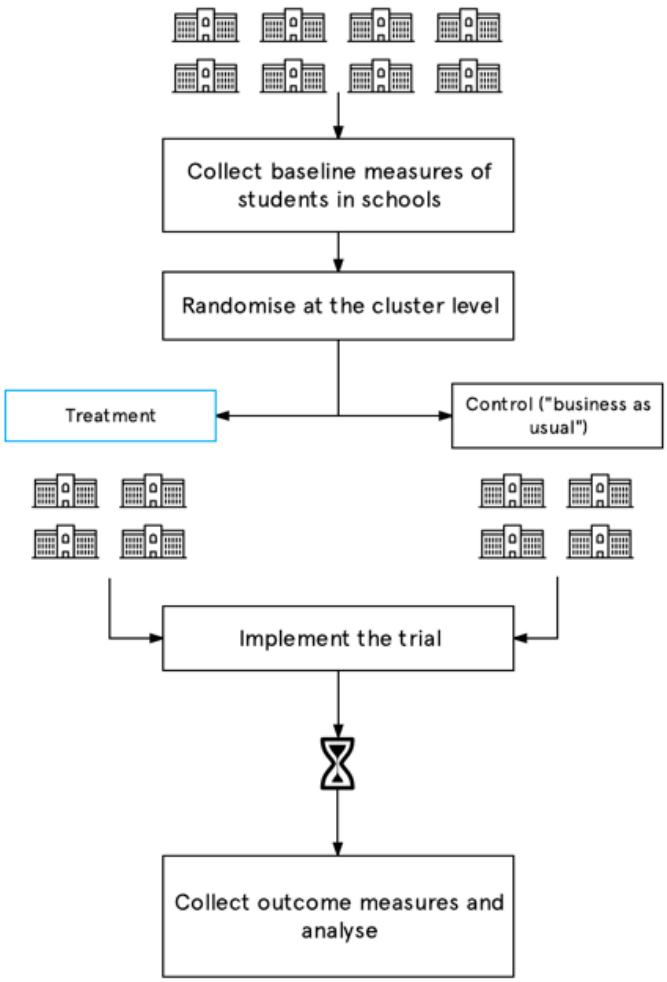
Cluster Randomised Controlled Trials (cRCTs)
In a cRCT, groups (or clusters) of participants are assigned together, and at random, to either a control condition or to one or more treatment groups.

In education, the two most common forms of clusters are the classroom—which may be appropriate in multi-form-entry primary schools, or in secondary schools where students are largely or entirely taught in the same groups—or the whole school.

In a cRCT, all schools have the same probability of being assigned to a given condition, and that assignment is independent of the assignment of other schools. For all participants within a school, however, the assignment of a given student predicts exactly the assignment of all other students in their school.

Figure 3.1, below, runs through the process of a cRCT.

Figure 3.1: Process of a cluster randomised controlled trial



3.2: Mitigating risks to RCT validity when evaluating complex whole-school interventions

Attrition, of pupils or schools, is a potential risk to the validity of all evaluations. This is discussed in existing EEF guidance. However, the risks may be particularly acute in the case of CWSIs because of:

- the increased length of evaluations; and
- the all-encompassing, and hence burdensome, nature of whole-school interventions.

This discussion focuses on the issue of school-level attrition based on the EEF's existing guidance (Education Endowment Foundation, 2013), experience with previous RCTs, and the particularly large implications for validity and statistical power. Nevertheless, it is important to consider issues of differential pupil attrition and potential crossover using existing guidance.

3.2.1 Considering the appropriate trial length

Key Message 4: Ensure evaluation is long enough to capture effects
To capture the full effect of an intervention, evaluations of CWSIs are likely to be longer than most of the evaluations currently funded by the EEF. However, they must not be allowed to become longer than necessary.

Key Message 5: Consider implications of trial length for statistical power
Longer trials are likely to be associated with higher rates of attrition and non-compliance. The length of the evaluation should be taken into consideration when estimating statistical power.

It is a feature of CWSIs that they may take longer to 'bed in' than simple interventions. However, a longer trial to capture this—for example a five-year-long trial (which would allow students who began a secondary school at the start of the trial to complete their GCSEs)—is more likely to suffer from substantial attrition and non-compliance. If an effect takes time to build, but the available sample gradually shrinks, there is a trade-off between treatment effects on the treated and the sample available to analyse. In the case of CWSIs, the former is likely to need to take primacy.

As such, evaluations of CWSIs are likely to be among the longest evaluations funded by the EEF. We envisage evaluations that could approach, or even exceed, five years in duration from setup to outcome measurement. As noted in Section 2, this is a key reason for the consideration of QED evaluations, rather than RCTs, because of the difficulty in maintaining a control group over this extended period.

Nevertheless, it is important to prevent this from going beyond what is necessary. Even if a QED impact evaluation is adopted, longer evaluations increase risk because of the increasing risk of non-compliance by treated schools the longer the evaluation goes on.

3.2.2 Attrition from an intervention

Key Message 6: Adopt measures to increase retention
Measures discussed in Section 3.2.2 should be considered to increase retention of schools.

Key Message 7: Investigate attrition as part of the IPE
The IPE (discussed further in Section 5) should explore why pupils/teachers/schools drop out and whether this was related to the intervention.

Attrition from an intervention can occur either before the trial or at any point during it. If it is not accompanied by attrition from data collection, it does not directly threaten the internal validity of the trial but does make it less likely that we will observe a significant treatment effect.

Before the intervention delivery starts, attrition is most likely to occur because schools lose interest. This is a particular risk if long lead-in times are required for recruitment and set-up. Losing schools during this period clearly poses issues for the project (including for statistical power and external validity) but, if it occurs prior to randomisation, does not directly pose a risk to the trial's internal validity. Furthermore, using this period productively to ensure schools are fully committed to the project and aware of the requirements of the evaluation can help to reduce the risk of attrition once the intervention delivery begins. This is particularly important in the case of CWSIs, given the particularly large commitment required to participate fully.

Throughout the trial, schools may cease the treatment due to a loss of enthusiasm among staff, key members of staff supporting an intervention leaving, or new teachers arriving and not being trained. Attrition from treatment might be avoided by using process monitoring data in real time, or close to real time, to determine which schools are dropping off from treatment. Specifically, if there are materials that need to be accessed regularly for treatment, then project teams can monitor which schools are, or are not, logging onto the resource website, and proactively reach out to the latter to offer support or refresher training courses.

We recommend considering the following measures to increase retention:

- allow longer lead-in times for on-boarding schools, perhaps as much as six to nine months before trial launch;
- plan a recruitment, communication, and engagement strategy to mitigate the risk of long lead times (see below); and
- if possible, build in real-time monitoring of process data to allow targeted support of at-risk schools.

Given the challenge that attrition may well still pose, we also recommend investigating it as part of an IPE. Although finding out why does not help us correct for attrition during the trial, it can help us interpret and understand outcomes. This will help evaluators understand potential unintended effects of the intervention, and may help improve the intervention if rolled-out further.

3.2.3 Attrition from data collection

Schools dropping out from data collection directly threatens the internal validity of the trial since it undermines attempts to estimate a true intention-to-treat effect. This issue is significantly reduced where administrative data is used as an outcome and this should be considered as part of the design process. However, this will not be right for every intervention. Ultimately, the other messages in this section, along with existing EEF guidance, are likely to be the best ways to minimise this issue.

3.2.4 Recruitment, communication, and engagement

Key Message 8: Agree a combined recruitment, communication, and engagement strategy
Project and evaluation teams should agree and implement a combined recruitment, communication, and engagement strategy during recruitment and into the period of implementation. This should include the elements discussed in Section 3.2.4.

Existing EEF guidance notes that a project communication plan may be useful. In the case of CWSIs, we see a combined recruitment, communication, and engagement strategy as vital to the successful completion of the evaluation. After theorising and designing the trial, but before recruitment begins, the delivery and evaluation teams should develop this jointly. It should include:

- close collaboration between delivery and evaluation teams during the recruitment process, but with clear areas of responsibility to avoid unnecessary duplication or things 'falling between the gaps';
- a clear explanation to schools of what will be involved in both the project and the evaluation: CWSIs are often time and resource intensive for schools to implement and to co-operate with evaluators; and
- a clear outline of the responsibilities of the schools, the delivery team, and the evaluation team: duration and intensity of evaluation activities should be clearly listed.

The costs, in terms of time and effort, associated with evaluation should not be underestimated. Schools should be aware when, and in what way, the evaluation team needs access. Especially where the delivery team routinely implements their 'product' in schools, they may not be aware of the additional responsibilities that come with careful evaluation.

Having a member of the evaluation team present during discussions with schools will:

- facilitate a clear and transparent agreement between schools and the project team;
- allow the research team to identify schools that will be good implementation partners; and
- allow the evaluation team to become familiar with the specific contexts of participating sites: this will help guide their thinking about how, when, and where to conduct the IPE.

As noted above, communication is essential to reduce attrition and helps evaluators sustain a longer lead time prior to the launch. Once the evaluation is underway, it may be helpful to provide each participating school with a single, named point of contact for all queries about the project (whether intervention or evaluation) to help develop a strong relationship. This person would then direct queries to the appropriate person on the evaluation or project team for action.

In summary, an effective recruitment, communication, and engagement strategy for evaluating CWSIs includes:

- assigning one contact person to each school to develop the relationship;
- having one line of communication between the delivery and evaluation teams, arranged as part of the communication strategy; and
- communicating timelines to schools at recruitment stage, being clear about what is expected and what schools can expect in return.

3.2.5 Recruitment and school readiness

Key Message 9: Assess school readiness to participate during recruitment
Use the checklist in Appendix 4 to ensure school readiness to participate in the intervention and evaluation.

In an ideal world, we would like to recruit a random sample of schools to participate in evaluations. However, in reality, it is important to narrow this to schools with a high

probability of completing the intended intervention. As such, when engaging with potential participants, it is helpful to chart their organisational readiness to implement a CWSI.

Ultimately, implementation depends on the receptivity of participating schools. Although it is difficult to assess a school's likelihood of implementing an intervention with high fidelity before launching the trial, it is important to be mindful of the factors that are relevant to the intervention delivery system.

In initial conversations with schools, determining whether the senior management and teachers have a shared vision and agree on the value and purpose of the intervention, is instructive. Furthermore, cultural norms regarding change will likely determine whether the intervention succeeds. When the organisation is open to change in a broad sense, and its staff is willing to try new approaches, a new intervention is more likely to be perceived as acceptable and to be integrated into praxis.

A checklist to assess school readiness to participate in evaluations of CWSIs is included in Appendix 4. Schools should meet most, if not all, of the criteria to be considered ready to participate in the evaluation.

4: Quasi-experimental impact evaluation for complex whole-school interventions

Although Section 3 argues that it is often possible to use cRCTs to estimate the impact of CWSIs, this will not always be the case (see Section 2). In cases where an RCT is infeasible or undesirable, a well-designed quasi-experimental evaluation design (QED) may be able to provide an impact estimate where one would otherwise simply not exist (Petticrew *et al*., 2005), or give a more credible impact estimate than from an RCT with implementation issues.

This section discusses important considerations for conducting a QED impact evaluation of CWSIs, including:

- considerations when planning a QED impact evaluation;
- discussion and exposition of one particular method of QED impact evaluation (matched difference in differences) that we argue is likely to be particularly useful in the context in which EEF might fund evaluations of CWSIs; and
- additional requirements for analysis and reporting of a QED impact evaluation.

4.1 Planning quasi-experimental evaluations

4.1.1 Pre-planned vs. natural QED impact evaluation

Key Message 10: Conduct only pre-planned QED impact evaluation
Where it pursues a quasi-experimental approach to evaluation, the EEF should concentrate on planned evaluations, rather than 'natural' experiments. This will guard against their use in settings where they are unlikely to produce credible estimates of impact.

There is a well-developed literature on estimating the impact of a policy change or intervention through QEDs, often making use of 'natural experiments'[2] rather than pre-planned evaluations of the type that the EEF is familiar with (Craig *et al*., 2012). This chapter discusses QEDs and how to design evaluations that make best use of these methods to provide robust estimates of impact.

Where these approaches are opportunistically applied to observational data ('natural experiments') it is sometimes questionable whether the assumptions needed for the QED to produce unbiased causal estimates are justified. However, this does not mean that the principles underlying the methods themselves are fundamentally problematic. If planned and carefully applied in cases where relevant assumptions are justified, it is possible to provide credible impact estimates.

[2] There are differences in the literature about what is generally meant by the term 'natural experiment' (Craig *et al*., 2012). Here we have in mind evaluations using pre-existing observational data in which the analyst identifies and exploits natural sources of variation to estimate causal impact (Petticrew *et al*., 2005, p. 752).

4.1.2 Intention-to-treat analysis in quasi-experimental evaluation

Key Message 11: Always use intention-to-treat analysis

Analysis of QEDs should be carried out on an intention-to-treat (ITT) basis to maximise comparability with this approach for RCTs.

Key Message 12: Identify a 'quasi-randomisation' date

To ensure there is clear identification of the analysis sample, the evaluation and project teams should agree a 'quasi-randomisation' date at which the sample of intention-to-treat schools is finalised.

All EEF evaluations are currently designed to recover intention-to-treat (ITT) effects because the sample is determined at the start of the evaluation at randomisation. It is important to consider what such an estimator looks like in a quasi-experimental evaluation. The most basic requirement remains to find comparators, and estimate results, using all schools who begin the intervention, not—

- all schools who express interest but are not fully signed up (who would drop out before randomisation in an RCT); or
- only schools who fully complete the intervention (since this would only be an on-treatment analysis).

To this end, we suggest that in QEDs the evaluation and project teams agree in advance a cut-off date analogous to the randomisation date in an RCT. This could be thought of as the 'quasi-randomisation' date. All schools and only schools who have completed all necessary pre-intervention steps by this date should be used as the basis for the evaluator's construction of a comparison group (i.e. completion of pre-intervention steps by the quasi-randomisation date should be treated as both a necessary and sufficient condition for inclusion in the treatment group).

In the case of an RCT, randomisation provides the basis for assuming that control group drop-out, had they been treated, would have been similar to that seen in the treatment group, making this a valid comparison. In the same way, when constructing a comparison group for all schools that we intended to treat (not just those that were actually treated), we are appealing to the identifying assumptions of the chosen QED to justify the implicit assumption that we would have seen a similar level of treatment drop-out from our comparison group. The plausibility of this assumption is likely to be similar to that of the QED approach more generally.

4.2 Approaches to quasi-experimental evaluation

Key Message 13: Consider matched difference in differences

We recommend that the EEF considers the use of a matched difference in differences design for future projects where RCTs are not feasible. It will not be suitable for all projects and nor is it the only possibility. The decision should be taken in conjunction with project and evaluation teams.

This section primarily focuses on one approach to QED impact evaluation: matched difference in differences. This is not because there are not others approaches that may be useable to evaluate certain CWSIs of interest to the EEF. However, we maintain this focus because we think this will often be an applicable and robust approach in this context.

4.2.1 Matching methods

Key Message 14: Selecting a preferred matching approach
We do not advocate one specific approach to matching but rather transparency about the approach used and ample robustness checking of that approach (see below).

Key Message 15: Robustness check the preferred matching approach
Evaluators should test the robustness of their findings to making different decisions about the matching approach. We suggest at least five robustness checks should be pre-specified at the design stage and the results reported in the evaluation report. We should not have much confidence in findings that are fragile and strongly dependent on the specific method used.

Key Message 16: Use a principled approach to selecting matching variables
Use the principles identified in Box 4.1 to select variables for inclusion in the matching model. It is especially important to ensure that matching variables could not have been affected by the treatment itself.

Key Message 17: Explore and, if necessary, impose common support
Failure of common support undermines the comparability between treatment and matched comparison samples. It should be explored and, if necessary, imposed as part of analysis.

One approach to QED impact evaluation is the use of matching methods. Rather than comparing individuals with all untreated individuals, matching restricts with whom individuals in the treatment group are compared.

This can be done in various ways (see below), but the aim is always to end up comparing treated schools with untreated schools that were, nevertheless, just as likely to have received the treatment. The intuition is that once we are comparing treated schools with untreated matched comparator schools that were equally likely to receive the intervention, these two groups of schools will have the same expected outcomes in the absence of treatment and, as such, any difference in their outcomes after the intervention has occurred is attributable to the treatment.

The strong assumption that is implied is that schools who are matched do genuinely have the same probability of ending up in the treatment group. In other words, a treated school and the school(s) with which it is matched would be expected to have the same outcomes in the absence of the treatment. Since we can only match on observed characteristics, the risk is that unobserved or unobservable differences could undermine this assumption.

Approaches to matching

There are many options available for matching schools in treatment and comparison groups. These range from relatively simple to more complex approaches, many of which can also be combined (for example, 'nearest neighbour' with 'caliper'). Five commonly used approaches are discussed in Table 4.1.

Table 4.1: Approaches to constructing a matched comparison group

Matching approach	Description	Notes
Nearest neighbour	Each treated school (or individual)	• Can be carried out 'with

	is matched with one school (or individual) in the comparison sampling frame whose propensity score is most similar.	replacement', where an untreated school can be used as the comparator for more than one treated school, or 'without replacement' where once a school has been selected as a comparator it is no longer available to be selected. • 1:1 matching is described here, but a larger number of matched untreated schools can be chosen per treated school.
Radius/Caliper	Treated schools (or individuals) are compared with an average of all schools (or individuals) in the comparison group whose propensity score is within a certain distance either side of the treated school's propensity score.	• The choice of the width of caliper involves a trade-off between bias (a narrower caliper limiting bias but potentially reducing the number of matches) and precision (a wider caliper interval that raises the possibility of bias but increases the number of potential matches). • Commonly combined with nearest neighbour to ensure that the nearest neighbour is only selected if it is not too different from the treated school.
Kernel	Treated schools (or individuals) are compared with a weighted average of all schools (or individuals) in the comparison group; weights depend upon how different the two propensity scores are.	• Weighting is determined by different kernel functions e.g. Epaneknikov.
Exact	Each treated school (or individual) is matched with comparison schools (or individuals) which precisely share all the characteristics on which matching is based.	• Often used for only a small number of very important characteristics and combined with other matching methods.

