

## Evaluation Handbook

---

Institute for Effective Education  
December 2020



# Contents

<i>Section</i>	<i>Page</i>
<b>Introduction</b>	<b>3</b>
<b>How to use this guide</b>	<b>4</b>
<b>Evidence-informed school improvement</b>	<b>5</b>
1. Really understand the problem or issue you would like to address	6
2. Review existing evidence to better understand your issue and to identify a potential solution	7
3. Write a research question	8
4. Plan a robust impact evaluation	11
a) Sample and recruiting the participants	11
b) Assignment to condition	12
c) Innovation	14
d) Outcome measures	15
e) Process evaluation	18
f) Analyses	19
g) Timeline	21
h) Budget	21
5. Carry out the evaluation as planned	22
6. Analyse your data	22
a) Outcome measures	22
b) Process evaluation analyses	23
7. Describe your findings	24
8. Identify the limitations of your evaluation	24
9. Draw conclusions	25
10. Decide on your next steps	26
a) Implications for practice	26
b) Implications for further evaluation	26
c) Communicate your findings	26
d) Reflections	27
<b>References</b>	<b>28</b>
<b>Sources of evidence</b>	<b>29</b>
<b>Glossary</b>	<b>30</b>
<b>Appendix A</b>	<b>34</b>
<b>Appendix B</b>	<b>38</b>

## Introduction

Teachers often try out innovative teaching practices with their pupils. Sometimes these are approaches that the teachers have developed themselves, either alone or in collaboration with colleagues in their school. Sometimes they are untested approaches created by other teachers. Sometimes they are adaptations of approaches that have been proven to work elsewhere, which the teachers may have altered to meet the needs of their pupils. However the ideas have developed, without robust evaluation it is not possible to tell whether the new approach is more effective than what it has replaced. By carrying out an evaluation, you can better understand:

- The likely impact of the new approach for the pupils in your school.
- What helped teachers implement the approach, as well as potential barriers to implementation.

Evaluating an approach before deciding whether to implement it more widely allows teachers and school leaders to make better-informed decisions about how best to allocate resources. Trialling a new way of working with a small group of pupils, classes or teachers also allows any barriers to implementation to be identified and addressed before the approach is introduced more widely.

There are many different types of research, all of which are well suited to answering different sorts of questions. You can find out more about different types of research and the types of questions they can answer in our [Engaging with Evidence Guide](#). When trying to determine the impact of a particular practice on pupil outcomes, it is best to carry out an *impact evaluation*. An impact evaluation combines an *outcome evaluation*, which compares the progress of pupils who received the innovative teaching approach with similar pupils who did not, with a *process evaluation*, which gathers information about how the approach was delivered and what key stakeholders thought of the innovation. By understanding both aspects of the evaluation, it is possible to understand whether the innovation has evidence of promise for improving outcomes in your context and why this finding may have occurred.

## How to use this guide

The purpose of this document is to support schools in planning and carrying out robust evaluations of practice as part of an evidence-informed approach to school improvement. Before you can carry out an evaluation you need to have a new approach to evaluate, so the guide starts by describing how to identify the issue you wish to address and how to use existing evidence to choose an approach to address this issue. It then describes how to plan and carry out an *impact evaluation* before providing a framework for analysing results and drawing conclusions.

We recommend carrying out each of the 10 steps in our evidence-informed school improvement process in order. We expect most people will want to dip into the guide to help them make decisions about the current stage of their project, so you may choose to read just the content relating to your current step in detail.

We have included a glossary at the end of this guide. This includes terms used in the guide which may not be familiar to all readers. It is not intended as a comprehensive glossary of education research terms, but we hope it will help readers to understand this guide.

You can use the template in Appendix A to help you plan your evaluation. Once you have completed your project, you may like to use our [writing up your innovation evaluation report](#) guidance to help you write up and share your findings.

Please note, in this document the word “school” is used to refer to any education setting, including early years settings, special schools, alternative provision and colleges. “Innovation” is used throughout the document to refer to the new teaching approach or intervention you wish to evaluate.

## Evidence-informed school improvement

Evidence-informed school improvement involves using school data to understand the issue your school is facing, looking at existing evidence about how to address the issue and evaluating the impact of the approach you implement.

Following this 10 step process will help you to apply an evidence-informed school improvement approach to making evidence-informed changes in your school and evaluating the impact these changes have:

1. Really understand the problem or issue you would like to address.
2. Review existing evidence to better understand your issue and to identify a potential solution.
3. Write a research question.
4. Plan a robust impact evaluation.
5. Carry out the evaluation as planned.
6. Analyse your data.
7. Describe your findings.
8. Identify the limitations of your evaluation.
9. Draw conclusions.
10. Decide on your next steps.

It is possible that you have already developed an innovation you would like to evaluate. We would still recommend going through stages 1 and 2 and, if necessary, making improvements to your innovation based on the data and evidence you review.

