

Implementation and process evaluation guidance for EEF evaluations



Updated August 2022

BACKGROUND

Implementation and process evaluation (IPE) is a crucial component of EEF's approach to evaluation, whether it is evaluating specific interventions or programmes, or wider strategies or practices. An IPE can help explain why an intervention or a project has or has not been successful, what factors have contributed to this result, what improvements or refinements could be made, and what lessons could be learned about wider educational practice and research. This guidance sets out the key principles for EEF evaluators in planning, conducting and reporting IPE, in line with recommended best practice for pragmatic mixed-method research. This is to ensure consistency and transparency in design decisions, and appropriateness of the design for the study. It 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; McCrudden et al., 2019; O'Brien et al., 2014; Pinnock et al., 2017; Tong et al., 2007), as well as ongoing discussions with our panel of evaluators and other experts. We encourage evaluators to follow this guidance when designing the IPE and writing the protocol and report. We also recommend that evaluators consult the introductory handbook (Humphrey et al., 2016b) for more detailed information and case-study examples from past EEF evaluations (trials and pilots), in addition to other references cited in this paper.

We are keen to achieve a similar level of quality and transparency for IPE as with our standards for implementing impact evaluations (IE) and encourage evaluators to design IPEs adhering to the key principles outlined in this document. We recognise that the intended IPE methods may require adapting or refining during the course of the project in response to emerging contextual or implementation challenges. We will support evaluators to make reasonable and appropriate adaptations. This guidance has been updated to incorporate lessons learned from the [IPE Quality Pilot](#),¹ feedback from our evaluation panel and EEF's internal reflections.² The EEF has updated its

¹ This study was commissioned to review and assess the quality of IPE designs in existing EEF evaluations (Maxwell et al., 2021).

² We will only expect projects that are set up after August 2022 to adhere to the updated guidance and templates.

protocol and report templates accordingly, and the updated templates will be used for all new projects.

SUMMARY OF KEY UPDATES

- We recommend including an IPE question to understand the impact of the intervention on disadvantaged pupils, including, for example, exploring barriers or facilitators that may be specific to disadvantaged pupils (p.6).
- An IPE protocol checklist is included to highlight the key information we expect evaluators to specify in the protocol (Appendix 4).
- The list of IPE dimensions has been updated to include perceived impact, cost, context/moderators and mediators (Appendix 2).
- The 'Logic model' principle has been removed from this update. The role of the logic model in IPE and IE integration is discussed under the 'Integrated' principle. A more in-depth guidance on how to develop a logic model for evaluations will be published as a separate stand-alone document.
- The 'IPE dimensions' principle has been merged and discussed under 'Relevance' in this update.
- We recommend specifying the IPE dimensions relevant for the evaluation before framing the IPE research questions (p.4 and Appendix 1).
- Additional guidance is included on specifying analysis approach (e.g., what data will be analysed, integrated or triangulated, and by what method) (p.8).
- Inclusion of guidance and worked examples around capturing usual practice, programme differentiation, context, unintended consequences and contamination (pp.5–6 and Appendix 3).
- Additional guidance is included on specifying sampling approaches and timing of data collection (p.7).
- Additional guidance is included on reporting and interpreting qualitative findings (e.g., avoid using numbers or proportions) (p.9).

IPE PRINCIPLES

1. **Integrated:** IPE and IE designs should be complementary and reflect the logic model.
2. **Relevant:** IPE design should be tailored and relevant to each individual study.
3. **Transparent and flexible:** IPE design and analysis should be planned early alongside the IE but be responsive and adaptive to programme implementation needs as the study progresses.

1. Integrated: IPE and IE designs should be complementary and reflect the logic model

Study rationale and research questions. The evaluator should specify what the overall objective of the study is and clarify how it is reflected in the IE and IPE research questions³ and methods, and how the methods will complement one another. The logic model developed in collaboration between the evaluation and delivery team should be used to support the design of the IPE research questions, and its integration with IE.⁴ Evaluators should consider including IPE research questions that will support the IE analysis to understand the impact of the intervention on disadvantaged pupils, including for example, exploring barriers or facilitators that may be specific to disadvantaged pupils.