Coarsened exact	Allows the analyst specifically to set a tolerance around exact matches of certain characteristics (particularly important with continuous variables where truly exact matches are unlikely) allowing ex-ante specification of balance rather than ex-post testing.	• A variant of exact matching.

We are reluctant to recommend any particular method since this depends on the nature of the evaluation. No one method will perform best in all circumstances. For this reason, we recommend that evaluators should test the robustness of their findings to making different decisions about the matching approach.

We suggest that a minimum of five robustness checks of alternative matching specifications should be conducted. This is just a rule of thumb based on the number of alternative specifications observed in previous trials in this field (for example, Allen and Allnutt, 2013, report seven; Alcott, 2017, only reports three) and should be driven by sensible alternatives that could have been chosen when choosing the preferred approach. Alterations to the approach might include:

- a different approach to matching entirely (for example, kernel rather than coarsened exact);
- changes to the number of nearest neighbours selected;
- changes to the width of the caliper;
- changes to the type of kernel used; and
- changes to the tolerance in coarsened exact matching.

Box 4.1: Principles for the selection of matching variables

- Variables used for matching should be ones than cannot have been affected by the intervention itself (Bryson *et al*., 2002, p. 35). Although this does not entirely rule out using variables collected after the intervention begins (since we could match only on individuals' fixed characteristics), in general we recommend the use of variables that are based on data collected before the intervention begins and, ideally, before schools are even approached about participation.
- Following previous evaluations in this context, we suggest considering matching on the basis of proportions of pupils in the following categories—those eligible for pupil premium, with special educational needs, having English as an additional language (EAL), of White British ethnic background, or female—other characteristics might include the average performance of pupils in national examinations (ideally at intake and outflow), school type, and region.
- Ideally, the variables used for matching should be guided by theory and previous evidence, i.e. what do we know that leads us to think that certain schools are likely to want to participate in evaluations of this type? We have provided some suggestions above but it may be productive to commission further work that explores the characteristics of schools (including their composition) in previous EEF evaluations in several broad categories. Such work should consider the issue of overfitting and consider techniques such as machine learning to provide a robust set of predictors of treatment and outcomes. Wong *et al*. (2017) suggest piloting an evaluation specifically for the purpose of determining important factors that predict participation and outcomes.
- A small number of exact matching criteria may be warranted if there are reasons to be concerned about effects that would differ significantly across types of school, such as academies vs. maintained schools, single sex vs. coeducational schools. In a similar vein, matching within region may be important (Wong *et al*., 2017) to achieve good balance on ethnicity and EAL composition (Allen and Allnutt, 2013, p. 12).
- Matching on pre-intervention trends in results may help to capture changes taking place at the school that might otherwise undermine the common trends assumption (Allen and Allnutt, 2013, p. 13; Cook, Shadish and Wong, 2008). This might also help to deal with issues of differential school motivation, otherwise incredibly difficult to capture as part of a matching approach, with pre-intervention trends in results acting as a proxy for school motivation.

Selecting variables on which to match

Previous evaluations have explicitly applied matching approaches to the NPD (specifically matched difference in differences, see further details below) to estimate, for example, the impact of Teach First (Allen and Allnutt, 2013), the provision of free school meals for all pupils (Kitchen *et al*., 2013), and a school building programme on academic outcomes (Thomson, 2016). Although the variables themselves differ, we use these previous studies and other sources of guidance (such as Bryson *et al*., 2002; Caliendo and Kopeinig, 2008) to develop a list of key principles regarding the selection of matching variables.

Common support

An important issue when estimating impact using a matching approach is exploring the extent to which there is overlap between the treatment and comparison groups in terms of the propensity to be selected for the intervention. This is known as 'common support'. Failure of common support undermines the comparability of the matched comparison group that is assembled.

To illustrate the issue, consider the case of simple 'nearest neighbour' propensity score matching with replacement. If this were to be conducted without considering common support, and there are large numbers of treated schools that have higher propensity scores than the untreated school with the highest propensity score, they would all, nevertheless, be matched against that one untreated school with the highest propensity score, no matter how much more likely to receive the treatment they were themselves.

This will be far less of a problem when individuals in the treatment and comparison groups have similar distributions of propensity scores. However, where it does occur, it is common practice to impose common support by dropping treated observations with a higher propensity score greater than the highest propensity score of untreated observations and dropping untreated observations with a smaller propensity score than the lowest propensity score of treated observations. We then estimate impacts concentrating on the sub-sample where we have seemingly comparable treatment and comparison schools.

When matching approaches produce credible estimates

Matching methods are most likely to be effective where untreated schools could easily have been treated. This is most likely where decisions to participate in the programme are in some way arbitrary, for example being based on proximity to where a project team has available trainers. A checklist to assess the plausibility of producing credible estimates using this approach is provided in Appendix 5. In general, all criteria should be met to support use of this method.

Matching approaches are more likely to be effective when evaluating programmes where:

1. informational constraints mean a large number of schools simply haven't heard of the intervention. The less well-known a programme, the more likely our untreated comparator schools would have signed up if they had heard of it and, thus, be better comparators; and
2. there are a number of potentially interchangeable programmes and we want to consider their relative effectiveness, while the purchase decision isn't closely related to the characteristics of the school so can be considered as good as random for the purposes of evaluation (e.g. Read Write Inc.).

By contrast, they are much less likely to be considered credible:

1. when evaluating well-known programmes promoted by high-profile organisations, since these may have already been considered (e.g. PiXL Club); or
2. where taking part has major effects on the organisation of the school and, thus, is much too important a decision to be considered random (e.g. School Direct; Teach First).

4.2.2 Difference in differences

Another approach is the use of 'difference in differences', essentially comparing trends in treated schools' performance with such trends in untreated schools. While difference in differences evaluations can make use of individual-level data, the intuition is clearer at the school-level.

To understand this method, it is helpful to consider a simpler version, known as the 'before-after approach. In this approach, we simply use treated schools' past performance as our comparator: our estimate of the treatment effect is the difference between their performance before and after its introduction. This makes the strong assumption that any change in a school's performance between the period before an intervention started and the period afterwards (beyond that we might expect from random variation) is caused by the intervention. However, we might well think that this is too strong an assumption. For example, changes in schools' performance might just reflect national trends in results.

The robustness of this method is enhanced by combining it with a comparison group; this is known as a difference in differences approach. Rather than just using the change in treated schools' performance, we compare the change in treated schools' performance with the change in untreated schools' performance over the same period.

When difference in differences produces credible estimates

Comparing the assumptions behind difference is differences and matching methods is instructive. A key advantage of the former is that it does not rely on the assumption that all the differences between treated and comparison schools (or individuals) are explained by the observable characteristics that we have available to us. Instead, we can rely to some extent on unobservable differences between these two groups being absorbed by the process of focusing on differential trends over time rather than outcomes directly.

It is, therefore, important that such interventions do not coincide with other school initiatives or changes with the potential to produce similar effects, for example the appointment of a new headteacher. A checklist to assess the plausibility of producing credible estimates using this approach is provided in Appendix 5. In general, all criteria should be met to support use of this method.

Difference in differences is more likely to be effective when evaluating interventions where:

1. there is a defined group of untreated schools that we would expect to be experiencing similar trends in our outcome of interest across the period of analysis; and
2. treated and comparison schools are similar in terms of observable characteristics, since this is likely to be associated with common trends in outcome variables.

By contrast, they are much less likely to be considered a credible approach when:

1. implementation is typically contemporaneous with other changes occurring in the school, since the impact estimate will pick these up as part of the treatment effect; or
2. there is reason to believe that trends in outcomes between treatment and comparison schools are diverging for reasons unrelated to the treatment.

4.2.3 Matched difference in differences

Matching and difference in differences approaches may be combined to leverage advantages from each (Stuart *et al*., 2014). This has previously been used in educational settings, for example to estimate the impact of employment while in school on educational outcomes (Buscha, Maurel, Page and Speckesser, 2012).

It is less demanding than simple matching because it is only *trends* that need to be the same between the treatment schools and the matched comparison schools, rather than *levels*. Likely for this reason, there is evidence that it performs better than simple matching approaches at replicating experimental results (Smith and Todd, 2005).

The basic idea of this hybrid approach is to:

- use the process of matching to select a comparison group of schools that is more likely (as noted above) to have the same trends in outcome variables as our treatment group, then
- apply the difference in differences approach to this panel of schools.

More complex hybrids are also possible, for example matching at both school and individual level so that we only compare the difference in performance between matched individuals in treatment and comparison schools after implementation with the difference in performance between matched individuals in treatment and comparison schools before intervention.

The National Pupil Database (NPD) is well suited to use in matched difference in differences evaluations, even though it may not be particularly suitable for either approach in isolation:

- Simply comparing the trends in treated schools to all other schools is still likely to suffer from bias due to the presence of schools with highly different characteristics.
- In addition, administrative data typically lacks rich enough background variables to conduct matching in isolation.

However, the large dataset means that we have a large pool of potential comparison schools and sufficient school-level characteristics to identify matched schools that are likely to experience common trends in our outcome of interest.

When matched difference in differences produces credible estimates

To produce credible impact estimates, matched difference in differences requires some of the assumptions from each of its elements (matching and difference in differences) but

these are able to rely on one another. For example, we noted above that difference in differences is likely to be more plausible when the treatment and comparison groups are similar in terms of observable characteristics: matching helps to ensure that this is the case. Likewise, matching benefits from being able to use a school's previous performance to deal with unobserved differences.

A checklist to assess the plausibility of producing credible estimates using this approach is provided in Appendix 5.

Matched difference in differences is more likely to be effective when evaluating interventions where:

1. informational constraints mean a large number of schools simply haven't heard of the intervention. The less well-known a programme, the more likely our untreated comparator schools would have signed up if they had heard of it and, thus, be better comparators.

By contrast, they are much less likely to be considered a credible approach when evaluating interventions where:

1. implementation is typically contemporaneous with other changes occurring in the school, since the impact estimate will pick these up as part of the treatment effect; or
2. there is reason to believe that trends in outcomes between treatment and comparison schools are diverging for reasons unrelated to the treatment.

4.2.4 Other QEDs

We have not discussed here the many other QEDs that exist—such as Synthetic Control methods and Regression Discontinuity Designs (RDDs)—but this does not mean that their use will never be appropriate. Evaluators should still be able to advocate for their use if they seem appropriate for a particular setting, however we judge that they are less likely to be generally applicable to the evaluation of CWSIs than the approaches we have advocated.

4.3 Implementation, analysis, and reporting

Key Message 18: Match blind to outcomes
Matching should be carried out blind to outcomes. Best practice would be to carry out matching at a roughly analogous time to when randomisation would be carried out in an RCT to ensure that this is the case.

Key Message 19: Maximise transparency of matching process
The matching process should be made as transparent as possible to ensure the credibility of findings and allow for replication. Recommended reporting requirements are outlined in Section 4.3.

There are some important differences in implementation, analysis, and reporting in QEDs. We describe these here. We also include an outline Statistical Analysis Plan (SAP) for our recommended approach of a matched difference in differences evaluation, in Appendix 3.

4.3.1 Timing and transparency of matching process

An important question that does not arise in RCTs concerns at what point, and under what conditions, matching should be carried out. It is important that matching should be carried out:

- blind to outcome variables; and
- ideally, at an analogous point in the process to randomisation (that is, ahead of implementation).

The impact estimates from matching analysis are more dependent on decisions made by the analyst than is the case for randomisation (meaning it is easier to make changes to the specification inadvertently or deliberately that affect the likelihood of finding a positive outcome). As such, it is particularly important that there is no potential for it to be claimed that the process has been manipulated to increase or reduce the impact estimates.

The matched sample chosen should be reported to the EEF (but not published) and full details of the approach adopted reported in a SAP (see Appendix 3).

4.3.2 Statistical inference in difference in differences

Key Message 20: Use appropriate statistical inference with difference in differences
Statistical inference in difference in differences analysis should be conducted in line with the principles highlighted by Brewer, Crossley and Joyce (2013).

Difference in differences has its critics, such as Bertrand, Duflo and Mullainathan (2004), who note that, compared to randomised designs, there is a risk that it might lead us to estimate seemingly statistically significant effects where none are present. This stems from serial correlation within clusters.

However, others have demonstrated that this risk is significantly reduced when appropriate statistical inference is used (Brewer, Crossley and Joyce, 2013). The use of appropriate statistical inference is, therefore, important. Brewer *et al.* highlight the importance of accounting for serial correlation. In many cases this can be achieved by calculating cluster-robust standard errors across time. Where there are a relatively small number of clusters (below roughly 42) alternative procedures based on cluster bootstrapping may give more reliable inference.

Depending on one's view of the purpose of matching (Ho *et al*., 2007), it is often also seen as adding complexity to statistical inference. In the past, a common way around this was to use bootstrapping techniques, however this approach may not be valid with all approaches to matching (Abadie and Imbens, 2008). Adjustments for simple nearest neighbour matching are currently only available for limited settings (Abadie and Imbens, 2011). As such, the potential implications for confidence in our estimates of the approach to statistical inference taken by the evaluator should be reported.

4.3.3 Balancing checks on observables

Key Message 21: Conduct balancing checks
Balancing checks on observables, as currently reported for RCTs, should also reported in this setting. This should be done for, at a minimum, all characteristics on which schools have been matched.

In RCTs, it is common practice to check the balance of observable characteristics between the treatment and control groups, both after randomisation and of the final analysis sample. This is even more important in the context of QEDs where lack of balance is more likely to occur. At minimum, balance of all characteristics included in the matching model should be checked. Despite the difference in differences element of the design, our confidence in the results of the evaluation should be considerably reduced if we have not achieved strong balance on observables.

This may be helpfully done using standardised differences rather than significance testing of differences (Imbens and Rubin, 2015, ch. 14) especially in this setting of using NPD data where large sample sizes may give us the sense that a small difference in characteristics is, nevertheless, statistically significant. Our confidence in studies where there are absolute standardised differences greater than 0.05 in important observable characteristics should be significantly reduced. This adopts the same definition and threshold as that for randomised experiments.

Nevertheless, we should remain mindful of the fact that there may still be imbalance on unobserved or unobservable characteristics that could be important for predicting participation in the intervention and our outcomes of interest, the identifying assumption matching ('unconfoundedness') requires to produce an unbiased estimate of the treatment effect.

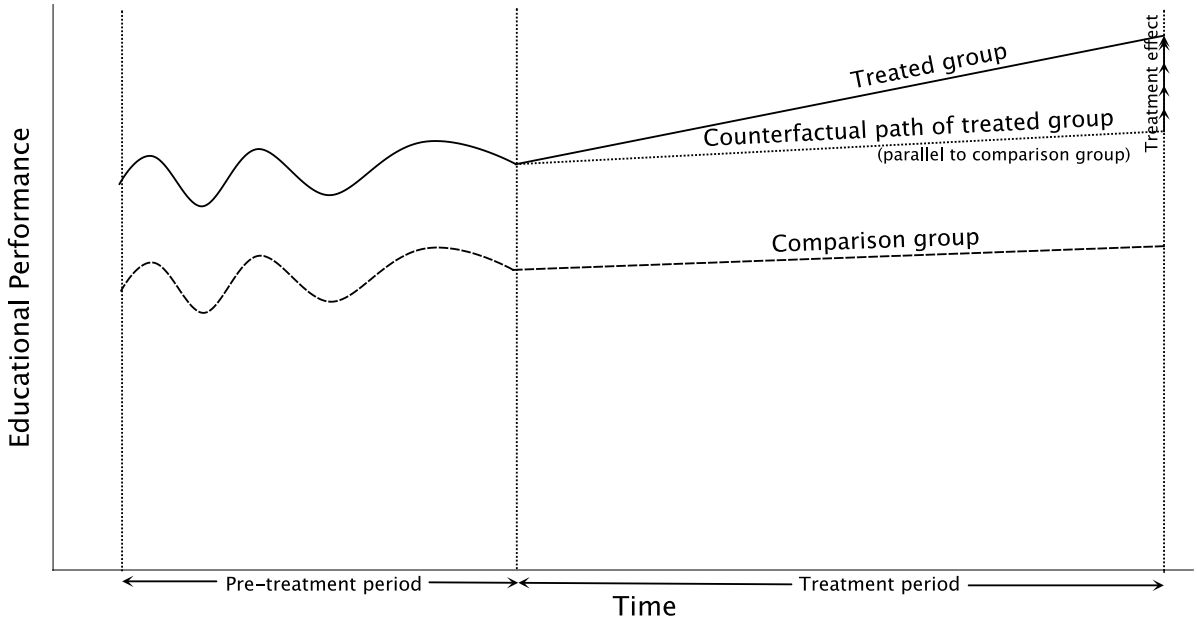
4.3.4 Analysis of common trends

Key Message 22: Analyse and report pre-treatment common trends
When using difference in differences, analysis of pre-treatment common trends in the outcome variable between the treated group and the matched sample should be reported. This should include both placebo tests of pre-treatment impact and graphical plots of pre-treatment trends in outcome variables among treatment and comparison groups.

