BACKGROUND

Implementation and Process Evaluation (IPE) is a crucial component of EEF’s approach to evaluation. Whether or not Impact Evaluation (IE) has found an effect, IPE can help explain why an intervention has or has not been successful, what factors have contributed to this result and what lessons we can learn about educational practice and research. We now have an established statistical analysis guidance which provides best practice advice for evaluators designing and implementing IE. For trials, evaluators also complete a statistical analysis plan (SAP) before data collection, describing in detail how IE statistical analysis will be conducted and reported, in line with pre-specification best practice for impact evaluation (e.g., Montgomery et al., 2018; Schulz, Altman, & Moher, 2010). A similar analysis plan is not required for IPE, and the EEF does not intend to introduce an additional (separate) document for this purpose at this stage. However, we are keen to achieve a similar level of quality and transparency for IPE, which has so far been typically pre-specified only in very general terms. We are, therefore, strongly encouraging evaluators to clarify and justify their IPE design decisions in the evaluation protocol and report, according to the principles outlined further down. We recognise that, by its very nature, IPE research is more exploratory than IE and that intended methods described in the protocol may sometimes be refined or redefined during the project. We encourage evaluators to state in the protocol whether this is anticipated and in what areas, and we recommend that evaluation reports describe clearly any changes made during the project, the reasons for these changes and the likely implications on the overall design and results. The EEF has updated the protocol and report templates accordingly, and the new templates will be used for all new and ongoing projects.

THIS PAPER

This paper recommends a set of key principles for EEF evaluators planning, conducting and reporting IPE, in line with recommended best practice for pragmatic mixed-method and qualitative research. We do not envisage these recommendations resulting in more resources necessarily being allocated to IPE. Rather, we would like to use these principles to encourage evaluators to be more consistently transparent about their design decisions and for more of the IPE planning to occur prior to the protocol being finalised. In some cases, this may mean some redistribution of resources from one stage of the

---

1 In this document, by ‘protocol’ we mean an evaluation protocol used for randomised controlled trials or an evaluation plan (sometimes called a study plan) used for pilots, quasi-experimental studies and other designs.
project to another, for instance by making key decisions about data collection instruments and analysis methods at the start of the project rather than immediately before data collection. This need not mean designing actual data collection instruments before completing the protocol. Rather, we encourage more transparency and early planning regarding the intended content, focus, approach or type of data collection instruments and analyses intended to be used, in line with the research questions and rationale. If IPE follows a sequential design, with one method building on another (e.g., the questions in an interview schedule depend on the results of an earlier survey), where the focus may not be easy to anticipate at protocol stage, this should be stated and justified accordingly. Whether or not anticipated, other details may also change during the research – for instance, the timing of collecting specific data, the types of participants interviewed, the characteristics of the schools sampled and so on. We encourage evaluators to be flexible and pragmatic, to discuss any significant protocol changes with the EEF (and the delivery team) before implementing them, and to describe any such changes in the final report.

This guidance builds on the IPE literature synthesis and introductory handbook written by Humphrey et al. (2016a, 2016b), relevant mixed-method literature and reporting standards (Leech & Onwuegbuzie, 2010; Levitt et al., 2018; Mcrudden, Marchand, & Schutz, 2019; O’Brien, Harris, Beckman, Reed, & Cook, 2014; Pinnock et al., 2017; Tong, Sainsbury, & Craig, 2007), as well as ongoing discussions with our panel of evaluators and other experts. We recommend evaluators consult the introductory handbook (Humphrey et al., 2016b) for more detailed information relevant to EEF evaluations (trials and pilots), in addition to other references cited in this paper.

We are very grateful to the five experts from our evaluation panel who provided feedback on drafts of this guidance.

**IPE PRINCIPLES**

1. **Integration**: IPE and IE should be complementary and fully integrated.
2. **Logic model**: IPE design and integration with IE should be driven by a logic model that includes specific core elements, is examined against the evaluation data and is revised as needed.
3. **Compliance, fidelity and usual practice**: All impact evaluations should include measures of compliance, fidelity and usual practice, in addition to any other relevant implementation dimensions.
4. **Relevance**: IPE design should be tailored and relevant to each individual study.
5. **Pre-specification**: IPE design and analysis should be planned early and pre-specified transparently.
1. Integration: IPE and IE should be complementary and fully integrated.