Analysis. It is important that the two evaluation strands (IE & IPE) are analysed independently of each other, to prevent one set of results influencing the analysis of the other (Moore et al., 2015; Oakley et al., 2006). IPE protocols should include detail on how data from different IPE methods will be analysed and how findings will be synthesised. If particular analyses are planned so that they depend on the results of other analyses (e.g., IPE analysis seeking to provide possible exploratory sub-group IE analyses), this should be stated clearly in the protocol and in the final report.

Reporting. The IE and IPE methods and results sections are kept separate in the report so that it is clear which method feeds into each research question. However, evaluators should ensure that the overall interpretation of the results draws on findings from the IE and IPE, highlighting areas in the logic model where findings are supported (or not). This helps ensure that IPE and IE elements are fully integrated.

The evaluator should comment on any convergent or divergent findings, 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 identified as an area where more research is needed.

2. Relevant: 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 the specific evaluation they are undertaking. Methods should enable exploration of how, for whom and in what circumstances the intervention leads to the intended outcomes, and these 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 IPE design should be tailored

³ We recommend that research questions are numbered for ease of reference (e.g., Appendix 1).

⁴ We are currently developing a separate guidance on Theory of Change and Logic Model.

according to the study type, and the implementation/process factors and dimensions that are the most useful and relevant for the study.

Study types: The nature and aims of the study should be considered when deciding the most appropriate research methods, including, but not limited to, the following considerations:

Pilot/developmental 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.

Efficacy trials emphasise implementation fidelity and testing the impact of the programme under ideal conditions (asking the question ‘could this programme be implemented as intended?’). Methods that explore the programme theory and relationships between different components of the logic model will be particularly helpful. It is important, in efficacy trials, to collect data that will help test the causal assumptions of the logic model, to inform potential scale-up. Evaluators should specify in advance how causal mechanisms will be identified and tested and how, in the case of a null result, it will be possible to differentiate between programme (theory) failure and implementation failure⁵ (Morris et al., 2016).

Effectiveness trials focus on real-life implementation, therefore allowing for greater variability and contextual adaptation. It is particularly important for effectiveness trial IPEs to capture information about changes that were made to the programme following efficacy trial, enabling it to operate at greater scale, the implications of these changes, and any variability in implementation. It is also important to explore any causal relationships between outcomes and moderators and/or contextual factors.

Scale-up evaluations focus on understanding how interventions are delivered at scale, where there is likely to be a high degree of heterogeneity in implementation and delivery. Understanding scaling mechanisms for reach, uptake, acceptability and adaptation is particularly important at this stage, as is a description of what the approach to scale-up looks like in practice.

⁵ In trials where there is a CPD component, implementation failure comprises both issues related to the delivery of the CPD and implementation of the learnings from the CPD failure as participants apply the learning in the school or other settings.

IPE dimensions: Evaluators and delivery teams should agree on the IPE dimensions that are the most relevant for the study at project set up and use these to structure the research questions (see list of IPE dimensions in Appendix 2 and Humphrey et al., 2016b, p.6). We advise developing a concise set of research questions with a coherent rationale, and clarity about the type and quantity of data that will be collected to provide adequate answers to them (White, 2008). All IPE designs should specify questions related to fidelity and usual practice, in addition to any other dimensions that the evaluation and/or the delivery team consider relevant to the intervention. The evaluator should also specify whether any of the implementation dimensions are expected to be outcome mediators or moderators and, if so, ensure that they are captured in the logic model.

Fidelity. By fidelity we mean the degree to which the intervention is implemented as intended or prescribed (Berkel et al., 2011; Carroll et al., 2007). In the IPE, fidelity should be considered more broadly than in the IE, where the focus is on quantitative measures of compliance. Compliance measures seek to understand the extent to which the critical ingredients of the intervention (such as attendance at training) are delivered to and/or received by the target participants. In contrast, fidelity can be captured by both quantitative and qualitative data and sits on a spectrum across different dimensions (e.g., quality, dosage, adaptability). 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. The parameters for all the fidelity components included must be clearly specified in the evaluation protocol and report. The evaluators should discuss and agree with delivery teams the extent to which the intervention should be faithfully adhered to and/or adapted.