As described above, the fundamental idea behind the difference in differences approach is to compare the change in schools' performance over the course of the programme with the change experienced by other schools. The robustness of this method rests on the assumption that the change in the schools where the programme was implemented would have been the same as that seen in the comparison group schools, had the programme not been operated there.

This is known as the 'common trends' assumption. Figure 4.1 provides an illustration of outcome measures with common trends in pre-treatment outcome measures. Failure of the assumption could result in over- or under-estimation of the effect of intervention.

Figure 4.1: Illustration of intuition underlying difference in difference evaluation



The common trends assumption is less demanding than the assumption inherent in the 'before and after' approach, but nevertheless requires justification. It cannot be tested directly because we do not know what would have happened to the treatment schools if the intervention had not been implemented.

However, assembling several years of pre-intervention trends in the outcome variable in treatment and comparison schools provides useful evidence on whether it is likely to be justified and, hence, how much confidence we should have in our approach. This can be done by:

- graphical inspection of trends in the data among treatment and comparison groups; and
- more formally using 'placebo' tests in which we estimate the 'impact' during the pre-treatment period: rejecting the null hypothesis of no impact in a placebo test should significantly reduce our confidence in the actual impact estimate.

Of course, we should acknowledge that common trends in the pre-treatment period does not guarantee that this would be maintained into the treatment period in the absence of the treatment, as the identifying assumption of difference in differences requires to produce an unbiased estimate of the treatment effect.

4.4 Second round recruitment

We touched above on the issue of differential school motivation and how matched difference in differences may be able to address this to some extent, by allowing matching on pre-trends. One way of identifying similarly-motivated schools would be to launch a "second-round recruitment"–identical to the first round–at the end of the initial treatment procedure, once the evaluation begins. This could either ask schools if they would, in principle, like to participate in a future round of the evaluation. A positive response could be used as a criterion for inclusion in the comparison group or could form a second round of schools for intervention. This latter approach would prevent use of this as group for long-term follow up, but would provide a comparison group that have followed through the process from recruitment to starting implementation.

While this seems attractive, it still poses challenges. In particular, motivation is not a factor that is constant over time. Just because our 'second round' recruits are motivated by the intervention at the time of second recruitment does not mean that they were similarly motivated at the time of the initial recruitment. This reduces our confidence in the assumptions behind the evaluation method (for example, of balance on important observables and unobservables in the case of simple matching and/or that there will be common trends between the two groups in the case of difference in differences) that we then go on to apply. The closer in time between the two rounds of recruitment the less likely this would be to raise a concern, although this clearly limits the length of time over which we would be estimating the impact of the intervention. Regardless, it remains important to undertake appropriate tests and robustness checks for the method being applied.

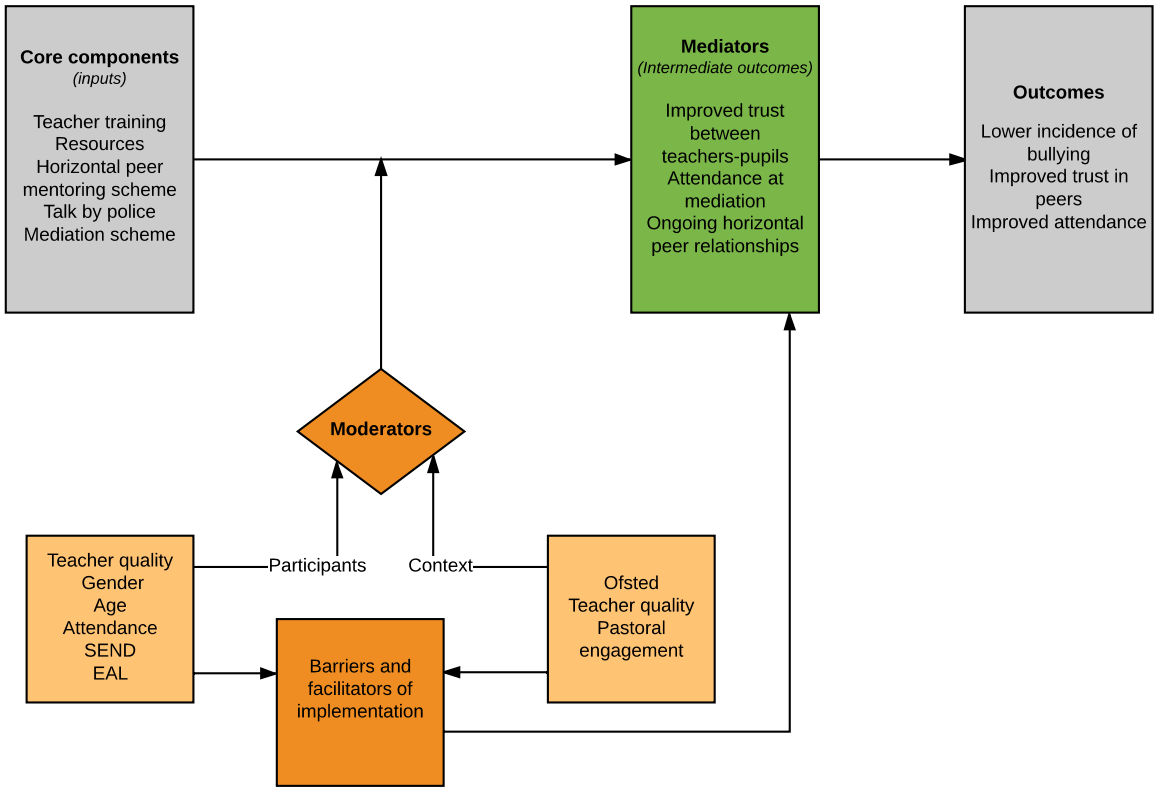
5: Implementation and process evaluation (IPE) for evaluations of complex whole-school interventions: moderators, mediators, and outcomes

In Sections 2 and 3, we highlighted experimental and quasi-experimental designs that can be used to evaluate CWSIs. This section outlines how evaluation of CWSIs should be enhanced by being embedded within a multi-phase mixed methods IPE (including a combination of additional quantitative and qualitative evaluation methods, incorporating multiple data collection points) to help explore not just 'what works', but 'what works, when, and for whom' (Bonell *et al*., 2012; Moore *et al*., 2015).

Much of this discussion builds directly on the EEF's guidance for Implementation and Process Evaluation (IPE) by Humphrey *et al*. (2016a). Here we emphasis points that are particularly important or relevant in the context of CWSIs. In the case of CWSIs, a fully integrated mixed-methods approach to evaluation is especially important, to the extent that we may not wish to think of there being a 'separate' IPE.

We posit a multi-phase mixed-methods approach to support evaluators to work within the complex whole-school context, using qualitative data to refine RCT or QED implementation or outcome measures. We identify three key elements that should be addressed in an IPE: a clear focus on (1) core components, (2) moderators, and (3) mediators. A fictional example of how these stages interact is outlined below in Figure 5.1.

Figure 5.1: Example logic model: reducing bullying in schools



These elements are important to consider throughout the design and implementation stages and should be integral to the evaluation (see Humphrey *et al*., 2016a, p. 3). For CWSIs, it would be useful to map out these three factors in conjunction with conducting a

TIDieR framework (Humphrey *et al*., 2016b, p. 19). We consider these at design stage and then provide suggestions to help measure these during implementation.

5.1 IPE design

5.1.1 Overall approach to IPE

Key Message 23: Conduct multi-phase mixed methods IPE
Use a multi-phase mixed methods IPE design to ensure findings are maximising the efficiency and accuracy of the evaluation. The order of the quantitative and qualitative data collection and analysis should be fully rationalised in the evaluation protocol.

Humphrey *et al*. (2016a, p. 8) state that the use of mixed methods in IPE should be standard. We see no reason to diverge from this view for the evaluation of CWSIs. Building on this, Humphrey *et al*. (2016a, p. 13) identify four mixed-methods designs that may be useful frameworks for an EEF-funded IPE:

- convergent parallel;
- explanatory sequential;
- exploratory sequential; and
- multi-phase (also referred to as 'embedded').

We characterise each of these four approaches in Figure 5.2, highlighting the strengths and limitations of each for the evaluation of CWSIs (Creswell, 2013). The order in which the IPE collects and analyses both the qualitative and quantitative data will affect what can be taken from the findings. As there is no 'one size fits all' approach to IPE, it is recommended that evaluators discuss the order in which they combine their data to ensure it is most appropriate for understanding the CWSI.

Based on this assessment, we advocate the use of multi-phase approaches for the evaluation of CWSIs. For similar reasons to our advocacy of using multi-stage protocols (see Section 2), we recommend the use of this approach because:

- of the importance of being able to use early findings from IPE to inform later analysis as the evaluation continues; and
- the likely length of evaluations of CWSIs allows the truly iterative process needed for this approach across the duration of the evaluation.

5.1.2 Timing of IPE

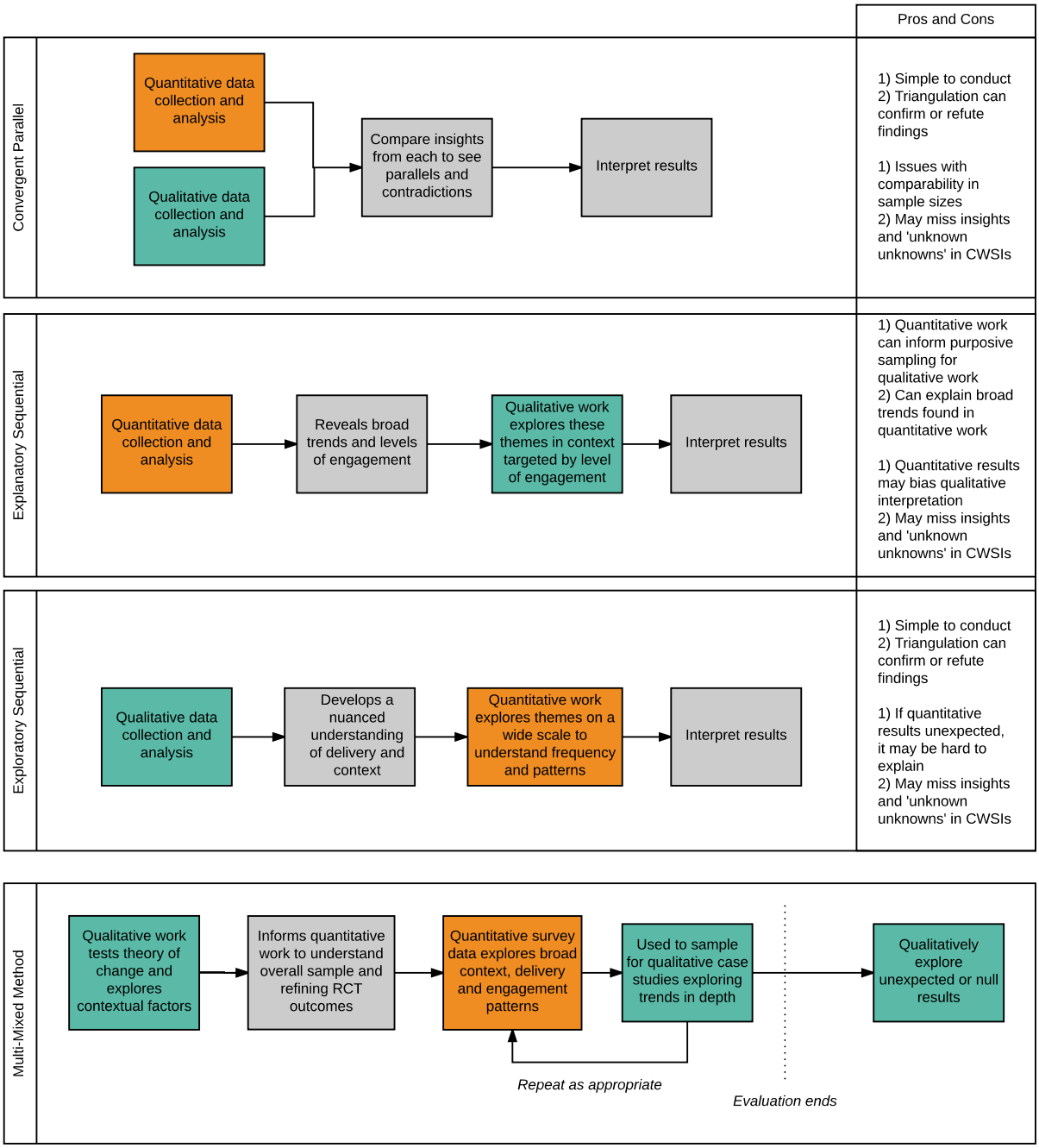
Key Message 24: Consider commissioning a follow-up IPE
In some cases, there may be value to commissioning a follow-up IPE where there are unexpected results from the impact evaluation, to help understand these findings.

Critiques have found that conducting IPE before the RCT outcomes can, if the result is found to be surprising, be of little help (Munro and Bloor, 2010), especially if the qualitative inquiry ran on the assumption of a positive effect. It is hoped that the multi-stage protocol process (Section 2) will reduce the instance of wholly unexpected findings from impact evaluation.

Nevertheless, this is still likely to occur in some instances, given the 'need to anticipate the results from the outcome evaluation' to make 'decisions on the topical foci of the process evaluation data collection' (Munro and Bloor, 2010, p. 709). Although, in general, knowing the findings of the impact evaluation may bias findings from IPE, there may be a case for commissioning follow-up IPE to understand findings more fully. It is important to invest in

understanding why something did not work as expected in order to learn as much as possible from the evaluation and inform future research and investment.

Figure 5.2: Mixed-methods designs for the evaluation of complex whole-school interventions



5.2 Developing a logic model

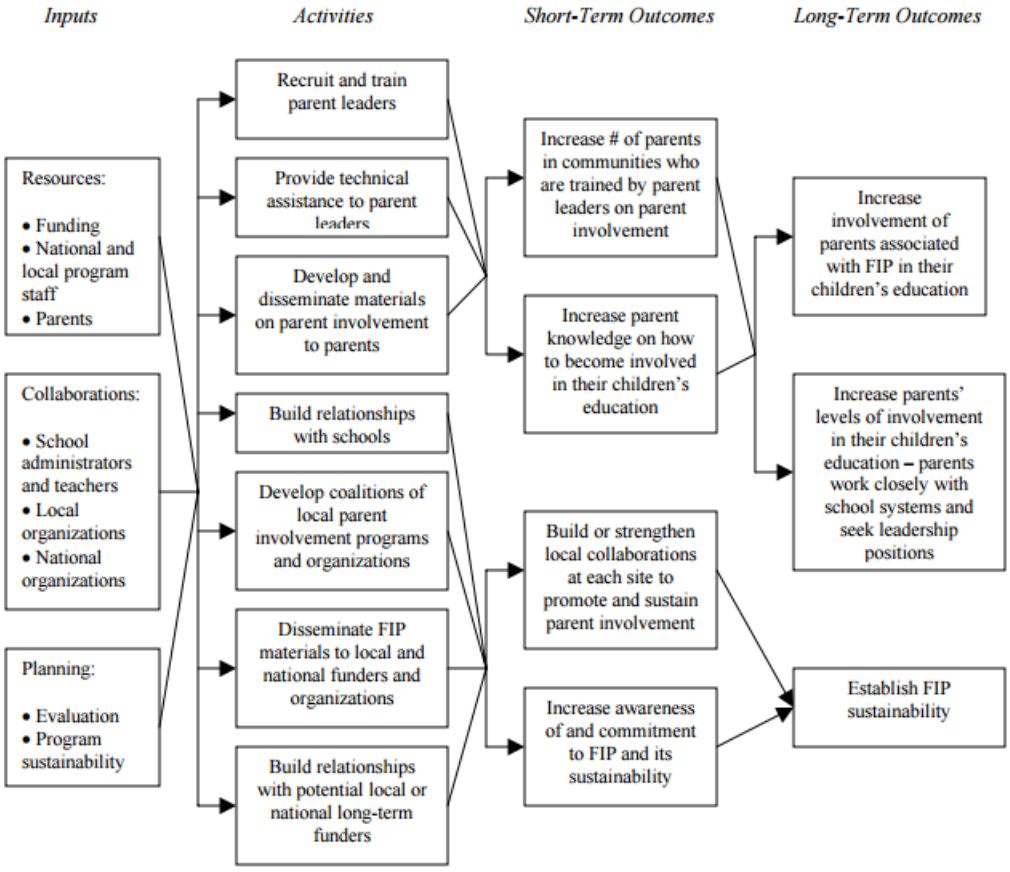
Key Message 25: Develop and continue to update a logic model
Developing a model of the logical relationships between the core components, moderators, mediators, and outcomes of a programme is particularly important in the context of CWSIs. In line with Key Message 2, the logic model should not be a static product, rather it should be updated as part of a multi-stage evaluation protocol approach.

The core components, and their hypothesised interactions with the context and environment in which they take place, should be mapped onto a logic model (see Figure 5.3 for an example). Delivery and evaluation teams should work closely together to identify the key principles of an intervention. Although the evaluation team can suggest areas of focus and data collection methods, they should aim to work closely with the delivery team to ensure the measures capture, and do justice to, the spirit of the intervention. This is in line with existing EEF guidance on IPE.

Particularly important questions to be considered during this exercise are:

- What are potential unintended (negative) effects from the intervention?
- What alternative hypotheses could explain the intervention effect?

Figure 5.3: An example logic model of a complex whole-school intervention



Reproduced from the Global Research Family Project, Harvard Graduate School of Education (Coffman, 1999).