The evaluation proposal, protocol and report will draw to a similar extent on IE and IPE, including, as far as possible:

a. **Rationale.** IPE and IE should not be regarded as two separate, parallel studies. The case for the study should be made with reference to previous impact, implementation and process evidence, discussing the strength of this evidence and clearly specifying the IE and IPE gaps that the study will address, and how the methods will complement one another.

b. **Research questions.** The IE and IPE research questions should be complementary and together aim to address the study’s overall objective. The evaluator should specify what the overall objective of the study is and clarify how it is reflected in the IE and IPE research questions and methods. We advise against including long lists of research questions without a coherent rationale, where it is not clear how the type and quantity of data collected could provide adequate answers to all the questions listed (White, 2008). We recommend that research questions are overarching but specific, summarising the main lines of enquiry rather than mirroring the actual questions asked in the data collection instruments. We recommend IE and IPE research questions are kept separate, ensuring it is clear what method feeds into each research question, and specifying how impact and process data will be integrated (Oakley, Strange, Bonell, Allen, & Stephenson, 2006). Evaluators should also ensure the IPE research questions reflect the stage of the evaluation (e.g., pilot, efficacy or effectiveness), and what is already known about the process/implementation of the programme from previous evidence.

c. **Research design.** The evaluator should describe the study design and how it addresses each research question, clarifying their rationale for mixing specific research methods in a particular way (Creswell & Creswell, 2018; Gorard, 2013; Morse, 2003; Teddlie & Tashakkori, 2009, 2012) and clarifying how the mix of methods will be integrated so as to explain what works for whom, why and in what circumstances (McCrudden et al., 2019; Plano Clark, 2019). We recommend that methods are selected and justified in line with the mixed-method principle of ‘complementary strengths and non-overlapping weaknesses’ (Johnson & Turner, 2003).

d. **Analysis and interpretation.** For practical reasons (e.g., different teams working on different evaluation areas), the IE and IPE analyses are often conducted and written up separately. Whether or not this is the case, it is important that the two analysis strands are independent

---

2 The level of detail included in the proposal, protocol and report will, of course, be proportional to the expected length of these documents.

3 We recommend that research questions are numbered for ease of reference (e.g., Appendix 1).
of each other, to prevent one set of results influencing the analysis or interpretation of the other (Moore et al., 2015; Oakley et al., 2006). If particular analyses are planned so that they depend on the results of other analyses (e.g., IPE analysis seeking to provide possible explanations for sub-group IE analyses), this should be stated clearly in the protocol and final report. It is equally important that the overall interpretation of the results draws on IE and IPE in equal measure, highlighting any areas where emerging evidence is particularly strong or insufficient.

e. **Reporting.** We encourage evaluators to think of our ultimate audience of teachers, school leaders and policy makers when reporting their findings. A coherent, comprehensive interpretation of the IE and IPE findings should help our audience distil a holistic message about the programme evaluated. The evaluator should comment on any convergent or particularly divergent findings resulting from the different strands of the evaluation, attempting to interpret these with reference to the existing evidence, nature of the programme and context of the intervention. Any apparent contradictions should be accounted for (Seale, 1999), if possible on the basis of the data collected, or at a minimum acknowledged and flagged as an area where more research is needed. Integration must also be visible in similarly high levels of rigour and quality assurance between IE and IPE, which should be planned and described from evaluation proposal stage.