Usual practice. All trials will capture usual practice at baseline and post-intervention in all trial arms (intervention and control). The purpose of collecting this data is to inform the following:

1. **Programme differentiation** – the extent to which key elements of the programme are different (or not) from what is ‘business-as-usual’ in schools.
2. **Evidence of contamination** – indications of programme content or pedagogy adopted by the control group.
3. **Unintended consequences** – indications of displacement of other practices or provision as a result of delivering the programme.
4. **Context** – circumstances under which there are deviations from usual practice (e.g., policy changes to the curriculum) and how this may influence the outcomes investigated.

Where possible, quantitative usual practice data (e.g., times per week or number of hours delivering a practice) should be collected and analysed in order to assess any changes to practice that have occurred in control schools. For intervention schools, the programme description should clarify

whether the programme delivery is in addition to, or instead of, usual practice and should clarify which are the key elements that set the programme apart from business-as-usual. For control arms, information about the nature of existing practice as well as existence of similar programmes should be captured, both in relation to programme content and pedagogical practices. For example, survey data should capture the types of activities that are similar in nature to the intervention as well as those that might achieve similar outcomes to the intervention. Details about the context in which the usual practices are described should also be included (e.g., changes to the curriculum or school funding).

We expect that evaluators will normally collect usual practice information through pre- and post-intervention surveys (see Appendix 3 for recommendations on what to include in the survey). However, evaluators may wish to collect additional data on usual practice through other methods.

3. Transparent and flexible: IPE design and analysis should be planned early alongside the IE but be responsive and adaptive to programme implementation needs as the study progresses

Evaluators should decide the specifics of their IPE approach at the set-up stage and include these in the evaluation protocol (and the final report). However, the design should be responsive and adaptive to programme implementation needs or unexpected changes in context.

At protocol stage, we require clear justification on the choices of methodology and analytical approaches, including indicative sampling criteria and sample sizes, recruitment approaches and ethics, data collection and intended analyses (i.e., what data will be analysed, integrated or triangulated, and by what method).

We encourage evaluators to be flexible and make decisions in the interest of conducting research that reflects the lowest risk of bias and can provide the most useful results. For example, if recruitment to the trial was a challenge, the evaluator may decide to collect additional data to understand the barriers to recruitment and/or implications to fidelity. Where adaptation in the IPE is reasonable and a clear justification is provided, EEF will take cost or timeline implications forward for consideration. We ask that evaluators document the rationale for changes in protocol updates and describe these clearly in the final reports.

Below is some guidance around methodology, data collection, analysis and reporting that evaluators should consider to ensure clarity and transparency of the IPE. We have also included a checklist of key IPE elements that we recommend including at the protocol stage (see Appendix 4).

Align IPE dimensions and research questions with IPE methodology. The study design should specify clearly which research methods will contribute to answering which research questions, reflecting on which IPE dimensions. It should also state how data will be collected and analysed for each research method, and how many participants or data sources each method is intended to draw on. We recommend that all protocols and reports should include the table in Appendix 1, adapted to the individual project, for a comprehensive overview of the IPE design.

Describe the methodological approach. If a mixture of research methods is used to answer the same research question, a clear rationale for why, when and how, and the plan for analysis should be included in the protocol. For example, the benefits of gathering multiple forms of data should be made clear and there should be clarity about whether the data will be collected and analysed sequentially or concurrently.

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 or between-case analysis if applicable, to inform their emerging interpretations (Seale & Silverman, 1997). For example, they should specify whether the focus is on configurations and associations *within* a specific case (case-oriented approach); or where broader patterns are identified by comparing variables *across* a number of cases (variable-oriented approach) (Miles et al., 2019; Ragin, 1987). 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. Evaluators should clearly define and justify the approach taken, with reference to the study rationale and research questions.