We also recommend that the delivery and evaluation teams should source potential users and implementers—such as school staff, the governing board, or parents—to validate the logic model. This could be achieved through exploratory, open-ended focus groups or semi-structured interviews. This should commence early in the design stage.

Generally, however, we would argue that the logic model is not a static product: it should be refined through continuous reflection and evaluation in a multi-stage protocol (see Section 2).

5.3 Core components

> For flexible, less prescriptive interventions, and/or those that are complex and multi-component in nature, the process of measuring implementation is more complicated. For example, fidelity/adherence takes on a different meaning when there is no prescribed model against which to assess the delivery. In this case, fidelity may be assessed by rating, for example, the extent to which the key *principles* of the intervention are being adhered to as opposed to specific *practices*. Agreeing such principles with delivery partners from the outset is an important consideration, and as such is an ideal topic for exploration in an IDEA workshop.
>
> *Humphrey* et al*. (2016a, p. 15)*

5.3.1 Identifying the core components at design stage

Before mapping the pathways of how an intervention may lead to hypothesised outcomes, the delivery and evaluation teams should have a good understanding of the intervention core components. Core components are the activities that take place as part of the intervention, the active ingredients are the ones which appear to trigger the mechanisms of change and, thus, drive the change in outcomes.

If it becomes clear that these have not been fully developed or clearly theorised, we suggest piloting the intervention on a small scale before rolling it out more widely. This pilot phase will allow the teams to identify the intervention core components.

There are several reasons why identifying the core components is an essential step in the evaluation of CWSIs.

1. If the delivery team can pinpoint the elements they believe will positively impact both the proximal outcomes (for example, attendance) and the longer-term outcomes (for example, attainment at the end of the 2-year programme), evaluators will be able to make decisions about appropriate outcome measures.
2. Clear and consistent measurement of the intervention core components is central to being able to explain the presence or absence of effects (Bertram, Blase and Fixsen, 2015).
3. Identifying core components early on in the design process will help distinguish these from elements that can be modified to fit within the context (Blase and Fixsen, 2013).

5.3.2 Measuring core components during implementation

Once the intervention core components have been identified, evaluators should think critically how to operationalise and measure these, and IPE should measure how participants respond to, and interact with them. Additionally, it is important to measure how these core components were delivered during implementation to ascertain if certain elements were more ‘active’ at engendering change than others. The IDEA workshop conducted in collaboration between the evaluation and delivery teams—as recommended by Humphrey *et al*. (2016b, p. 31)—can be used to explore how identified core components can be operationalised. During implementation, IPE should assess their relative presence or absence across intervention sites (Humphrey, 2016b, p. 43). After implementation, evaluators should have a clear idea which hypothesised core components were indeed essential, and which were not. Those components that ‘make the difference’ are the intervention’s ‘active ingredients’.

Programme ‘drift’ can occur if the delivery team is unable to focus on the core components across diverse settings. Adaptation of interventions is important, but one needs to know what elements can be changed to fit better with the new context, and what elements are central to the intervention effect.

5.4 Moderators

Key Message 26: Collection of a set of potential moderators for CWSIs
Create a preferred set of potential moderators that the EEF requires evaluators to collect—potentially via a standardised method of data collection—from all schools participating in evaluations of CWSIs.

Key Message 27: Scales of moderators
Isolate moderators that influence implementation to a greater or lesser extent by creating ‘scales’, hierarchies or definitions. An example of a moderator ‘dashboard’ that could be developed further as part of individual evaluations, or for a standardised EEF approach, is provided in Appendix 2.

5.4.1 Identifying the moderators at design stage

Understanding how the intervention is delivered and experienced in the presence of different moderators (that is, in different contexts)[3] can help to shed light on the mediators and allow us to isolate particular elements that are more or less effective, especially in multiple contexts. It is also important because a popular critique of RCTs and QEDs is that it is often unclear how effects vary in difference contexts (Jamal *et al*., 2015). Capturing moderator variability is essential for CWSIs. Alongside the IPE (as already described by Humphrey *et al*., 2016a), contextual data should be captured.

Moderators are of primary importance in CWSIs as they could influence the implementation of the intervention. Understanding whether two variables have the same relationship across groups (such as gender) can help isolate levers for change. Understanding how these differences manifest can ensure findings are both specific and generalisable beyond their immediate surroundings. Being able to generalise enhances the utility of the evaluation for policymakers and those looking to scale and replicate the intervention.

The challenge faced by CWSIs is how they are delivered and implemented across a variety of contexts: even within a single academy chain, for instance, leadership styles and staffing models can vary considerably. Capturing the types of participants, how they participate, the context within which they operate and, crucially, how these interact, involves a multifaceted approach.

5.4.2 Measuring moderators

Standardising how moderators are measured is valuable to ensure we are robustly capturing information important to an intervention. This could be collated to build a detailed directory of school characteristics, used longer term to assess and segment in order to understand what types of interventions are most effective in certain circumstances.

Critical documentation, whereby delivery teams and evaluators theorise what they perceive to be the key levers of implementation within the schools, can help explore

[3] ‘Context’ refers to the pre-existing set of social situations, norms, values and inter-relationships (such as organisation structure, geographic location, demographics of participants) within which an intervention is implemented (Jamal *et al*., 2015).

whether certain elements in the context may make a greater impact upon the intervention. Outlining answers to the following questions may also be helpful:

1. How will the intervention manifest for students, classroom teachers, senior leaders, or the school body as a whole?
2. What moderators underpin their experience?

Discerning further characteristics within these factors is also helpful, for example, the motivation level of a headteacher, percentage of students with English as an Additional Language (EAL), school ethos or policy, teacher retention, or Ofsted category.

For multiple criteria, data to monitor 'moderators' in the evaluation can be standardised by creating sliding scales to measure where a school falls. For CWSIs, multiple scales could be devised to help understand and isolate the combination of factors that promote or hinder implementation. This would involve significant investment in terms of data collection at the school recruitment stage, but it is recommended in order to provide robust measures around context. This could include school size, Ofsted rating, geographic location, characteristics of pupils, CPD programmes, tenure of staff, and length of service in teaching. Evaluators should, however, consider how burdensome this may be for schools and devise creative solutions to reduce the 'ask' on schools (for example, by making the most of non-school-led sources).

5.5 Mediators

Key Message 28: Monitor mediators throughout the evaluation
It is important to monitor mediators identified as important by the logic model throughout the evaluation. These should include elements such as teacher attitudes that go beyond tick-box measures of fidelity. Further suggestions on measurement are provided in Key Messages 30 and 31.

5.5.1 Identifying the theorised mediators at design stage

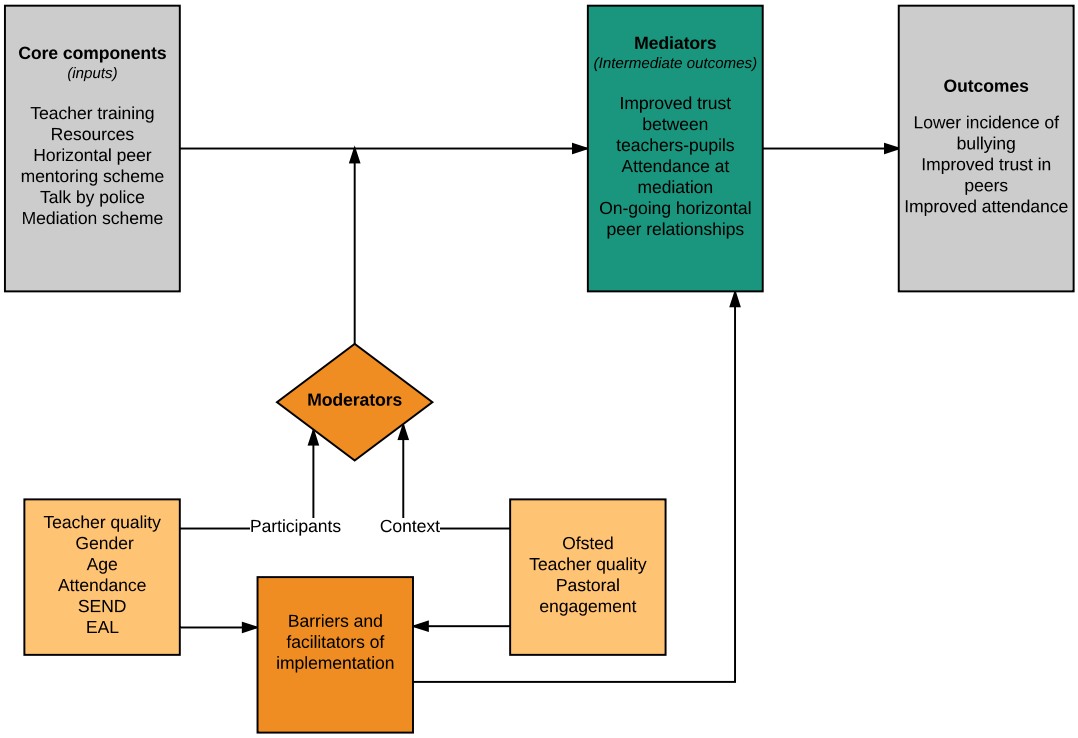
Once core components have been identified and then operationalised, evaluators should identify the mediators of an intervention. These mediators are intermediary outcomes that enable the evaluator to understand how the intervention is progressing, or not, toward achieving its ultimate aims. In other words, mediators are affected by the intervention and, in turn, affect the final outcomes of interest.

Although mediators will vary depending upon the intervention, some are likely to be common across CWSIs—for example, those linked to quality of school leadership. Section 6 returns to this point by discussing adaptation and re-use of existing measures that aim to capture such mediators.

5.5.2 Measuring mediators during implementation

The methods for capturing and measuring data on these mediators should also be discussed at planning stage. It is helpful to think about the various ways mediators can be operationalised. For example, when a proposed short term outcome is 'Increase parent knowledge on how to become involved in their children's education', indicators could include (1) the frequency of attending parents' evenings at schools, (2) the number of phone calls with school staff, or (3) a parent attitude survey administered after the intervention.

Figure 5.4: Example logic model demonstrating examples of mediators



Lack of clarity about the mediators of an intervention can lead evaluators to choose implementation measures that are too simple or reductionist. This may be particularly challenging for CWSIs since there are so many moving parts to consider and weigh up against one another. For example, an intervention designed to increase classroom cohesion and intra-class support networks, if implemented by a disengaged or sceptical teacher, could still tick all the fidelity assessment boxes and achieve a high level of fidelity. If the evaluation team failed to think through the logic model carefully, they may disregard the importance of examining teacher attitudes or buy-in.

Again, we recommend consulting Section 6 for examples of existing measures that aim to capture relevant mediators that may be considered for adaptation and re-use.

5.6 Considering the dimensions of implementation

Key Message 29: Dimensions of implementation
In the context of CWSIs, the IPE should pay particular attention to the importance of Reach, which should consider all persons potentially influenced by, or who could be influencing, the intervention.

Significant guidance is already available for measuring implementation in the EEF's existing IPE guidance (Humphrey *et al*., 2016a). When considering the barriers and facilitators of implementation, there are typically eight dimensions to consider when assessing the intervention:

1. Fidelity/adherence—the extent to which implementers (e.g. teachers) adhere to the intended treatment model.
2. Dosage—how much of the intended intervention has been delivered and/or received.
3. Quality—how well different components of an intervention are delivered.
4. Reach—the rate and scope of participation.
5. Responsiveness—the degree to which participants engage with the intervention.
6. Programme differentiation—the extent to which intervention activities can be distinguished from other, existing practice.
7. Monitoring of control/comparison groups (in a trial context)—determination of the 'counterfactual' (e.g. that which is taking place in the absence of the intervention).
8. Adaptation—the nature and extent of changes made to the intervention, and how these elements interact.
Source: Humphrey et al*., 2016.*

These dimensions, in the context of a CWSI, need to be considered in the broadest possible context.

'Reach' is a particularly important dimension to highlight as it assesses how CWSIs trickle down across various departments and classrooms. Reach often focuses on participant uptake, but, for CWSIs, should consider all persons potentially influenced by, or who could be influencing, the intervention. Monitoring the extent to which the intervention is transmitted throughout a whole-school will involve triangulating survey and interview/observational methods. It is recommended that data, such as training attendance records or web analytics, should be used to understand where an intervention is well, or poorly, diffused. Participants' engagement and context can subsequently be explored.

6 Measures to capture likely mediators in complex whole-school interventions (whole-school improvement and school leadership)

Key Message 30: Suggested measures of mediators
Evaluators should consider the adaptation and re-use of the measures of mediators summarised in Appendix 8 in IPE for CWSIs.

This section is informed by a non-systematic review of literature on approaches and measures to assess whole-school or leadership quality and enable better understanding of these factors as mediators (or interim outcomes) that might be consistently used across schools in an evaluation, and across different evaluations, of CWSIs. It specifically supports the suggestions regarding measurement of mediators in Section 5.

We find that there are currently very few measures and instruments which have been used in empirical research on the impact of school improvement across a number of schools. Standardised instruments, presented in Appendix 8, were nearly all developed for use in the U.S.A. for monitoring school improvement in individual schools or to inform the performance management of individual school leaders. Most studies of school improvement across multiple schools use bespoke instruments to measure mediators.

6.1 Methods of data collection

Key Message 31: Complement schools' existing approach to self-assessment
Where possible, measures and data-collection regarding mediating mechanisms should attempt to complement the types of self-assessment most schools conduct as a matter of course. Consultation with schools during recruitment may help to support this.

Data collection of instruments on school quality and leadership should be carefully planned to avoid both unnecessary burden on schools and contamination of control schools. Self-evaluation has the potential to be an intervention in itself. However, almost all schools are already engaged in self-assessment to at least a limited extent. We suggest tailoring measures and methods of collection to use this to the advantage of the evaluation by 'going with the grain' of schools' existing processes, insofar as this is possible.

Questionnaires are widely used in school self-assessment and are relatively light touch compared to some forms of measurement. This makes them more likely to be acceptable to schools as part of an evaluation, and less likely to cause contamination of the control group. Nevertheless, such an approach would allow for both a richer picture of how, where, and why interventions do, or do not, have an effect, and interim assessments of the impact of CWSIs where these might reasonably be expected to take several years to affect pupil outcomes.

Evaluators should consider capturing termly or annual data on factors that can significantly affect delivery, such as staff turnover, SLT turnover, headline pupil outcomes (GCSEs), Ofsted ratings, exclusions, and expulsions. The ease of access to such reports is an important factor: evaluators should consider what is already being reported to the governing body, for example, and aim to minimise burden on schools.

7 Lowering the assessment costs and burden

This section explores the issue of measuring pupil attainment in evaluations of CWSIs and offers possible solutions, particularly based around considering sources of assessment data as part of the process for selecting schools to be eligible for an evaluation. We consider—

- evaluations making use of commercial test-provider data;
- the use of assessments tied to pedagogical programmes; and
- the use of assessments tied to multi-academy trusts.

We also consider the advantages and challenges of using random sample testing of pupils across a whole school to capture outcomes. Ultimately, the selection of outcome measures should always be driven by the research question of a given evaluation. The aim of this section is to suggest possible solutions that we believe may have particular potential in the context of EEF evaluations of CWSIs.

These messages are unlikely to be relevant to evaluations using QEDs of the type described in Section 4 since much of their advantage stems from not needing to interact with comparison group schools. However, this is not to rule out the possibility of an alternative QED approach that would make use of data of this type. One could imagine, for example, using a commercial test provider's database as the sampling frame for a matched comparison group, however, no assessment of the feasibility of this has been made.

7.1 The problem

EEF guidance states that evaluations should have pupil attainment as their primary outcome, reflecting the organisation's dedication to breaking the link between family income and education achievement.

Relying on national standardised tests to measure the impact of treatment raises four issues:

- England does not have annual standardised testing, so it is only straightforward to collect test scores using nationally available administrative data at ages 11 and 16;
- without waiting seven years in primary schools, or eleven years in secondary schools (which has its own problems; see Section 2), the children taking these tests will only be partially treated by a CWSI;
- the knowledge that the programme will be judged on that year group's achievement may distort implementation away from being a truly whole-school endeavour; and
- we can learn nothing about the effects of the programme in different parts of the school.

These serious shortcomings in the suitability of these tests for evaluations explains why EEF evaluations have made extensive use of commercial tests to measure pupil attainment. However, these have their own issues:

- high costs to the EEF from evaluators administering these tests;
- non-response and attrition producing serious analytical difficulties for evaluation teams; and
- concerns that knowledge of the content of these tests has caused some distortion in programme implementation to target relatively narrow outcomes.

We propose designing evaluations to use assessments that we know participating schools will use with pupils across the school without the need for extensive cost and effort.

7.2 Data availability-led evaluations

Key Message 32: Assess tests administered in schools for efficacy trials
At efficacy trial stage, a lower value is generally placed on external validity compared to costs (since there is generally less evidence of promise at this stage). During design, implementation teams should consider whether it is possible to make use of tests that are already administered in schools, carefully documenting all the drawbacks of doing so.

7.2.1 Choose schools based on existing use of commercial tests

Pupil testing takes place every day in schools, but not via universal, standardised tests. However, many tests are simultaneously being used across large numbers of schools. We do not necessarily know the properties of all these tests, and they may not have been calibrated to national scales, but this does not mean that they are unusable as outcomes in EEF evaluations.