2. **Logic model:** *IPE design and integration with IE should be driven by a logic model that includes specific core elements, is examined against the evaluation data and is revised as needed.*

Terminology use varies in the evaluation literature, Funnell and Rogers (2011, pp. 23–24) listing 22 phrases that are sometimes used interchangeably with ‘logic model’ and sometimes argued to have distinct meanings by different proponents. For the purpose of this paper, by logic model we understand a **visual representation of a programme’s inputs, activities, outputs, outcomes and underlying causal mechanisms** (drawing on Coldwell & Maxwell, 2018; Cooksy, Gill, & Kelly, 2001; Funnell & Rogers, 2011; Kaplan & Garrett, 2005; Knowlton & Phillips, 2012; Renger & Titcomb, 2002). In the event of a null or negative result, a good logic model will guide the evaluator in distinguishing between **theory** failure (the programme does not work as hypothesised to achieve the expected outcomes), **implementation** failure (the programme was not implemented as intended) and

---

methodology failure\(^5\) (inadequate evaluation methods were selected, or suitable evaluation methods were used inadequately) (Coldwell & Maxwell, 2018; Stame, 2010).

The recommended process for developing logic models for EEF-funded evaluations is: i) the delivery team produce an initial draft at application stage focusing on implementation (inputs, activities, outputs and outcomes – see below), based on prior evidence and hypothesised causal mechanisms; ii) the evaluation team revises this initial draft, adding the remaining components following set-up discussions with the delivery team and the EEF; iii) the evaluation and delivery teams refine the model at the IDEA workshop\(^6\) (see Appendix 2, Humphrey et al., 2016b); iv) the evaluator finalises the model and includes it in the evaluation protocol; v) after data collection, the evaluator tests the model by conducting pre-specified quantitative and qualitative analyses to evaluate programme theory (hypothesised causal mechanisms), programme implementation and methodological validity (IE plus IPE); vi) the evaluator revises the model in line with the evidence; vii) the evaluator includes both the initial logic model developed with the delivery team and the revised version based on data\(^7\) in the final evaluation report, describing the revision process and the extent to which the model was supported by the evidence.

Logic models developed for EEF evaluations have so far tended to focus mainly on implementation. In order to improve the quality of our reports and their usefulness in identifying what works in improving teaching and learning, we recommend that evaluators capture both causality and programme implementation in the logic models developed for the evaluations we fund\(^8\). As Rogers (2008, p. 34) explains, simple logic models that focus on implementation and omit underlying programme theory, contextual factors and causal mechanisms risk overstating the contribution of the intervention to the outcomes, as well as providing less useful information for potential replication.

Evaluators may use their own preferred format when developing logic models\(^9\). However, we recommend that the following core components are included as standard, together with a description of the problem addressed and the assumed relationships between the logic model components.

---

\(^5\) Methodology failure can also lead to false positive results (Type I errors), concluding that a programme works when in fact it does not or there is insufficient evidence to support the positive conclusion. Likewise, methodology failure can lead to Type II errors, concluding that a programme does not work when, in fact, it does.

\(^6\) The IDEA (Intervention Delivery and Evaluation Analysis) workshop usually takes place after the second set-up meeting. However, depending on the stage of logic model development and evaluation design, evaluators and developers may decide to have the workshop between the two set-up meetings.

\(^7\) Further updates may be included, depending on project and supporting evidence available (e.g., Coldwell & Maxwell, 2018).

\(^8\) For guidelines on developing effective logic models, with examples of poor and good practice, please see Funnell and Rogers (2011, pp. 277–291) and other relevant references cited in this document.

\(^9\) For useful examples, please see Coldwell and Maxwell (2018), Funnell and Rogers (2011), Knowlton and Phillips (2012).
Depending on the amount of information that needs to be captured (which will vary from programme to programme), evaluators may choose to summarise some of these components in text boxes within the logic model itself or in the descriptive text accompanying it.

a. **Inputs**: the human, material, financial and other resources necessary for delivering the programme;

b. **Activities**: specific actions, processes, tools, technology or events through which the programme is delivered, thus contributing to the identified outputs;

c. **Outputs**: the direct results of the activities through which the programme is delivered to the participants (usually described in terms of number, size or scope of the services and products delivered by the intervention to its participants);

d. **Short-term outcomes**: specific changes in knowledge, skills, attitudes or behaviours experienced by individuals or groups in the short term as a result of programme outputs. In EEF evaluations, this may be the primary outcome, or a proximal outcome measured earlier than a distal (long-term) primary outcome. For instance, if a programme includes a post-test and a delayed post-test, the post-test may be the short-term outcome and the delayed post-test the long-term outcome. This will depend on each individual project, its theory of change and the length of time after which change can realistically be expected. Unintended short-term outcomes or consequences should also be anticipated and specified.