If interviews, surveys or observations are used, the evaluator should specify what form these will take (e.g., structured, unstructured or semi-structured for interviews; pre-coded, open or mixed for surveys and observations). 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.

Specify approach to sampling and data collection. Evaluators should also clarify their approach to selecting participants and any sampling criteria used, for each IPE method, including control group data sources. Ideally, the sample for qualitative research should either be random or purposive, rather than a convenience sample. A purposive sampling approach could be based on fidelity to the programme (e.g., sampling schools with high implementation fidelity and schools with low implementation fidelity), school characteristics (e.g., urban/rural, setting/streaming practice or

high/low proportion of FSM pupils), to identify barriers or enablers to implementation. If any data are collected to inform the sample, this should be justified and specified in the protocol and evaluation report. All decisions should be clearly justified, with limitations discussed and mitigated. This includes why the proposed sample size will be sufficient to address the research questions, how and why research participants or events are selected, and if any criteria are applied to identify sampling saturation.⁶ Some research suggests that 6–7 interviews will capture the majority of themes in a homogenous sample (6 interviews to reach 80% saturation), with the range extended to 11–12 interviews to reach a higher degree of saturation (Guest et al., 2020). However, this depends on the heterogeneity of the population, selection criteria and practical constraints (e.g., time, money and recruitment). The more heterogeneous the population is, the more interviews are required.

Protocols and reports should indicate when data will be collected, the sequence of data collection and whether any data collected will be used to inform sampling or further analysis.

Specify how quantitative data will be analysed and report findings appropriately. For quantitative IPE methods such as surveys or other methods using numerical data, the evaluator should specify in the protocol how the data will be analysed, for example, using descriptive, inferential, parametric or non-parametric statistics, cross-tabulation or data visualisation (Cohen et al., 2017; Field, 2017; Gray, 2017; Tabachnick & Fidell, 2007). The protocol should specify whether a cross-sectional or longitudinal approach to analysis is taken (e.g., comparing survey responses from a single or multiple timepoints), and should describe and justify any triangulation (e.g., of different types of samples such as pupils, teachers, parents to collect data on the same topic). 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.

Specify how qualitative data will be analysed. If responses or transcripts will be coded, the approach to thematic analysis and coding should be clarified in the protocol (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 a mix where new codes are added to a set of pre-determined codes or themes) (Boyatzis, 1998; Fereday & Muir-Cochrane, 2006). We expect most EEF evaluations to use a deductive approach with prespecified themes based on the implementation dimensions and research questions, or a mixed approach. Inductive approaches are usually used with exploratory studies, especially where there is a completely open, narrative/biographical interview approach, which are less appropriate for EEF studies.

⁶ Saturation refers to the point at which no new information is being found whereby the [researcher] can develop properties of the category (Glaser and Strauss, 1967, p.61)

If data collected via different methods is to be triangulated or integrated, the approach taken should be clearly specified in the protocol (e.g., observation and interview data integrated, with both datasets analysed thematically in the same coding frame).

Report qualitative findings appropriately. The EEF considers that quantitative reporting of qualitative data is generally not appropriate because the data is typically drawn from very small and unrepresentative samples (Ritchie et al., 2014). The purpose of gathering qualitative data is to understand the range and diversity, rather than the frequency, of experiences. However, evaluators are encouraged to provide some indication of the strength of findings, to avoid approaches or claims that the data do not support being applied more widely than appropriate. If evaluators report frequencies and proportions for qualitative data, then a justification should be provided.

Quotes should be used sparingly. Evaluators should consider how quotes are selected to minimise bias to support pre-established expectations or an emerging interpretation (Jackson & Bazeley, 2019; Miles et al., 2019; Saldaña, 2015). Long quotes or numerous quotes often indicate inadequate analysis – we recommend selecting a single quote that best illustrates a key point made following analysis of a body of data.

Describe the quality assurance process. This may include multiple independent scoring of data to check for inter-rater reliability, audit trail or triangulation to minimise bias and ensure 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).