The commercial tests that are used on an ad-hoc basis for evaluations funded by the EEF are already used by thousands of schools across England. Costs (and resources) are already covered by school budgets as they are central to those schools' pupil tracking and feedback systems. If it was possible to elicit a list of schools using a particular test from one of these commercial test providers, then schools on the list could be invited to participate in an EEF-funded evaluation.

The advantages of doing this, if it could be made to work, are clear:

- the schools using these tests are likely reasonably representative of all schools;
- the tests are known to be reliable measures of the major subject disciplines; and
- the commercial test providers almost always mark these tests and so would hold the assessment information in a way that could be straightforwardly and securely passed to a third-party organisation.

However, there are three obvious reasons why commercial test providers would be reluctant to provide such information:

- being sensitive commercial information, they would not want to risk it falling into the hands of competitors (although we know of several data-sharing agreements for research or commercial collaboration);
- they benefit from the status quo where a considerable number of their tests are purchased for EEF-funded evaluations; and
- they would be anxious to maintain good relations with their schools and so may not want to pressure them to participate in an evaluation.

We believe that at least some of these reservations could be overcome, particularly through safeguards on data use, for example:

- test providers could draw a sample of schools to contact for recruitment purposes and only those who agree to participate would then become more widely known (rather than directly sharing lists of customers);
- by committing those involved with the evaluation not to list the schools; or
- by restricting access to the EEF Data Archive that would contain information on the schools who participated for a number of years (subject to negotiation).

7.2.2 Use of assessments tied to pedagogical programmes

The most frequent tests used across schools are those that are tied to pedagogical or homework programmes. These could be used to closely observe accumulations in learning, but it would require some investment of time on the part of the EEF to decide how these tests could best be used. Examples include:

- IXL for maths and English;
- mymaths;
- Renaissance Learning reading and maths programmes;
- Educake; and
- Hegarty Maths.

It might be possible to use these tests for low-cost efficacy trials where external validity does not need to be as high. Although, all the difficulties and potential solutions to commercial sensitivities of data described above also apply here.

For trials of this type, it is important to note that these programmes are designed to alter the path of student learning on their own. As such, trials using these as outcome measures could only formally be interpreted as the impact of the intervention *alongside* whatever package is being delivered to collect assessment material.

Whether this matters will rather depend on the nature of the intervention and of the assessment package. For example, it would not be appropriate to test the impact of a scheme that encourages parents to read with their child at home using the Renaissance Learning test scores since Accelerated Reader aims to influence the motivation of the child itself to read. On the other hand, if there is an intervention that supports the development of logical thinking, such as Chess in Schools, it may be informative to run a trial using a maths homework package to collect attainment information.

7.2.3 Use of assessments created by multi-academy trusts (MATs) or other collaborative groups

Most MATs, and many other organisations that bring schools together, are working towards creating common assessments so that schools can judge their performance in every major subject and in each year group. Some make use of commercially available tests, but others are developing their own, particularly in subjects where commercial tests are unavailable. Where this type of data is found to be reliable and valid for schools in a chain or organisation, this provides a framework to work to develop a trial that is randomised within the chain.

There are a number of potential downsides:

- few groups of schools are currently large enough to make this approach viable, but this approach is likely to be increasingly viable in the future;
- even so, it is likely to be hard to achieve sufficient sample sizes within a single organisation, so pooling across multiple small trials using slightly different assessment outcomes would have to be considered; and
- the schools in any one MAT are not nationally representative, so a trial would face threats to external validity.

More radically, the EEF could support a group of schools in developing some assessments and then use them as part of a subsequent trial. For example, a group of schools could be recruited who all agree to develop and implement a set of tests throughout Key Stage 3 geography to support pupil learning and feedback. After the first year, the schools in the group could then be randomised in a pedagogical trial focused on geography learning. The

RCT could use the new geography test as its outcome. The impact of all schools participating in an assessment programme could eventually be observed in outcomes.

This approach would be risky because a trial would have to be commissioned without knowing the predictive validity of the assessments at the outset. For this reason, it is most appropriate where existing standardised, commercially available tests are not satisfactory and where there is a secondary attainment measure available (such as a GCSE exam) if the developed assessment is not satisfactory.

7.3 Random sample testing

Key Message 33: Consider random sample testing
Evaluators should consider testing a random sample of pupils from across the school to collect outcome measures. Samples of at least 30 pupils per school are unlikely to reduce statistical power much relative to whole school testing. However, this restricts outcome measures to those suitable for use across the whole age range. Furthermore, the benefits in terms of reducing costs and burden of testing may not be as large as the reduced number of pupils tested would suggest.

An alternative approach to collecting pupil assessment data for evaluations of CWSIs is to conduct random sample testing across schools. Given that our aim is to capture outcomes from across a school, this seems an obvious approach if we accept that testing of all pupils in a school is unlikely to be cost-effective. However, there are some potential issues with this approach (some of which mirror existing issues with testing in EEF-funded evaluations) including:

- the complexity of testing across a whole school;
- the need for testing that is suitable for all ages; and
- implications for statistical power.

We discuss each of these in turn.

7.3.1 Testing across a whole school

Organising testing across a whole school is likely to be significantly more complex than testing a single year because of the increased amount of co-ordination involved. Without trialling this, we are unable to provide guidance on how much more effort per school this is likely to require. Although it will, of course, be less than trying to test all pupils in the school.

7.3.2 Testing suitable for all ages

Testing a wide age-range requires the use of adapted or adaptive assessments designed to produce comparable outcome data in these circumstances. As such, it is likely to require the use of established commercial tests, such as GL Assessment's Progress Test series for primary age children. In secondary education there are fewer obvious options: CEM's offering for this period is divided into MidYIS (ages 11–14) and YELLIS (ages 14–16) which cannot be treated as a single assessment.

7.3.3 Implications for statistical power

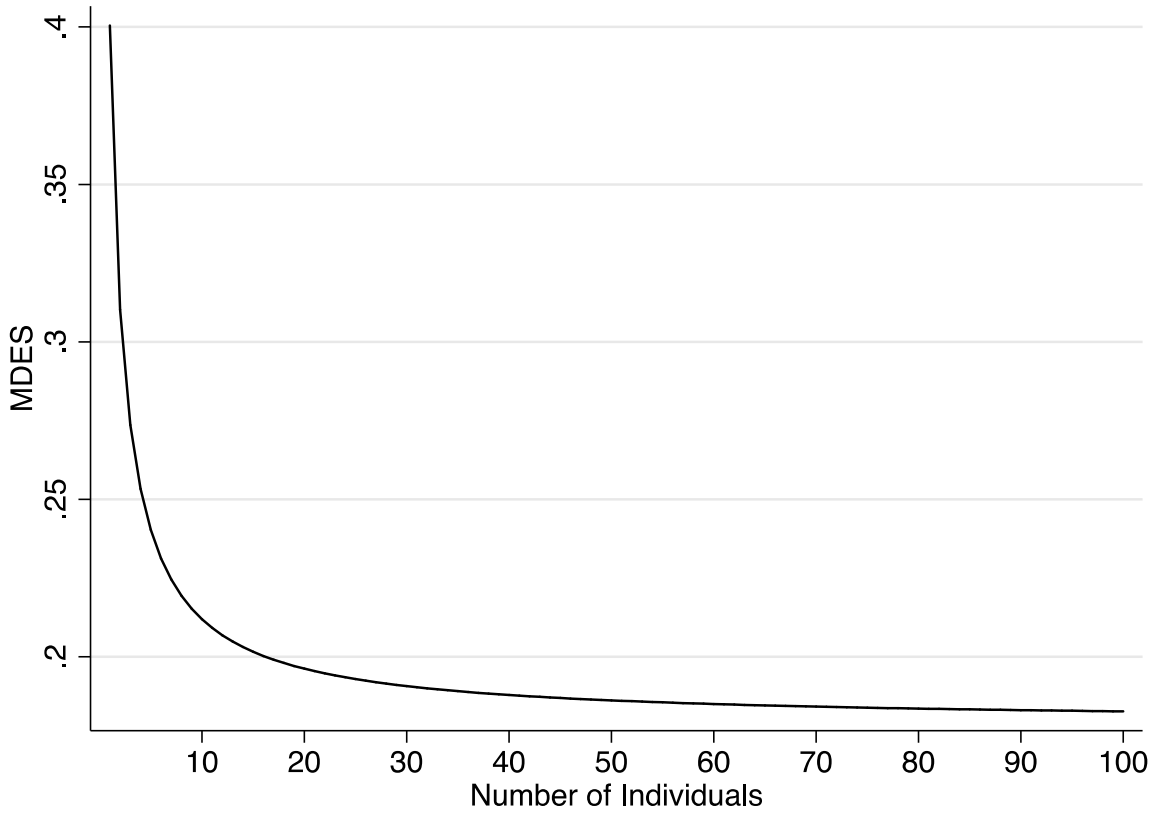
The implications for statistical power of using a random sample are unlikely to be particularly large once a critical mass is reached. Power calculations, based on standard assumptions regarding clustering within schools and for effect sizes of similar magnitude to those detected in previous EEF-funded evaluations, highlight the diminishing returns associated with increasing the number of pupils tested within a school (see Figure 7.1).

While increasing the number of pupils tested from 10 to 20 reduces the minimum detectable effect size from 0.212 to 0.196, increasing it from 20 to 30 only reduces it to 0.191. The precise trade-off would depend upon the particular circumstances of the

evaluation, but this analysis (along with others where other assumptions are varied) suggests that there is little cost-effective gain to sampling more than 30 pupils from across each school.

Of course, this assumes that the pupils are truly randomly sampled (blocking by year group is also advisable). Biased sampling, for example due to non-response by some individuals, would be as problematic (if not more so, as each individual would be proportionally more important) as it is for current EEF testing, especially if it varied systematically across trial arms.

Figure 7.1: Trade off-between pupils tested per school and minimum detectable effect size



Note: Based on a trial involving 50 treated and 50 untreated schools, ICC of 0.2, proportion of variance explained by single pre-test variable of 0.5 at individual and cluster levels, and seven blocks with a two-tailed 0.05-level significance test at 0.8 power.

Appendix 1: Glossary

Active ingredients: Active ingredients are 'intervention-specific components serving as key levers of change' (Abry *et al.*, 2015, p. 320). In contrast to core components—defined as 'the intervention elements that are hypothesised to create the desired change' (*ibid.*, p. 321)—active ingredients are identified during implementation and analysis (whereas core components are proposed at the design stage). Distinguishing active ingredients is important for future iterations of an intervention and for policymakers more broadly.

Context: Context refers to the pre-existing set of social situations, norms, values, and inter-relationships (such as organisation structure, geographic location, demographics of participants, economic factors) within which an intervention is implemented (Jamal *et al.*, 2015). Context, by interacting with mechanisms, determines the direction of outcomes and change (Lacouture *et al.*, 2015).

Core components: As defined above, core components are 'the intervention elements that are hypothesised to create the desired change' (Abry *et al.* 2015, p. 321). The core components can be entered into the intervention logic model under 'activities'. Core components, and how they interact with the context and participants, are also the target of fidelity assessments during Implementation and Process Evaluation (IPE).

Implementation and process evaluation (IPE): This may consist of both qualitative and quantitative measures of the functioning of an intervention, by examining implementation, mechanisms and context.

Logic model: A logic model outlines the implementation of an intervention, and how to measure the implementation and its outcomes over time. The logic model describes the resources applied to ensure implementation, intervention components, mechanisms or intermediary outcomes, and long-term intended outcomes.

Moderator: If a relationship between the intervention (X) and outcome (Y) is different at different values of a variable (Z), we call this 'moderation'. Examples of moderating variables are stable characteristics such as gender, socioeconomic status (SES), or intelligence. Z influences the strength of the relationship between X and Y (MacKinnon, 2011) and identifies stable contextual factors that interact with the intervention.

Mediator: If a variable (Z) is intermediate in a causal sequence between intervention (X) and outcome (Y), that is, X → Z → Y, we call this 'mediation'. It intervenes between the intervention and the outcome, and focuses on participants' responses to the programme or intervention. For example, in a drug prevention program, the intervention (X) could be training and improved access to information, which targets expectations about drug use (mediating, Z), to ultimately change drug use (Y).

Quasi-experimental design (QED): A design that employs statistical techniques to assemble a comparison group against which a treated group may be reasonably compared to estimate the effect of the intervention.

Randomised controlled trial (RCT): In its simplest form, individual participants are assigned either to a control condition (where participants receive 'business as usual' or a placebo intervention), or to one of a number of treatment groups, at random. The groups are followed up to see how effective the experimental intervention was.

Mixed-methods RCT: This refers to an RCT embedded with a broader evaluation design which involves quantitative and qualitative methods. It aims to combine the strengths of the counterfactual analysis of an RCT with the use of qualitative and quantitative data to understand not only ‘what works’ but also ‘how’ and ‘why’. It incorporates implementation and process evaluation and triangulates quantitative and qualitative data to understand better how contextual factors interact with the intervention (Spillane *et al.*, 2010).

Appendix 2: Example outline protocol for evaluation of a complex whole-school intervention

Intervention and Significance

Challenge the Gap is a school-to-school approach co-ordinated by Challenge Partners that aims to narrow the attainment gap between pupils from advantaged and disadvantaged backgrounds. Challenge Partners is a network of schools committed to taking forward and developing successful school improvement approaches. The Challenge the Gap intervention contends that 'Radical improvements can be achieved by identifying outstanding individuals within struggling schools and providing them with the coaching, support and networks with outstanding colleagues in other schools. With the right training and tools, these schools can work together to learn from each other and boost their performance'. The EEF originally carried out an evaluation of this intervention from 2012 onwards.

Methods

A challenge for evaluating this model is that it explicitly involves groups of schools working together—and fully committed to working together. Since these groups were already in place—formed as part of the piloting process—schools could not be randomised into groups. Even if such randomisation had been possible in this study, there is a question mark about whether this would be compatible with the level of mutual commitment necessary for the model to work. In addition, there are difficulties in randomising groups into treatment or control since, having established the groups, it may not be credible to ask the groups allocated to the control condition now not to work together (as would need to be the case for this to be a business as usual comparison).

As such, the previous evaluation took a matching approach to impact evaluation (matching on school characteristics including attainment, numbers of FSM pupils, Ofsted rating, and a measure of motivation to engage such as Ofsted leadership ratings), supported by a qualitative analysis aimed at understanding the aspects of collaboration important to bringing about improvements, the organisational conditions necessary for the approach to be effective, and barriers to these conditions.

While these research questions remain the right ones, considerations highlighted in this review suggest a number of changes would strengthen the evaluation, improving the potential for credible impact estimates and a more thorough understanding of the potential mediating mechanisms. We recognise the inherent difficulties of using randomisation for this approach, although if it is the case the intervention only works 'with the right training and tools' then it may be possible to get an estimate of the impact of only these training and tools through randomising self-nominated groups of schools either to receive these or not. The impact of schools working together without these training and tools would not be picked up by this estimate, since schools in the control group would also be doing this. However, if this impact is negligible this might still be a viable approach.

Participants

The existing groups (formed for the pilot) were used in the study itself. However, had new groups been needed, and given the need for groups of schools committed to working together, it is likely that recruitment would be a complex and time-consuming process.

The previous evaluation's IPE explored what organisational factors must be in place for this approach to work. However, some of the likely factors will be clear in advance of the

evaluation beginning. Recruitment for a trial of this type should include an assessment of the organisational readiness of schools to engage in the intervention and trial.

Based on both of these factors, we suggest the need for plenty of time, perhaps as much as a whole academic year, for the process of recruitment to trials where formation of groups is required.

Analysis Plan

If we continue to pursue a matching approach, then we suggest augmenting this with a matched difference in differences design. Thus, we estimate impact by looking at how trends in attainment outcomes in these schools diverge once the intervention is implemented. This reduces the potential for bias by relying only on there being common trends in outcomes between treatment and comparison, rather than common levels. There is also more scope for testing the plausibility of this assumption by exploring pre-treatment trends in outcome variables. For the purposes of this exercise we obviously cannot explore all aspects of the matching approach, but for the sake of an example we might want to consider a school-level coarsened exact matching approach on the following characteristics:

- proportions of pupils eligible for pupil premium, with SEN, with EAL, or of White British ethnic background;
- average performance of pupil intake at KS2; and
- average performance of pupils at KS2.

This might be combined with exact matching on school type and government office region.

Implementation and process evaluation (IPE)

The complexity of the intervention makes the integration of the impact evaluation within a mixed method design particularly important. While programme developers and evaluators may have strong hypotheses about the channels through which the intervention will improve outcomes, as reflected in the development of the logic model during design phase, it is important that these be tested systematically, with the logic model updated during the course of the intervention as part of a multi-stage protocol model. From the start, we suggest identifying likely mediators that may be possible to measure at regular intervals, in line with the key messages of Sections 4 and 5.

The first-stage of the multi-stage protocol would contain all details currently included in a traditional one-stage protocol, based on the CONSORT statement. As such, it would define features including the intervention, its significance, research questions, impact estimation design (in this case matched difference in differences), the participants (in this case only those included in the pilot were eligible for the trial), primary and secondary outcome measures, sample size calculations, an outline analysis plan, cost data collection, IPE methods, and timetable. However, it would include in its timetable the publication of a second protocol to be published after completion of IPE but before impact evaluation. This would draw on the IPE to define additional hypotheses that it will be possible to test with data collected as part of the trial. For example, it might emerge from IPE that mutual trust between heads of schools within groups is an important mediator. If it were possible to capture data on this factor we could then list this as a hypothesis and conduct confirmatory analysis of the relevance of this factor to impact.

Timetable

The timetable would include usual elements, such as time of randomisation, but would formally define:

- date for completion of IPE (before impact evaluation analysis); and
- date for publication of second stage protocol (before impact evaluation analysis).