e. **Long-term outcomes**: changes experienced by individuals, groups or organisations over time as a consequence of programme outputs. In EEF evaluations, this may be the primary outcome or a longitudinal follow-up of the primary outcome. Any unintended long-term outcomes or consequences should be specified as well.

f. **Mediators and causal mechanisms**. Mediators are intermediate variables in a causal sequence from intervention to outcome (intervention X causes mediator Z, mediator Z causes outcome Y). An example could be a mentoring programme (X) that aims to increase student attainment (Y) by increasing student attendance (Z). Or a parenting intervention (X) aiming to increase pupil literacy (Y) by enhancing parental literacy awareness (Z). A good evaluation will specify and test the assumed causal mechanisms linking these three elements, as well as the direction of any causality (Coldwell & Maxwell, 2018; Morris, Edovald, Lloyd, & Kiss, 2016). Mediators help answer the question: ‘How does the intervention work?’ and, together with the causal mechanisms or processes, represent the programme theory. In EEF trials, they are often defined as secondary outcomes. Short-term outcomes can also be mediators for long-term outcomes, in programmes where the short-term outcome must be achieved in order to trigger the long-term outcome. Mediators and

---

10 Though, of course, a secondary outcome will only be a mediator if in line with this definition.
causal mechanisms (or ‘change mechanisms’) are sometimes used interchangeably in the evaluation literature, though this is not necessarily appropriate. As Tryon (2018, p. 626) explains, establishing a mechanism requires three steps: ‘First, investigators must provide statistical evidence of mediation. Second, investigators must provide additional experimental evidence of cause. Third, investigators must then be able to explain the process(es) by which causation works.’ (For more information about mediation analysis and causal mechanisms, please see Gardner, Hutchings, Bywater, & Whitaker, 2010; Imai, Keele, Tingley, & Yamamoto, 2011; Judd & Kenny, 1981; Kazdin, 2007; MacKinnon, 2007; MacKinnon, Fairchild, & Fritz, 2007; MacKinnon, 2011; Myrberg & Rosén, 2008; Raviv, Kessenich, & Morrison, 2004. Wu & Zumbo, 2008)

g. **Moderators and other contextual factors.** Moderators are variables that modify the form or strength of the relation between intervention and outcome. These may include any individual characteristics (e.g., gender, age, socio-economic status, disability), institution characteristics (e.g., school type, geographic location) or baseline measures that may result in differential impact on different types or sub-groups of participants (Baron & Kenny, 1986; Howe, 2019). Moderators may also be various contextual factors (e.g., workload, unconscious bias, supportive leadership) or implementation dimensions such as fidelity, quality of delivery, adaptation or responsiveness. Different contextual factors will have a moderating effect at different points in the process (Coldwell, 2019), and evaluators should discuss and test their hypotheses in this regard. Moderators help answer the question: ‘For whom does the intervention work and under what circumstances?’

The logic model should support every stage of the evaluation, from design to interpretation and reporting (Cooksy et al., 2001; Knowlton & Phillips, 2012; Markiewicz & Patrick, 2016). The research design, research methods and research questions should clearly reflect the logic model developed by the evaluator at the start of the intervention, in collaboration with the delivery team. Data collection and analysis should be planned so that the components, relationships and causal assumptions included in the initial logic model can be tested, and the extent to which they have been found to be supported by evidence should be discussed in the report. The evaluation protocol and report should include a description of the logic model with reference to the problem or evidence gap being addressed and the model should be referred to in the final analysis and interpretation of results. The report should include both the initial logic model developed by the evaluator in collaboration with the delivery team and the revised logic model updated following data analysis (unless the data confirmed the logic model fully and no revision was deemed necessary). The reasons and process for updating

---

11 Depending on the programme content and outcomes, these may also be conceptualised as mediators (Berkel, Mauricio, Schoenfelder, & Sandler, 2011).
the logic model should be described in the report, highlighting any areas where supporting evidence was particularly strong or particularly weak (McLaughlin & Jordan, 1999; Morell, 2018).