APPENDIX 1: EXAMPLE OF A METHODS OVERVIEW TABLE

The following table provides an example on how to populate the Methods Overview Table. The IPE dimensions and research approach taken should be tailored to the specific programme or projects. Appendix 2 includes a long list of IPE dimensions that should be considered. Evaluators should ensure that any data related to causal mechanisms in the logic model and contextual variables that can affect the causal mechanisms and outcomes are also captured in the table.

IPE dimension	RQ addressed	Research methods	Data collection methods	Sample size and sampling criteria	Data analysis methods
Fidelity	RQ1, 2	Observations	structured observations	10 schools (urban/rural)	frequency counts; deductive coding;
		Surveys (pre/post)	online questionnaires	150 headteachers (all)	descriptive statistics; correlations; exploratory factor analysis
Quality	RQ 3	Interviews	structured interviews	20 parents (high/low fidelity)	deductive coding; thematic analysis
		Surveys (pre/post)	online questionnaires	150 headteachers (all)	descriptive statistics; correlations
Perceived impact	R4,5	Surveys (pre/post)	online questionnaires	50 teachers 300 pupils (all)	descriptive statistics
		Interviews	structured interviews	20 parents (high/low fidelity)	deductive coding; thematic analysis

APPENDIX 2: IPE DIMENSIONS

1. **Fidelity/adherence** – the extent to which implementers (e.g., teachers) adhere to the intended treatment model
2. **Adaptation** – the nature and extent of changes made to the intervention
3. **Dosage** – how much of the intended intervention has been delivered and/or received
4. **Quality** – how well different components of an intervention are delivered
5. **Reach** – the rate and scope of participation
6. **Responsiveness** – the degree to which participants engage with the intervention
7. **Programme differentiation** – the extent to which intervention activities can be distinguished from other, existing practice
8. **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)
9. **Perceived impact** – perception from stakeholders on whether the programme has achieved its intended outcomes
10. **Cost** – perceived cost-benefit of the programme from key stakeholders
11. **Context/moderators** – perceived factors that may influence programme delivery or impact (e.g., who did it work for and under what circumstances?)
12. **Mediators** – perceived factors that may have caused the desired outcomes

APPENDIX 3: CAPTURING ‘USUAL PRACTICE’ THROUGH PRE/POST-INTERVENTION SURVEYS

The table below summarises aspects of ‘usual practice’ that should be captured in pre- and post-intervention surveys. The survey should be kept brief to minimise burden for schools. Where possible, collect quantitative data that can be directly compared across the two time-points.

Pre-intervention survey		Post-intervention survey
Control schools	Programme differentiation – what common programmes and practices already take place in schools in this subject area? What does ‘business-as-usual’ look like?	Contamination – compare results from the pre-intervention survey. is there evidence that control schools changed their practice (e.g., increased provision or adoption of new practices)?
Intervention schools	Usual practice – how different is ‘business-as-usual’ provision from the programme?	Unintended consequences – is there evidence of displacement of provision?
All schools	Context – are there any local/national contextual issues that could affect delivery or lead to adaptation of usual practice/programme delivery? (e.g., inspections, change in regulations, covid disruption)?	

Worked example 1 – First Thing Music

This intervention involves getting teachers to deliver 15 minutes face-to-face music sessions at the start of each day to a whole class of Year 1 pupils, over three academic terms. Teachers receive online training and resources, and head teachers are also asked to participate in at least one session per term. The evaluation was designed as a two-arm classroom-level clustered RCT. The table below includes some examples of questions to be addressed through the pre- and post-surveys. Given that the randomisation unit was ‘class’, it is particularly important for post-intervention surveys to capture the degree of contamination and change in music provision.