Appendix 3: Example communications pro-forma

Outlining a schedule for communication will enable delivery and evaluation partners to communicate effectively so both our aware what the other needs, and by when.

Item	Communicated by	Communication date	Action for school	Deadline
Consent forms	ER	11/10/17	Post forms to ER	November 17
Training dates	RG	18/10/17	Sign up to training date	31/10/17
ASBI Forms				
RA Visit schedule				

Appendix 4: Example checklist for assessing organisational readiness of schools (during recruitment)

Charting the broader policy landscape		
1	Does the proposed intervention go with the grain of current school policy?	
2	Are important school policy changes happening that could influence implementation and evaluation both in terms of proximal and long-term outcomes?	
Leadership		
3	Is the problem the intervention targets of interest to the school at the time of recruitment?	
4	Are the impact/outcomes that the school leadership are hoping to achieve through the intervention in line with its stated aims?	
5	Is there evidence that the leadership team will support and encourage staff to engage with the project?	
Communication		
6	Is there evidence of effective communication of change from the leadership team to the wider organisation?	
7	Is there evidence of effective mechanisms for teaching staff to share concerns regarding the project with the leadership team?	
8	Is there a plan in place for the briefing/training of new teachers who join the school during the intervention by someone strongly invested in the intervention?	
Processes		
9	Has the school put in place plans for how administrative staff, teachers, and the leadership team would collaborate during intervention implementation?	
10	Is the school able to appoint an individual within the school who can maintain support for the intervention for the duration of the evaluation?	
11	Do staff have clear roles and responsibilities for implementation of the intervention?	
Technical assistance (if applicable to intervention)		
12	Does the school have an IT team/person?	
13	Is there clear technical assistance resource easily available for teachers to drawn upon?	

Note: Schools should meet most, if not all, of the criteria to be considered ready to participate in the evaluation.

Appendix 5: Checklist for assessing plausibility of quasi-experimental methods

Matching		
1	Data on a large number of potentially comparable untreated schools is or will be available to use for the purposes of modelling the probability of participation in the intervention.	
2	Data (as per question 1) includes factors relevant to predicting a school's participation in the intervention.	
3	Participation in the programme is likely to be affected by arbitrary factors (e.g. informational constraints; personal preference) that do not predict school performance in our outcomes of interest.	
Difference in differences		
1	Participation in the programme is likely to be affected by arbitrary factors (e.g. informational constraints; personal preference) such that there exists a group of schools with comparable trends in outcomes that do not participate in the project.	
2	Outcomes data from this group of comparable untreated schools as well as the treated schools is, or will be, available for several years before the introduction of the intervention, as well as for the following years.	
3	The timing of the introduction of the intervention is not likely to be coincident with other factors changing within schools, e.g. change of leadership.	
Matched difference in differences		
1	Data on a large number of potentially comparable untreated schools is or will be available to use for the purposes of modelling the probability of participation in the intervention.	
2	Data in question 1 includes factors relevant to predicting a school's participation in the intervention.	
3	When new teachers join throughout the year, will they be briefed/trained? By whom?	
4	Participation in the programme is likely to be affected by arbitrary factors (e.g. informational constraints; personal preference) such that there exists a group of schools with comparable trends in outcomes that do not participate in the project.	
5	The timing of the introduction of the intervention is not likely to be coincident with other factors changing within schools e.g. change of leadership.	

Note: In general, all criteria should be met to support use of the relevant method.

Appendix 6: Template for statistical analysis plan (SAP) for matched difference in differences evaluations

This template is designed for use in evaluations involving matched difference in differences impact evaluation of complex whole-school interventions using data from the National Pupil Database. See Section 4, 'Evaluation of Complex Whole-School Interventions: Methodological and Practical Considerations' for further discussion of this method. We do not advocate its use in all settings but suggest that it will often be a useful approach in this setting.

For some evaluation teams, the trial manager and statistician may be the same person and for others not. In all cases, the SAP should be written for a statistician or analyst to be able to carry out the analysis without prior knowledge of the trial. This is important to avoid bias. Describing the analyses in sufficient detail for someone else to carry them out with certainty avoids conscious or sub-conscious decisions being made based on results seen. The SAP, if written sufficiently early, also provides continuity should key members of the evaluation team leave during the trial. Depending on the level of detail within the trial protocol, some sections of the SAP can be cut and pasted from it. Others will require further detail.

The SAP should be completed once the approach to selecting a matched sample of schools has been agreed and the matching process has been completed. This process should be completed before post-intervention outcomes data become available. If this is not possible for some reason, steps should be taken to ensure it is conducted blind to post-intervention outcomes data. It should also be completed at least three months before planned impact analysis to allow time for review of the analytical approach by one of a panel of EEF SAP reviewers. This timing allows testing of the balance of the matched sample on observables and common pre-intervention trends in outcome variables before final agreement of the SAP and agreement of additional outcomes as part of a second-stage trial protocol.

INTERVENTION	
DEVELOPER	
EVALUATOR	
TRIAL REGISTRATION NUMBER	
TRIAL STATISTICIAN	
TRIAL CHIEF INVESTIGATOR	
SAP AUTHOR	
SAP VERSION	
SAP VERSION DATE	
EEF DATE OF APPROVAL	
DEVELOPER DATE OF APPROVAL	

Table of contents

[INSERT TABLE OF CONTENTS HERE]

Introduction

This should contain a brief description of the intervention and evaluation, including the purpose of the analyses to be performed.

It should note that this is a quasi-experimental evaluation using a matched difference in differences approach.

Study design

This should include:

- description of population including eligibility criteria;
- description of quasi-experimental evaluation design (i.e. matched difference in differences);
- number of treated schools;
- description of intervention; and
- outcome measure(s).

Protocol and SAP changes

If any changes to the trial protocol affect the SAP, these should be highlighted here.

Sampling frame

Full details of the sampling frame should be reported here:

- detailed specification of the pool of potential matches.

Matching approach

Full details of the matching approach should be reported such that the same process could be replicated by a third party. As such, a full description of this stage is essential including:

- specification of preferred matching approach (e.g. 1:1 nearest neighbour matching without replacement) including whether this is purely at school-level or includes both levels (purely individual-level is not possible for difference in differences); and
- specification of matching model(s) (including list of variables—including NPD variable names if possible—to be included, and rules for inclusion of interaction terms, if appropriate).

If possible, the code used to construct the matched sample should be reported as an appendix.

Difference in differences approach

Full details of the difference in differences approach should be reported. This should include:

- precise timing of the pre-intervention period and the post-intervention period.

Outcome measures

Primary outcome

A full description of the outcome variable to be used should be given here. Evaluations using administrative data from the NPD should give precise description as several variables can be used for loosely specified outcomes such as ‘GCSE maths’.

As NPD outcomes are available from each year, the evaluator should specify the precise comparison of time periods which will be used to calculate the primary outcome difference in differences treatment effect estimate.

The expected range of the primary outcome should be described in the analysis plan.

Secondary outcomes

As for the primary outcome, a full description of variables is required. If a multi-stage evaluation protocol approach is being used it may be necessary to update this after later stage protocols.

Analysis

In addition to a written analysis plan, evaluators may wish to publish analysis syntax in advance of running it. This approach guarantees the absence of post-hoc decisions better than a written plan.

The statistical approach should be clarified and justified.

The analysis description should follow EEF guidelines insofar as these are applicable to any impact estimation model.

Primary intention-to-treat (ITT) analysis

The following elements should be included:

- model coefficient that will constitute the estimated intention-to-treat effect on the primary outcome;
- specify the chosen analysis model in full including level(s) of analysis, covariate(s), and their source instruments/datasets;
- confirm the inclusion in the model of all pre-specified covariates regardless of whether they are significant; and
- software used to run the model.

Imbalance

The analysis used to explore the imbalance between the treatment and matched comparison groups, in terms of background characteristics, should be specified. This should include a list of the variables that have been tested for balance between the intervention group and matched comparison group (these should include, at least, all variables included in matching model(s)).

Robustness checks/sensitivity analysis of matching approach

This should include:

- at least five alternative matching specifications (e.g. alternative caliper widths; alternative matching approaches) that will be used to test the robustness of results to deviations from the preferred approach; and
- details of additional analyses that will be carried out using alternative specification matched comparison groups to verify that the findings are not highly dependent upon the approach chosen.

Robustness checks/sensitivity analysis of difference in differences

Description of analysis of pre-treatment common trends in primary outcome including:

- ‘placebo’ pre-treatment impact estimates; and
- graphical analysis of pre-treatment trends in primary outcome.

This may also include additional robustness checks such as:

- alternative specification of pre-intervention and post-intervention years to estimate impact.

Non-compliance with intervention

A suitable analysis to estimate treatment effects for compliers should be included here, except where intervention uptake is expected to be close to 100%. This is likely to require additional matching. It will not be possible to identify the matched sample ahead of analysis, but the approach to be taken should be finalised in advance and generally analogous to the main matching model.

This should include:

- description of variable used to describe extent of intervention ‘dosage’; and
- description of on-treatment analysis model.

Secondary outcome analyses
The level of description should match that of the primary ITT analysis.

Subgroup analyses
These should include details of:

- subgroup analyses specified in the protocol;
- subgroup analyses not specified in the protocol; and
- description of model, including whether an interaction term is used or a separate dataset containing only members of the subgroup.

Effect size calculation
This should include details of:

- formula for calculation of effect size (Hedges' *g*) including exact specification of the numerator;
- how confidence intervals/Bayesian credibility intervals will be calculated.

Additional analyses
Further planned analysis should be described in this section.

Report tables
Applicable tables from the EEF trial report template should be pasted into the SAP and populated with relevant variables.

Templates for any tables and charts additional to those in the report template should also be specified in the SAP.

Appendix 7: Example moderator dashboard

Name: [School X]

Ofsted	Special M	RI	**Good**	Outstanding
Staff turnover	Low (0–10%)	**Average (10–20%)**	High (20–30%)	Very high (30%)
Head in post	Less than 1 yr	1–3 yrs	4–6 yrs	**6+ yrs**
Head engagement	**Low**	Average	High	Very high

Engagement Scale (devised with delivery partner):

- Low = attends less than 50% of training sessions, fails to assign intervention to member of SLT, whole-school training yet to be organised;
- Average = attends at least 50% of training sessions, assigned intervention to member of SLT, whole-school training organised;
- High = attends at least 75% of training sessions, assigned intervention to member of SLT, whole-school training organised; and
- Very High = attends over 90% of training sessions, assigned intervention to member of SLT, whole-school training organised.

Appendix 8: Summary table of standardised instruments for measuring factors linked to whole-school improvement and leadership.

Instrument	Type of measure	How delivered/who completes it	Where used	What dimensions does it measure?	Method and frequency of administration.	Can some factors be omitted or added?	Suitable for baseline?	Suitable for interim outcomes?	Evidence of reliability and validity	Advantages	Disadvantages
Instruments for measuring factors linked to whole-school improvement											
5Essentials	Questionnaires	Students, teachers, parents report on school-level factors	U.S.A.	Effective leadership; collaborative teachers; involved families; supportive environment; ambitious instruction.	Information not available	Yes	Yes	Yes	Yes	Widely tested and used in U.S.A. Based on established theoretical model.	Limited evidence for use in interventions across a number of schools.
GallupQ12	Questionnaire	Employees self-report	Worldwide	Employee engagement	Information not available	No	Yes	Yes	Yes	Widely-used international tool in a range of organisations, provides actionable issues for managers.	No examples found for use in school interventions.
Organizational Climate Index	Questionnaire	Teachers' self-report	U.S.A., with a small number of examples found of use elsewhere	Collegial leadership; professional teacher behaviour; achievement press.	Information not available	Yes	Yes	Yes	Yes	Widely used in U.S.A. Based on an established theoretical model, part of a suite of questionnaire tools.	Limited evidence of use in interventions.
Comprehensive school climate inventory	Questionnaires	Student, staff and parents report on school factors	U.S.A.	Safety; teaching and learning; interpersonal relations;	Information not available	Yes	Yes	Yes	Yes for student survey, staff and parent surveys yet to	Widely used in U.S.A.	No evidence found from outside the U.S.A. Case study evidence

				institutional environment; leadership (staff survey only); professional relationships (staff survey only).					be independently verified		of use only on developer's website
School Climate Instrument—Revised	Questionnaire	Staff report on school factors	U.S.A.	Order; leadership; environment; involvement; instruction; expectations; collaboration.	Information not available	Yes	Yes	Yes	Yes	Small number of examples in studies of interventions and longitudinal studies combined with other tools in mixed methods designs.	Examples only found from the U.S.A.
School Climate assessment inventory (ASSC_SCAI)	Questionnaires	Reports by students, parents, and staff on school-level factors	U.S.A.	Physical environment; faculty relations; student interactions; leadership and decisions; discipline; environment; learning and assessment; attitude and culture; community relations.	Information not available	Yes	Yes	Yes	Yes, by developer team	Has been used in school improvement research in U.S.A. by developer team members	No examples found of research other than by developer team.
Organizational Health Inventory (OHI)	Questionnaire	Teachers report on school factors	U.S.A.	Goal focus; communication adequacy; optimal power equalization; resource utilisation; cohesiveness; morale; innovation;	Information not available	Yes	Yes	Yes	Yes		No research found of use in intervention studies

				autonomy; adaptation; problem-solving adequacy.							
Instruments designed to measure leadership attributes											
Principal Instructional Management Rating Scale (PIMRS)	Questionnaire	Self-report by leader, teacher form, supervisor form to enable triangulated perspective	International	School performance: defines the school mission; managing the instructional programme; developing the school climate programme. Processes of leadership: frames the school's goals; communicates the school's goals; co-ordinates the curriculum; supervises and evaluates instruction; monitors student progress; protects instructional time; provides incentives for teachers; provides incentives for learning; promotes professional development; maintains high visibility.	Information not available	Yes	Yes	Yes	Yes	Has been widely tested and used in numerous school systems, based on well-developed theoretical model.	No examples found of use in intervention studies.
Vanderbilt Assessment of	Questionnaire	Self-report by leader, teacher	U.S.A.	Core-components of	Information not available	Yes	Yes	Yes	Yes	Recommended in a report by	No evidence found for use

Leadership in Education (VAL-ED)		form, supervisor form to enable triangulated perspective		school performance: high standards for student learning; rigorous curriculum; quality instruction; culture of learning and professional behaviour; connections to external communities; performance accountability. Processes of leadership: planning; implementing; supporting; advocating; communicating; monitoring.						Wallace Foundation which compared tools for leader performance management; has been used to monitor changes in leadership behaviour during leadership development programme and in comparison study.	outside U.S.A., only one example of intervention study located.

Bibliography and other relevant references

Abadie, A., Diamond, A. and Hainmueller, J. (2010) 'Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program', *Journal of the American Statistical Association,* 105 (490), pp. 493–505. doi:10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A. and Hainmueller, J. (2015) 'Comparative Politics and the Synthetic Control Method', *American Journal of Political Science,* 59 (2), pp. 495–510. doi:10.1111/ajps.12116

Abadie, A. and Gardeazabal, J. (2003) 'The Economic Costs of Conflict: A Case Study of the Basque Country', *American Economic Review,* 93 (1), pp. 113–132. doi:10.1257/000282803321455188

Abadie, A. and Imbens, G. (2008) 'On the failure of the bootstrap for matching estimators', *Econometrica*, 76 (6), pp. 1,537–1,557. doi: 10.3982/ECTA6474

Abadie, A. and Imbens, G. (2011) 'Bias-corrected matching estimators for average treatment effects', *Journal of Business and Economic Statistics*, 29 (1), pp. 1-11. doi: 10.1198/jbes.2009.07333

Abry, T., Hulleman, C. and Rimm-Kaufman, S. (2015) 'Using Indices of Fidelity to Intervention Core Components to Identify Program Active Ingredients', *American Journal of Evaluation*, 36 (3), pp. 320–338. doi: 10.1177/1098214014557009

Alcott, B. (2017) 'Does Teacher Encouragement Influence Students' Educational Progress? A Propensity-Score Matching Analysis', *Research in Higher Education*. doi:10.1007/s11162-017-9446-2

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B. and Pianta, R. (2013) 'Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary', *School Psychology Review*, 42 (1), p. 76.

Allen, R. and Allnutt, J. (2013) 'Matched panel data estimates of the impact of Teach First on school and departmental performance', QSS Working Paper 13-11, London: Department of Social Science, UCL Institute of Education. Retrieved from: http://repec.ioe.ac.uk/REPEc/pdf/qsswp1311.pdf

Allen, R., Parameshwaran, M. and Thomson, D. (2016) 'Social and ethnic inequalities in choice available and choices made at age 16', London: Social Mobility Commission. Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/574708/SMC_social_and_ethnic_inequalities_in_post_16_report.pdf

Anders, J. (2014) 'Does an aptitude test affect socioeconomic and gender gaps in attendance at an elite university?', QSS Working Paper 14-07, London: Department of Social Science, UCL Institute of Education. Retrieved from: http://repec.ioe.ac.uk/repec/pdf/qsswp1407.pdf

Antoniou, P. and Kyriakides, L. (2013) 'A dynamic integrated approach to teacher professional development: Impact and sustainability of the effects on improving teacher behaviour and student outcomes', *Teaching and Teacher Education*, 29, pp. 1–12.

Avolio, B. J., Bass, B. M. and Jung, D. I. (1999) 'Re-examining the components of transformational and transactional leadership using the Multifactor Leadership Questionnaire', *Journal of occupational and organizational psychology*, 72 (4), pp. 441–462.