3. Compliance, fidelity and usual practice: All impact evaluations should include measures of compliance, fidelity and usual practice, in addition to any other relevant implementation dimensions.

a. Compliance: All impact evaluations will include a measure of compliance defined specifically for the project in question, with all parameters clearly defined. This includes defining the variable (or variables\textsuperscript{12}) to be used in compliance analysis, the type of variable included in the analysis (e.g., continuous, categorical), the level it will be measured at (e.g., pupil, teacher) and any threshold for the participant to be considered compliant, if applicable. In the context of EEF trials, by compliance we mean \textit{the extent to which the critical ingredients of the intervention are delivered to and/ or received by the target participants}. This is often a measure of dosage (dose plus frequency), but different variables may be used, depending on the intervention. Examples of compliance measures relevant to EEF trials include: the number of CPD sessions attended by the teacher, the number of lessons attended by the pupil in which x teaching approach was used, the cumulative length of exposure to y input, the frequency of z event during the intervention. Compliance analysis – usually part of IE – will require good quality data in order to be robust. It is important, therefore, that evaluators and developers decide at set-up what compliance indicators will be used, so provision to collect these is built into the evaluation design and protocol.

b. Fidelity. All evaluations will include a measure of implementation fidelity defined specifically for the project. By implementation fidelity we mean \textit{the degree to which the intervention is delivered as intended or prescribed} (Berkel, Mauricio, Schoenfelder, & Sandler, 2011; Carroll et al., 2007). The evaluation and delivery teams will agree a definition of fidelity for their evaluation, along with the specific measures by which fidelity will be assessed. Depending on project, these may include content, coverage, dose, frequency, duration or moderators such as quality, adaptation or responsiveness. In some trials, it may be difficult to differentiate between compliance and fidelity, or specific fidelity markers could be used as measures of compliance in CACE\textsuperscript{13} analysis). Evaluators, in collaboration with the intervention delivery team, should clarify whether this is the case. The parameters for all the components included must be clearly specified in the evaluation protocol and report.

\textsuperscript{12}If several measures of compliance are used, an aggregate value will need to be used in compliance analysis (see latest EEF statistical analysis guidance).

\textsuperscript{13}complier average causal effect (Panayiotou, Humphrey, & Hennessey, 2019; Peugh, Strotman, McGrady, Rausch, & Kashikar-Zuck, 2017); see also the latest EEF Statistical Analysis Guidance
c. **Usual practice**: All trials will capture usual practice at baseline and post-intervention in all trial arms (intervention and control). For intervention arms, this should clarify whether the intervention is in addition to or instead of usual practice. Pilots will include a measure of usual practice focusing on differentiation from the intervention.

The evaluator may add other implementation dimensions they and/or the delivery team consider relevant to the intervention\(^4\). The evaluator should also specify whether any of the implementation dimensions are expected to be mediators or moderators and, if so, should ensure that they are captured in the logic model. They should also specify how they expect the implementation dimensions to interact and how these interactions will be tested (e.g., structural equation modelling, path analysis).

4. **Relevance**: IPE design should be tailored and relevant to each individual study.