Pre-intervention survey		Post-intervention survey
Control classes	Programme differentiation – Comparing between control vs intervention group on frequency and intensity of music activities (e.g., How many times a week, for how many minutes?). Questions on the type and timing of music activities may also be relevant.	Contamination – Comparing between pre- and post-survey in the control group (e.g., is there any indication of change in music provision?).
Intervention classes	Usual practice – How much does this differ from 15 mins, face-to-face, at the start of the day?	Unintended consequences – Was there anything that was displaced as a result of implementing this programme? What did pupils do at the start of the day prior to this programme?
All schools	Context – Level of training and resources that teachers received, school buy-in or Head teacher engagement	

Worked example 2 – Catch-up numeracy

This intervention targets pupils in Years 3–5 who are struggling with numeracy. Teaching assistants deliver two 15-minute one-to-one sessions per week outside of the classroom for up to three terms. The evaluation was designed as a two-arm school-level clustered RCT.

To establish programme differentiation, the pre-intervention survey for both intervention and control schools should capture information on the nature of existing maths or numeracy programmes or other related practices that teaching assistants typically deliver to support struggling pupils in these schools for these year groups, as well as specific details about the frequencies and intensities of pupils receiving one-to-one support in maths inside and outside of the classroom. The same questions should be asked at the end of the programme to identify contamination or change in practice as a result of participating in the trial in the control group, and to determine whether any unintended displacement of provision took place in the intervention schools.

Pre-intervention survey		Post-intervention survey
Control schools	Programme differentiation – e.g., Does your school run one-to-one maths/numeracy programmes for struggling pupils in Years 3–5? How frequently do these take place? How long is each session? Are these inside or outside of the classroom?	Contamination – e.g., Does your school run one-to-one maths/numeracy programmes for struggling pupils in Years 3–5? How frequently do these take place? How long is each session? Are these inside or outside of the classroom?
Intervention schools	Usual practice – e.g., Does your school run one-to-one maths/numeracy programmes for struggling pupils in Years 3–5? How frequently do these take place? How long is each session? Are these inside or outside of the classroom?	Unintended consequences – Was there anything that was displaced as a result of implementing this programme?
All schools	Context – School size, single/multi-form entry schools, support from local authorities or Maths hubs	

APPENDIX 4: IPE PROTOCOL CHECKLIST

1. The IPE and IE designs are well integrated and complement one another.
2. The IPE design clearly relates to the logic model and tests any important/unanswered questions related to this.
3. The IPE design is tailored and relevant to the type of study (e.g., pilot, efficacy, effectiveness).
4. There are good linkages between IPE dimensions, research questions and research methods.
5. Research questions and evaluation design support understanding the impact of the intervention on disadvantaged pupils, including, for example, exploring barriers or facilitators that may be specific to disadvantaged pupils.
6. Measures of fidelity, usual practice and any other implementation dimensions that are relevant to the specific intervention are clearly specified.
7. Business-as-usual in control schools and usual practice in intervention schools is assessed prior to intervention delivery and at least once more post-intervention.
8. The IPE should collect data on important contextual factors and describe the approach to collecting this data in the protocol and report.
9. The IPE design and analysis plans are reported transparently including:
 - Sufficient data is planned to be collected from all groups necessary to answer the research questions.
 - Methods of collecting data, timing and sequences of data collection are clearly specified.
 - Sampling approach is clear, justified and appropriate to methods used.
 - Methods of analysis are clearly set out and appropriate to the type of data and to answer the research questions.
 - Approach taken to synthesis, triangulation or integration of data and/or findings from different methods should be clearly specified.
 - Approaches to minimising bias and ensuring rigour are described.