Balch, R. (2012) 'The Validation of a Student Survey on Teacher Practice', Nashville, Tennessee: Vanderbilt University. Retrieved from https://mystudentsurvey.com/wp-content/uploads/2012/06/Balch-Student-Surveys-2012.pdf

Banerjee, R. (2010) 'Social and emotional aspects of learning in schools: Contributions to improving attainment, behaviour and attendance', Sussex: National STRATEGIES Tracker School Project.

Bell, C. D., Raczynski, K. A. and Horne, A. M. (2010) 'Bully Busters abbreviated: Evaluation of a group-based bully intervention and prevention program', *Group Dynamics: Theory, Research, and Practice*, 14 (3), p. 257.

Belot, M. and James, J. (2014) 'A new perspective on the issue of selection bias in randomised controlled field experiments', *Economics Letters,* 124 (3), pp. 326–328.

Berry, V., Axford, N., Blower, S., Taylor, R. S., Edwards, R. T., Tobin, K. *et al.* (2016) 'The effectiveness and micro-costing analysis of a universal, school-based social-emotional learning programme in the UK: a cluster-randomised controlled trial', *School Mental Health*, 8 (2), pp. 238–256. doi:10.1007/s12310-015-9160-1

Bertram, R. M., Blase, K. A. and Fixsen, D. L. (2015) 'Improving programs and outcomes implementation frameworks and organization change', *Research on Social Work Practice,* 25 (4), pp. 477–487.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004) 'How much should we trust difference-in-differences estimates?', *Quarterly Journal of Economics,* 119 (1), pp. 249–275. doi:10.1162/003355304772839588

Bevans, K., Bradshaw, C., Miech, R. and Leaf, P. (2007) 'Staff- and school-level predictors of school organizational health: A multilevel analysis', *Journal of School Health*, 77 (6), pp. 294–302.

Blase, K. and Fixsen, D. (2013) 'Core Intervention Components: Identifying and Operationalizing What Makes Programs Work', ASPE Research Brief, U.S. Department of Health and Human Services.

Blundell, R., Dearden, L. and Sianesi, B. (2005) 'Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168 (3), pp. 473–512. doi:10.1111/j.1467-985X.2004.00360.x

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. and Moore, L. (2012) 'Realist randomised controlled trials: a new approach to evaluating complex public health interventions', *Social Science & Medicine,* 75 (12), pp. 2,299–2,306.

Borkan, B., Capa, Y., Figueiredo, C. and Loadman, W. E. (2003) 'Using Rasch Measurement To Evaluate the Organizational Climate Index', Paper presented at the Annual Meeting of the MidWestern Educational Research Association (Columbus, OH), October 15–18, 2003.

Borman, G. D., Hewes, G. M., Overman, L. T. and Brown, S. (2003) 'Comprehensive school reform and achievement: A meta-analysis', *Review of educational research*, 73 (2), pp. 125–230.

Breiman, L. and Cutler, A. (2004) *Random Forests manual*.

Brewer, M., Crossley, T. F. and Joyce, R. (2013) 'Inference with Difference-in-Differences Revisited', IZA Working Paper 7742, Bonn, Germany: Institute for the Study of Labor. Retrieved from: http://ftp.iza.org/dp7742.pdf

Brown, C., Daly, A. and Liou, Y-H. (2016) 'Improving Trust, Improving Schools: Findings from a Social Network Analysis of 43 Primary Schools in England', *Journal of Professional Capital & Community*, 1, 1, pp. 69–91.

Brown, C. and Rogers, S. (2015) 'Knowledge creation as an approach to facilitating evidence informed practice: Examining ways to measure the success of using this method with early years practitioners in Camden (London)', *Journal of Educational Change*, 16 (1), pp. 79–99.

Bryson, A., Dorsett, R. and Purdon, S. (2002) 'The use of propensity score matching in the evaluation of active labour market policies' (DWP Working Paper 4), London: Department for Work and Pensions. Retrieved from:

http://eprints.lse.ac.uk/4993/1/The_use_of_propensity_score_matching_in_the_evaluation_of_active_labour_market_policies.pdf
Buscha, F., Maurel, A., Page, L. and Speckesser, S. (2012) 'The Effect of Employment while in High School on Educational Attainment: A Conditional Difference-in-Differences Approach', *Oxford Bulletin of Economics and Statistics,* 74 (3), pp. 380–396. doi:10.1111/j.1468-0084.2011.00650.x
Burgess, S. and Greaves, E. (2013) 'Test scores, subjective assessment, and stereotyping of ethnic minorities', *Journal of Labor Economics,* 31 (3), pp. 535–576.
Butler, E. D. and Alberg, M. J. (1991) *Tennessee School Climate Inventory: A Resource Manual*, Memphis, TN: Center for Research in Education Policy.
Caliendo, M. and Kopeinig, S. (2008) 'Some Practical Guidance for the Implementation of Propensity Score Matching', *Journal of Economic Surveys,* 22 (1), pp. 31–72. doi:10.1111/j.1467-6419.2007.00527.x
Camburn, E. M., Goldring, E., Sebastian, J., May, H. and Huff, J. (2016) 'An Examination of the Benefits, Limitations, and Challenges of Conducting Randomized Experiments With Principals', *Educational Administration Quarterly*, 52 (2), pp. 187–220. doi:10.1177/0013161X15617808
Campbell, T. (2015) 'Stereotyped at seven: Biases in teachers' judgements of pupils' ability and attainment', *Journal of Social Policy*. 44 (3), pp. 517–547. doi:10.1017/S0047279415000227
Card, D. (1992) 'Using Regional Variation in Wages to Measure the Effects of the Federal Minimum Wage', *Industrial and Labor Relations Review,* 46 (1), pp. 22–37.
Card, D. and Krueger, A. B. (1994) 'Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania', *The American Economic Review,* 84 (4), pp. 772–793.
Chaplin, D., Gill, B., Thompkins, A. and Miller, H. (2014) 'Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools' (No. 49fdd049478d40b5a294589acf76530a), Mathematica Policy Research.
Chapman, C., Muijs, D. and MacAllister, J. (2011) 'A study of the impact of school federation on student outcomes', National College for School Leadership.
Charalambous, E., Kyriakides, L. and Creemers, B. P. (2016) 'Promoting quality and equity in socially disadvantaged schools: A group-randomisation study', *Studies in Educational Evaluation*.
Christoforidou, M., Kyriakides, L., Antoniou, P. and Creemers, B. P. (2014) 'Searching for stages of teacher's skills in assessment, *Studies in Educational Evaluation*, 40, pp. 1–11.
Clifford, M., Menon, R., Gangi, T., Condon, C. and Hornung, K. (2012) 'Measuring School Climate for Gauging Principal Performance: A Review of the Validity and Reliability of Publicly Accessible Measures', American Institute of Research. Retrieved from http://www.air.org/.
Coffman, J. (1999) 'Learning From Logic Models: An Example of a Family/School Partnership Program', Harvard Family Research Project, Cambridge MA: Harvard Graduate school of Education. Retrieved from http://www.hfrp.org/publications-resources/browse-our-publications/learning-from-logic-models-an-example-of-a-family-school-partnership-program
Collins, J. B. and Pratt, D. D. (2010) 'The teaching perspectives inventory at 10 years and 100,000 respondents: Reliability and validity of a teacher self-report inventory', *Adult Education Quarterly*, p.0741713610392763.
Cook, T. D., Shadish, W. R., and Wong, V. C. (2008) 'Three conditions under which experiments and observational studies produce comparable causal estimates: New

findings from within-study comparisons', *Journal of Policy Analysis and Management*, 27 (4), pp. 724–750. doi:10.1002/pam.20375

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. and Petticrew, M. (2008) 'Developing and evaluating complex interventions: the new Medical Research Council guidance', *BMJ*, 337, p. a1,655.

Craig P., Cooper C., Gunnell D. *et al.* (2012) 'Using natural experiments to evaluate population health interventions: new Medical Research Council guidance', *Journal of Epidemiology and Community Health*, 66, pp. 1,182–1,186.

Creemers, B., Kyriakides L., Panayioutou, A., Bos, W., Holtappels, H., Pfeifer, M., Venneman, M. *et al.* (2013) *Establishing a Knowledge Base for Quality in Education: a Dynamic Theory for Education: Handbook on designing evidence-based strategies and actions to promote quality in education*, Münster: Waxmann. ISBN 978-3-8309-2933-8.

Creemers, B. and Kyriakides, L. (2009) 'Situational effects of the school factors included in the dynamic model of educational effectiveness', *South African Journal of Education*, 29 (3), pp. 293–315.

Creemers, B. and Kyriakides, L. (2008) *The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools*, London: Routledge.

Creswell, J. W., and Clark, V. L. P. (2007) *Designing and conducting mixed methods research*, SAGE Publications Inc.

Datnow, A. (2005) 'The sustainability of comprehensive school reform models in changing district and state contexts', *Educational Administration Quarterly*, 41 (1), pp. 121–153.

Day, C., Hopkins, D., Harris, A. and Ahtaridou, E. (2009) 'The impact of school leadership on pupil outcomes. Final Report', Nottingham, U.K.: National College for School Leadership.

Dearden, L., Emmerson, C., Frayne, C. and Meghir, C. (2009) 'Conditional Cash Transfers and School Dropout Rates', *The Journal of Human Resources,* 4 4(4), pp. 827–857.

de Jong, R. and Westerhof, K. J. (2001) 'The quality of student ratings of teacher behaviour', *Learning Environments Research*, 4 (1), pp. 51–85.

Diamond, A. and Sekhon, J. S. (2013) 'Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies', *Review of Economics and Statistics,* 95 (3), pp. 932–945. doi:10.1162/REST_a_00318

Durlak, J. A. and DuPre, E. P. (2008) 'Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation', *American Journal of Community Psychology,* 41 (3-4), pp. 327–350.

Dynarski, S. (2000) 'Hope for whom? Financial aid for the middle class and its impact on college attendance', NBER Working Paper 7756, Cambridge MA: National Bureau of Economic Research. Retrieved from: http://www.nber.org/papers/w7756

Education Endowment Foundation (2013) *Recruitment and Retention pack*, London: EEF. Retrieved from: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/Recruitment_pack_v2.0.docx_FINAL.pdf

Ehren, M. C. M., Gustafsson, J. E., Altrichter, H., Skedsmo, G., Kemethofer, D. and Huber, S. G. (2015) 'Comparing effects and side effects of different school inspection systems across Europe', *Comparative Education*, 51 (3), pp. 375–400.

Evans, R., Murphy, S. and Scourfield, J. (2015) 'Implementation of a school-based social and emotional learning intervention: understanding diffusion processes within complex systems', *Prevention Science,* 16 (5), pp. 754–764.

Eyles, A., Machin, S. and McNally, S. (2016) 'Unexpected School Reform: Academisation of Primary Schools in England', CEP Discussion Paper 1455, London: Centre for Economic Performance, London School of Economics. Retrieved from: http://cep.lse.ac.uk/pubs/download/dp1455.pdf

Fairchild, A. J. and MacKinnon, D. P. (2009) 'A general model for testing mediation and moderation effects', *Prevention Science*, 10 (2), pp. 87–99.

Ferguson, R. (2010) 'Student perceptions of teaching effectiveness', discussion brief from the National Center for Teacher Effectiveness and the Achievement Gap Initiative, Cambridge MA: Harvard University (source for Tripod's 7Cs framework).

Ferguson, R. F. and Danielson, C. (2014) 'How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching', *Designing Teacher Evaluation Systems*, pp. 98–143.

Finnigan, K. S. and Daly, A. J. (2012) 'Mind the gap: Organizational learning and improvement in an underperforming urban system', *American Journal of Education*, 119 (1), pp. 41–71.

Fredricks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B. and Mooney, K. (2011) 'Measuring Student Engagement in Upper Elementary through High School: A Description of 21 Instruments', *Issues & Answers*, REL 2011-No. 098, Regional Educational Laboratory Southeast.

Gaertner, H. (2014) 'Effects of student feedback as a method of self-evaluating the quality of teaching', *Studies in Educational Evaluation*, 42, pp. 91–99.

Garg, P. and Rastogi, R. (2006) 'Climate profile and OCBs of teachers in public and private schools of India', *International Journal of Educational Management*, 20 (7), pp. 529–541.

Gareis, C. R. and Tschannen-Moran, M. (2005) 'Cultivating principals' sense of efficacy: Supports that matter', in *annual meeting of the University Council for Educational Administration,* Nashville, TN.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. (2010) 'Variable selection using random forests', *Pattern Recognition Letters,* 31 (14), pp. 2,225–2,236. doi:10.1016/j.patrec.2010.03.014

Goe, L., Bell, C., and Little, O. (2008) 'Approaches to evaluating teacher effectiveness: A research synthesis', Washington, DC: National Comprehensive Center for Teacher Quality.

Goff, P., Edward Guthrie, J., Goldring, E. and Bickman, L. (2014) 'Changing principals' leadership through feedback and coaching', *Journal of Educational Administration*, 52 (5), pp. 682–704.

Golan, S., Cassidy, L. and Woodworth, K. (2016) 'The McKnight Foundation Pathway Schools Initiative Phase I Extended Report'.

Goldring, E., Cravens, X., Porter, A., Murphy, J. and Elliott, S. (2015) 'The convergent and divergent validity of the Vanderbilt Assessment of Leadership in Education (VAL-ED) Instructional leadership and emotional intelligence', *Journal of Educational Administration*, 53 (2), pp.177–196.

Grissom, J. A., Kalogrides, D. and Loeb, S. (2014) 'Using student test scores to measure principal performance', *Educational Evaluation and Policy Analysis*, p.0162373714523831.

Grossman, P. (2011) 'Protocol for Language Arts Teaching Observations. Available at http://www. gse. harvard. edu/ncte/news/NCTE_Conference_PLATO_Grossman. pdf

Guo, P. Choe, J. and Higgins-D'Alessandro, A. (2011) 'Report of Construct Validity and Internal Consistency Findings for the Comprehensive School Climate Inventory', Fordham University (unpublished paper).

Hall, G. E., Dirksen, D. J., and George, A. A. (2006) 'Measuring implementation in schools: Levels of use', Southwest Educational Development Laboratory.

Hallinger, P. (1984) *Principal instructional management rating scale*, New York, NY: Leading Development Associates.
Hallinger, P. (2012) 'A data-driven approach to assess and develop instructional leadership with the PIMRS', *Tools for improving principals' work*, pp. 47–69.
Hallinger, P. (2015) 'Developing the PIMRS Instrument', in *Assessing Instructional Leadership with the Principal Instructional Management Rating Scale* (pp. 47–61), Springer International Publishing.
Hallinger, P. and Heck, R. H. (2011) 'Leadership for learning: Does collaborative leadership make a difference?', *Educational Management Administration & Leadership*, 38 (6), pp. 654–678.
Hallinger, P. and Lee, M. (2013) 'Mapping instructional leadership in Thailand: Has education reform impacted principal practice?', *Educational Management Administration & Leadership*, pp. 174–196.
Hanover Research (2013) *Student Perception Surveys and Teacher Assessments*. Retrieved from https://dese.mo.gov/sites/default/files/Hanover-Research-Student-Surveys.pdf
Hapfelmeier, A. and Ulm, K. (2013) 'A new variable selection approach using Random Forests', *Computational Statistics & Data Analysis*, 60, pp. 50–69. doi:10.1016/j.csda.2012.09.020
Harter, J. K., Schmidt, F. L., Killham, E. A. and Asplund, J. W. (2013) *The relationship between engagement at work and organizational outcomes*, Washington: Gallup Poll Consulting University Press.
Hill, H. C., Umland, K., Litke, E. and Kapitula, L. R. (2012) 'Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice', *American Journal of Education*, 118 (4), pp. 489–519.
Ho, D., Imai K., King, G. and Stuart, E. (2007) 'Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference', *Political Analysis*, 15, pp. 199–236.
Holmes, J. O. (2015) 'An Investigation of the Impact of an Aspiring Principals Preparation Program on Principal Leadership Effectiveness', Doctoral dissertation, University of Maryland.
Mullis, I. V. S., Cotter, K. E., Fishbein, B. G. and Centurino, V., 'TIMSS 2015 Context Questionnaire Framework', from Chapter 3 of Hooper, M., Mullis, I. and Martin, M. (eds) (2016) *Methods and Procedures in TIMSS 2015*, Boston College, TIMSS & PIRLS International Study Center.
Hoy, W. K., Smith, P. A. and Sweetland, S. R. (2002) 'The development of the organizational climate index for high schools: Its measure and relationship to faculty trust', *The High School Journal*, 86 (2), pp. 38–49.
Hoy, W. K. and Sabo, D. J. (1998) *Quality Middle Schools: Open and Healthy*, Thousand Oaks: Corwin Press.
Hoy, W. K., Tarter, C. J. and Kottkamp, R. B. (1991) *Open schools, healthy schools: Measuring organizational climate*, Thousand Oaks: Corwin Press.
Hoy, W. K. and Clover, S. I. (1986) 'Elementary school climate: A revision of the OCDQ', *Educational Administration Quarterly*, 22 (1), pp. 93–110.
Huijgen, T., van de Grift, W., van Boxtel, C. and Holthuis, P. (2017) 'Teaching historical contextualization: the construction of a reliable observation instrument', *European Journal of Psychology of Education*, 32, p.159. doi:10.1007/s10212-016-0295-8
Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. and Kerr, K. (2016a) *Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook*, University of Manchester. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Guidance_Final.pdf

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. and Kerr, K. (2016b) *Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature*, University of Manchester. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Review_Final.pdf

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

Jamal, F., Fletcher, A., Shackleton, N., Elbourne, D., Viner, R. and Bonell, C. (2015) 'The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-example', *Trials,* 16 (1), p. 1.