The evaluator should propose a mix of methods that they consider best suited to each individual study, which will explore how, for whom and in what circumstances the intervention leads to the intended outcomes, and which will be agreed at set-up with the delivery team and the EEF. The relevance of the methods should be clear from the study rationale and approach to addressing the evidence gaps identified. The nature and aims of the study should be taken into account when deciding the most appropriate research methods, including, but not limited to, the following considerations:

a. **Pilot studies** are most likely to focus on preliminary evidence of promise, feasibility of the intervention and its scalability (or readiness for trial). Methods exploring intervention facilitators and barriers to implementation are likely to be particularly helpful. Programme theory is likely to be in an early developmental stage, as the relationships between its components may not yet be entirely clear, though they may be related to established causal mechanisms from other contexts. However, evaluators should endeavour to begin to test the relationships between the various components included in the logic model. Methods and research questions that differentiate between intervention feasibility and evaluation feasibility will help inform decisions about scalability or trialling.

b. **Efficacy trials** emphasise implementation fidelity and testing the impact of the programme under ideal conditions. Methods that explore the programme theory and relationships between different components of the logic model will be particularly helpful. Collecting data that will help test the logic model causal assumptions is also important in efficacy trials, particularly to inform scale-up. Evaluators should specify in advance how causal mechanisms

\(^4\) For other implementation dimensions, please see Humphrey et al. (2016b). Evaluators should include only additional dimensions that are directly relevant to the intervention being evaluated.
will be tested and how they will differentiate between programme (theory) failure and implementation failure in the case of a null result (Morris et al., 2016).

c. **Effectiveness trials** focus on real-life implementation, therefore allowing for greater variability and contextual adaptation. Capturing implementation variability and causal relationships with moderators/ contextual factors will therefore be particularly important at this stage.

5. **Pre-specification:** *IPE design and analysis should be planned early and pre-specified transparently.*

Evaluators should aim to decide the specifics of their IPE at the set-up stage and include these in the evaluation protocol (and final report). Bearing in mind that we do not require an analysis plan for IPE, it is important that key decisions about instruments and analysis are fully discussed and pre-specified in the protocol. We recognise that evaluators have more time to develop the statistical analysis plan for IE, but we intend IPE pre-specification to be far less onerous. At protocol stage, we require transparency about the intended overall approach, methods, types and numbers of participants or data sources, and not the actual data collection instruments, coding frameworks or other specific details. As indicated above, if any aspects of IPE are expected to change, or are planned so that their design will depend on methods implemented earlier, this should be decided and clarified at protocol stage. We recognise that sometimes unexpected changes are necessary later on. We encourage evaluators to be flexible and make decisions in the interest of conducting the best research possible. For example, if, despite all efforts, the response rate for a post-intervention survey is too low, the evaluator may decide to replace the survey with a qualitative method that will require fewer participants. Or an unexpected contextual variable proves important and the evaluator decides to collect additional data for a stronger interpretation. Such changes may not have been anticipated at protocol stage and updating the protocol for this reason alone may not always be justified. In such cases, we ask that evaluators discuss any significant protocol changes with the EEF (and delivery team) before implementing them, and describe the changes in the final report, including a clear rationale, with likely implications on the overall design and results.

In order to improve the quality, usefulness and comparability of our evaluation reports, we recommend at least the following details are pre-specified in the protocol and then included – in more detail, as appropriate – in the final evaluation report. As with IE, if aspects of IPE design change during the evaluation, the protocol may need to be revised accordingly, depending on the magnitude of the changes. The final report should acknowledge and justify any differences from the protocol, reflecting on any implications that these changes may have had on the evaluation.
a. **Methods and instruments.** The evaluator will specify clearly which research methods will contribute to answering which research questions, how data will be collected and analysed for each research method, how many participants or data sources each method is intended to draw on, what implementation dimensions are being explored and how they relate to the research questions and logic model. We recommend the table in Appendix 1 is adapted to the project and included in the evaluation protocol and report, for a comprehensive at-a-glance overview of intended methodological complementarity.