REFERENCES

- Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science*, 12(1), 23–33.
<https://doi.org/10.1007/s11121-010-0186-1>
- Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London: Sage.
- Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, 2(1). <https://doi.org/10.1186/1748-5908-2-40>
- Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8th edn.). London: Routledge.
- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1), 80–92. <https://doi.org/10.1177/160940690600500107>.
- Field, A. (2017). *Discovering statistics using IBM SPSS* (5th edn.). London: Sage.
- Glaser, B.G., & Strauss, A. (1967) Discovery of grounded theory. Strategies for Qualitative Research. Sociology Press.
- Guest, G., Namey, E., & Chen, M. (2020) A simple method to assess and report thematic saturation in qualitative research. PLoS ONE 15(5): e0232076. <https://doi.org/10.1371/journal.pone.0232076>
- Gray, D. E. (2017). *Doing research in the real world* (4th edn.). Los Angeles: Sage.
- Hamilton, L., & Corbett-Whittier, C. (2012). *Using case study in education research*. London: Sage.
- Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016a). *Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature*. London: The Education Endowment Foundation.
- Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016b). *Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook*. London: The Education Endowment Foundation.
https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/IPE_Handbook.pdf
- Jackson, K., & Bazeley, P. (2019). *Qualitative data analysis with NVivo* (3rd edn.). Thousand Oaks, CA: Sage.
- Kincheloe, J. L., & Berry, K. S. (Eds.). (2004). *Rigour and complexity in educational research*. Maidenhead: Open University Press.
- Leech, N. L., & Onwuegbuzie, A. J. (2010). Guidelines for conducting and reporting mixed research in the field of counseling and beyond. *Journal of Counseling & Development*, 88(1), 61–69.
<https://doi.org/10.1002/j.1556-6678.2010.tb00151.x>
- Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 26–46.
<https://doi.org/10.1037/amp0000151>
- Maxwell, B., Stevens, A., Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S., Stiell, B. & Lortie-Forgues, H. (2021). Review: EEF implementation and process evaluation (IPE) quality pilot. <https://educationendowmentfoundation.org.uk/projects-and->

- evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/eef-implementation-and-process-evaluation-ipe-quality-pilot
- McCrudden, M. T., Marchand, G., & Schutz, P. (2019). Mixed methods in educational psychology inquiry. *Contemporary Educational Psychology*, 57, 1–8.
<https://doi.org/10.1016/j.cedpsych.2019.01.008>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2019). *Qualitative data analysis: A methods sourcebook* (4th edn.). London: Sage.
- Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., ... Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *The British Medical Journal*, 350, 1–7. <https://doi.org/10.1136/bmj.h1258>
- Morris, S. P., Edovald, T., Lloyd, C., & Kiss, Z. (2016). The importance of specifying and studying causal mechanisms in school-based randomised controlled trials: Lessons from two studies of cross-age peer tutoring. *Educational Research and Evaluation*, 22(7–8), 422–439.
<https://doi.org/10.1080/13803611.2016.1259113>.
- Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *The British Medical Journal*, 332.
<https://doi.org/10.1136/bmj.332.7538.413>
- O’Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., & Cook, D. A. (2014). Standards for reporting qualitative research: A synthesis of recommendations. *Academic Medicine*, 89(9), 1245–1251. <https://doi.org/10.1097/ACM.0000000000000388>
- O’Cathain, A. (2010). Assessing the quality of mixed methods research: Toward a comprehensive framework. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (2nd edn., pp. 531–555). Thousand Oaks, CA: Sage.
- Pinnock, H., Barwick, M., Carpenter, C. R., Eldridge, S., Grandes, G., Griffiths, C. J., ... Taylor, S. J. C. (2017). Standards for Reporting Implementation Studies (StaRI) Statement. *The British Medical Journal*, 356, 1–9. <https://doi.org/10.1136/bmj.i6795>
- Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Oakland, CA: University of California Press.
- Ritchie, J., Lewis, J., McNaughton Nicholls, C., Ormston, R. (2014). *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. London, UK: SAGE Publication Ltd.
- Saldaña, J. (2015). *The coding manual for qualitative researchers* (3rd edn.). London: Sage.
- Seale, C. (1999). *The quality of qualitative research*. London: Sage.
- Seale, C., & Silverman, D. (1997). Ensuring rigour in qualitative research. *European Journal of Public Health*, 7(4), 379–384. <https://doi.org/10.1093/eurpub/7.4.379>
- Stake, R. E. (1995). *The art of case study research*. London: Sage.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th edn.). Boston, MA: Pearson Education.
- Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6), 349–357. <https://doi.org/10.1093/intqhc/mzm042>
- White, P. (2008). *Developing research questions: A guide for social scientists*. Basingstoke: Palgrave Macmillan.
- Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th edn.). Los Angeles: Sage.