Jones, A. and Shindler, J. (2016) 'Exploring the School Climate-Student Achievement Connection: Making Sense of Why the First Precedes the Second', *Educational Leadership and Administration*, 27, p. 35.

Kam, C. M., Greenberg, M. T. and Walls, C. T. (2003) 'Examining the role of implementation quality in school-based prevention using the PATHS curriculum', *Prevention Science,* 4 (1), pp. 55–63.

Kane, T. J. and Staiger, D. O. (2012) 'Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains', MET Project research paper, Seattle, WA: Bill & Melinda Gates Foundation.

Kitchen, S., Tanner, E., Brown, V., Payne, C., Crawford, C., Dearden, L., Greaves, E. and Purdon, S. (2013) 'Evaluation of the Free School Meals Pilot: Impact Report', London: Department for Education, Research Report DFE-RR227. Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/184047/DFE-RR227.pdf

Kitchen, S., Tanner, E., Turczuk, O., Finch, S., Crawford, C., Dearden, L. and Purdon, S. (2009) 'Evaluaton of the Free School Meal Pilots: Baseline Report', London: Department for Education. Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/184047/DFE-RR227.pdf

Klugman, J., Gordon, M. F., Sebring, P. B. and Sporte, S. E. (2015) 'A first look at the 5Essentials in Illinois schools' (research summary), University of Chicago Consortium on Chicago School Research.

Korkmaz, M. (2007) 'The effects of leadership styles on organizational health', *Educational Research Quarterly*, 30 (3), p. 22.

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S. and Sutton, M. (2016) 'Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units', *Health Economics,* 25 (12), pp. 1,514–1,528. doi:10.1002/hec.3258

Krueger, A. B. (2003) 'Economic considerations and class size', *The Economic Journal*, 113 (485), F34–F63.

Kyriakides, L. (2005) 'Extending the Comprehensive Model of Educational Effectiveness by an Empirical Investigation', *School Effectiveness and School Improvement* 16 (2), pp. 103–152.

Kyriakides, L., Creemers, B., Panayiotou, A., Vanlaar, G., Pfeifer, M., Gašper, C. and McMahon L. (2014) 'Using Student Ratings to Measure Quality of Teaching in Six European Countries', *European Journal of Teacher Education*, 37 (2): pp. 125–143.

Kyriakides, L., Creemers, B. P., Papastylianou, D. and Papadatou-Pastou, M. (2014) 'Improving the school learning environment to reduce bullying: An experimental study', *Scandinavian Journal of Educational Research*, 58 (4), pp. 453–478.

Kyriakides, L., Creemers, B., Antoniou, P. and Demetriou, D. K. (2010) 'A synthesis of studies searching for school factors: Implications for theory and research', *British Educational Research Journal*, 36 (5), pp. 807–830.

Kyriakides, L. and Creemers, B. (2008) 'Using a Multidimensional Approach to Measure the Impact of Classroom-level Factors Upon Student Achievement: A Study Testing the Validity of the Dynamic Model', *School Effectiveness and School Improvement*, 19 (2): pp. 183–205.

Lacouture, A., Breton, E., Guichard, A. and Ridde, V. (2015) 'The concept of mechanism from a realist approach: a scoping review to facilitate its operationalization in public health program evaluation', *Implementation Science*, 10 (1), p. 153.

Leithwood, K., Day, C., Sammons, P., Harris, A. and Hopkins, D. (2006) 'Seven strong claims about successful school leadership', research report, Nottingham UK: National College for School Leadership. Retrieved from: http://dera.ioe.ac.uk/6967/1/download%3Fid=17387&filename=seven-claims-about-successful-school-leadership.pdf

Leithwood, K. and Jantzi, D. (2008) 'Linking leadership to student learning: The contributions of leader efficacy', *Educational Administration Quarterly*, 44 (4), pp. 496–528.

List, J., Shaikh, A. and Xu, Y. (2016) 'Multiple Hypothesis Testing in Experimental Economics', NBER Working Paper 21875, Cambridge MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w21875

Machin, S., Telhaj, S. and Wilson, J. (2006) 'The Mobility of English School Children', *Fiscal* Studies, 23 (3), pp. 253–280. doi:10.1111/j.1467-8578.2006.00035.x

Machin, S. and McNally, S. (2008) 'The literacy hour', *Journal of Public Economics,* 92, pp. 1,441–1,462. doi:10.1016/j.jpubeco.2007.11.008

MacKinnon, D. P. (2011) 'Integrating mediators and moderators in research design', *Research on Social Work Practice*, 21 (6), pp. 675–681.

MacNeil, A. J., Prater, D. L. and Busch, S. (2009) 'The effects of school culture and climate on student achievement', *International Journal of Leadership in Education,* 12 (1), pp. 73–84.

Marchal, B., Westhorp, G., Wong, G., Van Belle, S., Greenhalgh, T., Kegels, G. and Pawson, R. (2013) 'Realist RCTs of complex interventions – an oxymoron', *Social Science and Medicine*, 94, pp. 124–128. doi:10.1016/j.socscimed.2013.06.025

Maulana, R., Helms-Lorenz, M. and van de Grift, W. (2015) 'Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach', *School Effectiveness and School Improvement*, 26 (2), pp.169–194.

McAleavy, T., Riggall, A. and Fitzpatrick, R. (2015) 'Rapid School Improvement', Education Development Trust.

McDonald, A. J., Ross, S. M., Abney, J. and Zoblotsky, T. (2008) 'Urban school reform: Year 4 outcomes for the Knowledge is Power Program in an urban middle school', in *Annual Meeting of the American Educational Research Association*, 2008.

Melick C, Feldman B, Wilson R. (2008) 'SY07-08 School climate data: Students, parents, and staff'. Retrieved from: http://www.baltimorecityschools.org

Moolenaar, N. M., Daly, A. J. and Sleegers, P. J. (2011) 'Ties with potential: Social network structure and innovative climate in Dutch schools', *Teachers College Record*, 113 (9), pp. 1,983–2,017.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W. *et al.* (2015) 'Process evaluation of complex interventions: Medical Research Council guidance', *BMJ,* 350, p. h1258.

Muijs, D. (2006) 'Measuring teacher effectiveness: Some methodological reflections', *Educational Research and Evaluation*, 12, pp. 53–74.

Muijs, D., Chapman, C. and Armstrong, P. (2011) 'The pedagogy of teach first teachers: What are graduates in an alternative certification programme like as teachers?', in

International Congress for School Effectiveness and Improvement, Limassol, CY, 4–7 Jan 2011.

Mulford, B. and Silins, H. (2003) 'Leadership for organisational learning and improved student outcomes—What do we know?', *Cambridge Journal of Education*, 33 (2), pp. 175–195.

Munro, A., and Bloor, M. (2010) 'Process evaluation: the new miracle ingredient in public health research?', *Qualitative Research*, 10 (6), pp. 699–713. doi: 10.1177/1468794110380522

Nathanson, L., Cole, R., Kemple, J. J., Lent, J., McCormick, M. and Segeritz, M. (2013) 'New York City School Survey 2008–2010: Assessing the Reliability and Validity of a Progress Report Measure' (technical report).

New Schools Network (2015) 'Free School Self-evaluation tool'. Retrieved from http://www.newschoolsnetwork.org/academy-resource-hub/resources/free-school-self-evaluation-form

Oakley, A., Strange, V., Bonell, C., Allen, E. and Stephenson, J. (2006) 'Process evaluation in randomised controlled trials of complex interventions', *BMJ*, 332 (7,538), pp. 413–416.

O'Brien, N., Heaven, B., Teal, G., Evans, E. H., Cleland, C., Moffatt, S., Sniehotta, F. F., White, M., Mathers, J. C. and Moynihan, P. (2016) 'Integrating Evidence From Systematic Reviews, Qualitative Research, and Expert Knowledge Using Co-Design Techniques to Develop a Web-Based Intervention for People in the Retirement Transition', *Journal of Medical Internet Research*, 18 (8).

OECD (2013) *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*, Paris, France: OECD Publishing.

Ofsted (2014) *Preparing a school self-evaluation*, London, Ofsted. Retrieved from http://dera.ioe.ac.uk/20169/7/Guidance for school self evaluation%5B1%5D.doc

Panayiotou, A., Kyriakides, L., Creemers, B., McMahon, L., Vanlaar, G., Pfeifer, M., Matevz, B. and Rekalidou, G. (2014) 'Teacher Behavior and Student Outcomes: Results of a European Study', *Educational Assessment, Evaluation and Accountability*, 26, pp. 73–93.

Pandey, P., Goyal, S. and Sundararaman, V. (2009) 'Community participation in public schools: impact of information campaigns in three Indian states', *Education Economics,* 17 (3), pp. 355–375. doi:10.1080/09645290903157484

Parisi, J. M., Ramsey, C .M., Carlson, M. C., Xue, Q. L., Huang, J., Romani, W. A., McGill, S., Seeman, T. E., Tanner, E. K., Barron, J. and Tan, E.J . (2015) 'Impact of Experience Corps® Participation on School Climate', *Prevention Science*, 16 (5), pp. 744–753.

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, London: Sage Publications Ltd.

Petticrew, M., Cummins, S., Ferrell, C., Findlay, A., Higgins, C., Hoy, C., Kearns, A. and Sparks, L. (2005) 'Natural experiments: an underused tool for public health?', *Public Health*, 119, pp. 751–757. doi: 10.1016/j.puhe.2004.11.008

Porter, A. C., Polikoff, M. S., Goldring, E. B., Murphy, J., Elliott, S. N. and May, H. (2010) 'Investigating the validity and reliability of the Vanderbilt Assessment of Leadership in Education', *The Elementary School Journal*, 111 (2), pp. 282–313.

Posner, B. Z. (2016) 'Investigating the Reliability and Validity of the Leadership Practices Inventory®', *Administrative Sciences*, 6 (4), p. 17.

Raman, A., Ling, C. C. and Khalid, R. (2015) 'Relationship between school climate and teachers' commitment in an excellent school of Kubang Pasu District, Kedah, Malaysia', *Mediterranean Journal of Social Sciences*, 6 (3 S1), p. 163.

Reform Support Network (2015) 'Leading Indicators for School Improvement', A Review of State Agency Practices. Retrieved from

https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/leadingindforschoolimprov.pdf
Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C. and Stringfield, S. (2014) 'Educational effectiveness research (EER): A state-of-the-art review', *School Effectiveness and School Improvement*, 25 (2), pp. 197–230.
Rhode Island Department of Education (2014) *Facilitator's Guide to Quarterly Monitoring of School Reform Plans*. Retrieved from: http://www.ride.ri.gov/Portals/0/Uploads/Documents/Information-and-Accountability-User-Friendly-Data/Transformation/Facilitators-Guide-Quarterly-Monitoring-Fall-2014.pdf
Rogers, P. J. (2008) 'Using programme theory to evaluate complicated and complex aspects of interventions', *Evaluation*, 14 (1), pp. 29–48.
Rosenbaum, P. R. (1989) 'Optimal Matching for Observational Studies', *Journal of the American Statistical Association,* 84 (408), p. 1,024. doi:10.2307/2290079
Rosenbaum, P. R., and Rubin, D. B. (1983) 'The central role of the propensity score in observational studies for causal effects', *Biometrika,* 70 (1), pp. 41–55. doi:10.1093/biomet/70.1.41
Rosenbaum, P. R., and Rubin, D. B. (1984) 'Reducing Bias in Observational Studies Using Subclassification on the Propensity Score', *Journal of the American Statistical Association,* 79 (387), pp. 516–524. doi:10.1080/01621459.1984.10478078
Rowan, B., Miller, R. and Camburn, E. (2009) *School improvement by design: Lessons from a study of comprehensive school reform programs*, Consortium for Policy Research in Education (CPRE), University of Pennsylvania.
Scheerens, J. (2013) 'What is effective schooling? A review of current thought and practice', paper for the International Baccalaureate Organization. Washington, DC.
Scheerens, J. (ed.) (2012) *School leadership effects revisited*. *Review and meta-analysis of empirical studies*, Dordrecht: Springer.
Schlotter, M., Schwerdt, G. and Woessmann, L. (2010) 'Econometric Methods for Causal Evaluation of Education Policies and Practices: A Non-Technical Guide' (4725). Retrieved from London: http://ftp.iza.org/dp4725.pdf
Scott, S. and McNeish, D. (2013) *School leadership review: using research evidence to support school improvement*, London: Department for Education Research Report.
Sebring, P. B., Allensworth, E., Bryk, A. S., Easton, J. Q. and Luppescu, S. (2006) 'The Essential Supports for School Improvement', research report, Consortium on Chicago School Research.
Shadish, W. R. (2012) 'Propensity score analysis: promise, reality and irrational exuberance', *Journal of Experimental Criminology,* 9 (2), pp. 129–144. doi:10.1007/s11292-012-9166-8
Shindler, J., Jones, A., Williams, A., Taylor, C. and Cardenas, H. (2009) 'Exploring below the surface: School climate assessment and improvements as the key to bridging the achievement gap*',* paper presented at the annual meeting of the Washington State Office of the Superintendent of Public Instruction, Seattle, WA.
Shindler, J., Taylor, C., Cadenas, H. and Jones, A. (2003) 'Sharing the data along with the responsibility: Examining an analytic scale-based model for assessing school climate', paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
Smith, P. A. and Sean Kearney, W. (2013) 'The impact of achievement press on student success in elementary schools', *International Journal of Educational Management*, 27 (4), pp. 387–401. doi: 10.1108/09513541311316322
Smith, J. and Todd, P. (2005) 'Does matching overcome LaLonde's critique of nonexperimental estimators?', *Journal of Econometrics*, 125 (1-2), pp. 305–353. doi:10.1016/j.jeconom.2004.04.011

Snijders, T. and Bosker, R. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, London: Sage.

Spillane, J. P., Pareja, A. S., Dorner, L., Barnes, C., May, H., Huff, J. and Camburn, E. (2010) 'Mixing methods in randomized controlled trials (RCTs): Validation, contextualization, triangulation, and control', *Educational Assessment, Evaluation and Accountability*, 22 (1), pp. 5–28. doi:10.1007/s11092-009-9089-8

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E. and Barry, C. L. (2014) 'Using propensity scores in difference-in-differences models to estimate the effects of a policy change', *Health Services and Outcomes Research Methodology,* 14 (4), pp. 166–182. doi:10.1007/s10742-014-0123-z

Vaux, N. D. (2015) 'Academic optimism and organizational climate as predictors of academic achievement and school effectiveness', Ph.D. dissertation, University of Alabama.

Teddlie, C. and Tashakkori, A. (eds.) (2009) *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*, Sage Publications.

Thomson, D. (2016) 'The Short Run Impact of the Building Schools for the Future Programme on Attainment at Key Stage 4', QSS Working Paper 16-07, London, UK: Department of Social Science, UCL Institute of Education. Retrieved from: http://repec.ioe.ac.uk/REPEc/pdf/qsswp1607.pdf

Treadaway, M. (2015) 'We worry about teachers inflating results; we should worry more about depression of baseline assessments', in *Seven things you might not know about our schools*', Education Datalab. Retrieved from: http://educationdatalab.org.uk/2015/03/we-worry-about-teachers-inflating-results-we-should-worry-more-about-depression-of-baseline-assessments/

Turner, J. C. and Meyer, D. K. (2000) 'Studying and understanding the instructional contexts of classrooms: Using our past to forge our future', *Educational Psychologist*, 35 (2), pp. 69–85.

Vanlaar, G., Kyriakides, L., Panayiotou, A., Vandecandelaere, M., McMahon, L., De Fraine, B. and Van Damme, J. (2016) 'Do the teacher and school factors of the dynamic model affect high- and low-achieving student groups to the same extent? A cross-country study', *Research Papers in Education*, 31 (2), pp. 183–211.

Voight, A. and Hanson, T. (2012) 'Summary of existing school climate instruments for middle school', San Francisco: REL West at WestEd.

Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L. and Marder, M. (2012) 'Development of the UTeach observation protocol: A classroom observation instrument to evaluate mathematics and science teachers from the UTeach preparation program', unpublished paper, Southern Methodist University.

The Wallace Foundation (2009) 'Assessing the effectiveness of school leaders: new direction and new processes'.

Wallace, T. L., Kelcey, B. and Ruzek, E. (2016) 'What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod Student Perception Survey', *American Educational Research Journal,* 53 (6), pp. 1,834–1,868. doi: 10.3102/0002831216671864

Werner, H., Karel, S., Jan, V. and Wait, K. (2011) 'Evaluating the difference in employee engagement before and after business and cultural transformation interventions', *African Journal of Business Management*, 5 (22), p. 8,804.

Wholey, J. S., Hatry, H. P. and Newcomer, K. E. (2010) *Handbook of practical program evaluation*, John Wiley & Sons.

Williams, A. D. (2016) 'The School Climate-Student Achievement Connection: If We Want Achievement Gains, We Need to Begin by Improving the Climate', *JSARD*, p. 9.

Wong, V., Valentine, J. and Miller-Bains, K. (2017) 'Empirical performance of covariates in education observational studies', *Journal of Research on Educational Effectiveness*, 10 (1), pp. 207–236. doi:10.1080/19345747.2016.1164781

Yazzie-Mintz, E. (2010) *Charting the path from engagement to achievement: A report on the 2009 High School Survey of Student Engagement*, Bloomington, IN: Center for Evaluation & Education Policy.