If a *case study* approach is taken, evaluators should clearly define their case unit (e.g., school, classroom, region), describe their criteria for case selection and explain how each research method will inform their understanding of the case units (Hamilton & Corbett-Whittier, 2012; Stake, 1995; Yin, 2018). Evaluators should specify how they will use within-case, between-case and deviant/ negative case analysis (Seale & Silverman, 1997), if applicable, to inform their emerging interpretations. Simply collecting data through a variety of methods in a variety of settings does not in itself represent case-study research unless the methods are used holistically to inform understanding of the case unit. As Miles et al. (2019) explain, drawing on Ragin (1987), a *case-oriented approach* considers the case as a whole entity, looking for associations and configurations within the case, before turning to comparative analysis of a small number of cases. The researcher taking this approach will look for underlying similarities and differences in order to develop an interpretation. In contrast, a *variable-oriented approach* is conceptual and theory-driven, casting a wider net over a number of cases, but with a focus on variables and their relationships, rather than cases. In this latter approach, the details of any specific case are less important than the broad patterns found across a variety of cases, and little explicit case-to-case comparison is done. Evaluators should clearly define and justify the approach taken, with reference to the study rationale and research questions.

As for IE outcome measures, the type and broad content (focus) of the IPE data collection instruments should be decided at set-up, to ensure adequate integration and complementarity. If interviews, questionnaires or observations are used, the evaluator should specify whether these will be structured, unstructured or semi-structured. An indication of likely content and key purpose within the IPE design should be provided at protocol stage, along with a brief description of the process for developing the data collection instruments. Any piloting or validation exercises should be described.

The intended approach to data analysis should also be specified in advance, explaining how the decision has been made and why this is the most appropriate approach for the project.
If responses or transcripts will be coded, the approach to coding should be clarified (e.g., deductive – if the coding framework is developed before the transcripts are examined, inductive – if the codes are derived by examining the transcripts, or mixed) (Boyatzis, 1998; Fereday & Muir-Cochrane, 2006). When deductive coding is used, the evaluator may already have a coding framework based on prior research evidence, practical experience or other factors (e.g., a thematic funding round focusing on specific outcomes). In such cases, it would be helpful to include the coding framework in the protocol, even if very high level. If the coding framework has not yet been produced, but the likely areas of interest are known (e.g., logic model components, unintended consequences, engagement barriers or facilitators), these should be specified in the protocol. If a mixed (deductive-inductive) approach is used, the evaluator starting with some pre-determined codes but also looking for new codes to emerge from data analysis, this should also be described in as much detail as available at protocol stage. For inductive coding, the intended approach to interrogating the data and minimising bias should be described. Likewise, if thematic analysis is used, the evaluator should clearly describe their approach to developing the themes, along with the measures taken to minimise bias and selective choice of quotations to support pre-established expectations (Jackson & Bazeley, 2019; Miles et al., 2019; Saldaña, 2015).

We strongly recommend that evaluators avoid collecting qualitative data without planning to analyse them (i.e., selectively picking quotations to support an emerging interpretation). If qualitative methods such as interviews or focus groups are used without analysing the transcripts, the evaluator must critically consider and justify the usefulness of such qualitative methods within the project, as well as the financial, ethical and data protection implications of collecting data that will not actually be used in the evaluation.

For quantitative IPE methods such as surveys, observations, tests or other methods using numerical data, the evaluator should clearly specify in the protocol how the data will be analysed, for example, using descriptive, inferential, parametric or non-parametric statistics, cross-tabulation, data visualisation (Cohen, Manion, & Morrison, 2017; Field, 2017; Gray, 2017; Tabachnick & Fidell, 2007).

Any post-hoc analytical decisions should be described and justified in the final report, detailing the measures taken to minimise bias, as well as implications on the overall design and results.

b. Participants and sampling. Details should be provided about the type and number of participants and other data sources selected for each research method (see overview table in Appendix 1). Evaluators should clarify their approach to selecting participants and any
sampling criteria used, for each IPE method, including control group data sources. (The
nature of participants and sampling approaches will be different for, e.g., surveys, case
studies or individual interviews, while the unit of analysis for sampling purposes may be a
school, a class, an individual etc.) All decisions should be clearly justified, with limitations
discussed and mitigated.

c. **Rigour.** The protocol and report should also describe evaluators’ approach to minimising
bias and ensuring rigour (i.e., reliability and validity for quantitative research; dependability
and trustworthiness for qualitative research; legitimacy for mixed-methods) (Kincheloe &
Berry, 2004; Levitt et al., 2018; O’Cathain, 2010; Seale, 1999; Seale & Silverman, 1997). This
may include measures of inter-rater reliability, member checking, audit trail, triangulation
etc.
### APPENDIX 1: METHODS OVERVIEW

Table x. IPE methods overview (example – please adapt)

<table>
<thead>
<tr>
<th>Research methods</th>
<th>Data collection methods</th>
<th>Participants/data sources (type, number)</th>
<th>Data analysis methods</th>
<th>Research questions addressed</th>
<th>Implementation/logic model relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tests</strong></td>
<td>paper tests (post-intervention)</td>
<td>pupils (6,000)</td>
<td>regression; path analysis</td>
<td>RQ2</td>
<td>secondary outcomes; mediators; causal mechanisms</td>
</tr>
<tr>
<td><strong>Surveys (pre/post)</strong></td>
<td>online questionnaires</td>
<td>headteachers (150)</td>
<td>descriptive statistics; correlations; exploratory factor analysis</td>
<td>RQ3, RQ5</td>
<td>usual practice; moderators; fidelity; quality; cost</td>
</tr>
<tr>
<td><strong>Interviews</strong></td>
<td>structured interviews</td>
<td>parents (20)</td>
<td>deductive coding; thematic analysis</td>
<td>RQ4</td>
<td>responsiveness; quality</td>
</tr>
<tr>
<td><strong>Document analysis</strong></td>
<td>attendance register</td>
<td>CPD attendance registers (20)</td>
<td>frequency counts; regression</td>
<td>RQ5</td>
<td>compliance; moderators</td>
</tr>
<tr>
<td><strong>Case studies</strong></td>
<td>unstructured interviews</td>
<td>teachers (10)</td>
<td>inductive coding; thematic analysis</td>
<td>RQ2</td>
<td>context; moderators</td>
</tr>
<tr>
<td>(10; case study unit = school; analytical approach = methodological and participant triangulation)</td>
<td>semi-structured focus groups</td>
<td>pupils (4 focus groups x 5 pupils = 20)</td>
<td>deductive/inductive coding; thematic analysis</td>
<td>RQ2, RQ4</td>
<td>responsiveness; adaptation</td>
</tr>
<tr>
<td></td>
<td>structured observations</td>
<td>intervention lessons (2 lessons x 10 case study schools) = 20</td>
<td>frequency counts; deductive coding; cross-case analysis</td>
<td>RQ2</td>
<td>fidelity; adaptation</td>
</tr>
<tr>
<td></td>
<td>document analysis</td>
<td>school websites (10); Y9 syllabus (1); Y9 lesson plans (2 x 10 case study schools) = 20</td>
<td>within-case analysis; cross-case analysis</td>
<td>RQ2</td>
<td>context</td>
</tr>
</tbody>
</table>
**Appendix 2: IDEA Workshop Checklist**

We recommend evaluators and developers use this checklist to guide their IDEA workshop discussion, in addition to the TiDieR checklist. The agreed content will then be included in the evaluation protocol (and final report).

1) **Logic model:** Have all core components been clearly defined and represented in the model? Please summarise:
   - inputs:
   - activities:
   - outputs:
   - short-term outcomes (including unintended):
   - long-term outcomes (including unintended):
   - mediators and causal mechanisms:
   - moderators and other contextual factors:

2) **Compliance and fidelity:** Have compliance, fidelity and any other relevant implementation dimensions been clearly defined and pre-specified, including any sub-components? Please summarise:
   - compliance
   - fidelity
   - other?

3) **Usual practice:** Is it clear how usual practice will be elicited, pre- and post-intervention, in all trial arms? Please summarise:
   - intervention
   - control

4) **IPE methods overview:** Has the methods overview table been completed?

<table>
<thead>
<tr>
<th>Research methods</th>
<th>Data collection methods</th>
<th>Participants/data sources (type, number)</th>
<th>Data analysis methods</th>
<th>Research questions addressed</th>
<th>Implementation/logic model relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

15
REFERENCES