## 1. Really understand the problem or issue you would like to address

It is likely that you have started this process with an issue in your school or across a group of schools you work with. This issue could relate to any aspect of school life: pupil learning, teacher retention, parental engagement, pupil behaviour, attendance or mental health, or a whole host of other outcomes. Make sure you identify a single, clear problem as it is easier to understand and address a specific problem (for example, Key Stage 4 (KS4) pupils with speech, language and communication needs are receiving a high number of fixed term exclusions) than a broad problem (eg. behaviour in KS4 isn't good enough). Make sure the problem you have identified is something that you can influence.

In order to address the problem it is important to have a good understanding of what the issue is, who it affects and any possible underlying causes. For example, the average reading age of Year 2 pupils at your school may be much lower than expected. In order to address the issue, it is important to know which elements of reading pupils are having difficulties with (phonics? reading fluency? sight word recognition? comprehension? vocabulary? etc.), if any groups of pupils are disproportionately affected and whether any other factors could be contributing to these poor outcomes (for example, poor attendance, poor pupil behaviour or engagement while reading is being taught etc.).

If you do not already have a good understanding of what might be causing or sustaining the problem, it can be helpful to use a hypothesis testing approach. To do this, list all of your *hypotheses* about what might be causing the problem: it can be useful to use external evidence to identify likely hypotheses and to ask colleagues for their theories about the problem. Then gather data to test your key hypotheses. You can use a range of sources of information; this might include teacher assessment, routinely collected school data, *standardised measures* of potential areas of difficulty, staff perceptions, pupil views and parent views. Once you have gathered all of your data, you should have a better understanding of what is, and what is not, contributing to your problem. If the data you have gathered suggests that multiple factors are contributing to the problem, we suggest selecting one of these factors which is relatively **easy to change and which you are able to influence**.

## 2. Review existing evidence to better understand your issue and to identify a potential solution

You should now have an idea of the potential cause of your issue. Read about the problem and the potential cause to establish what is already known. Many of the problems frequently encountered in schools have been extensively researched, and this research can help teachers and school leaders decide how to address the issue and improve outcomes.

A number of organisations provide accessible summaries of research which will help you to understand the evidence and to identify possible solutions. Links to some of these organisations are listed in the sources of evidence section below. You can use the IEE's [Engaging with Evidence Guide](#) to help you understand different types of evidence, the questions each type of evidence can answer and how to gauge the quality and *reliability* of the evidence.

Use your review of the external evidence to decide on a potential solution to address your problem. Consider what existing research evidence makes you believe that this innovation will be effective. Create a *theory of change*, explaining the mechanisms by which you think it will work.

### Box 1: Theory of change

A theory of change gives a clear description of the process by which you think your chosen solution could lead to the desired change in your context. Theories of change show a hypothesised path between your solution and your desired outcome, so creating a theory of change makes you think through exactly how and why you believe your approach will address the problem you have identified. Creating a theory of change takes time and mental effort, but without going through this process you risk selecting a fashionable intervention or a solution which you are biased towards, but which is unlikely to change the outcomes you want to improve.

There is no correct structure for a theory of change; the best way to represent the sequence of changes you hope will happen may be a flow chart, a concept map, a web, a table, or it could take a whole myriad of other forms. There are lots of resources available about how to make a theory of change. You might like to start by looking at resources produced by [Community Tool Box](#) or [The Center for Theory of Change](#).

Once you have selected an innovation, you should consider how you will maximise the chance of it being implemented correctly and consistently. The EEF's [Putting evidence to work: a school's guide to implementation](#) may be helpful in planning for implementation.

### 3. Write a research question

Write the specific research question that you would like the evaluation to address. Having a well-defined research question is vital in planning and carrying out a coherent evaluation. The research question should include the practice that will be evaluated, for how long it will be delivered, the outcomes to be measured and the people who will experience the new practice. You will have identified the practice from your review of existing evidence (step 2 above), while your outcome is likely to be the issue you pinpointed in step 1. A research question for an impact evaluation should take the form:

What impact does [what practice?] delivered [over how long?] have on [what outcome?] for [whom]?

Each element of the research question should be clearly defined. For example:

What impact does Toddler Talk delivered for 10 minutes a day, four days a week for four months have on the expressive vocabulary of two year olds whose home language is not English?

You may wish to look at the impact of your innovation on more than one outcome. In this case you should write separate research questions for each outcome. However having too many research questions can make an evaluation unwieldy. Between one and three research questions is generally manageable; we would not recommend trying to answer more than three research questions in a single evaluation.

You may also like to make a specific prediction about what you expect to happen (a *hypothesis*). This should include the comparisons being made (eg, I predict that after four months, the average expressive vocabulary of 2 year olds who participate in Toddler Talk will be larger than the average expressive vocabulary of 2 year olds in the control condition) and should be based on what existing research suggests would be a likely outcome.

## Box 2: When research questions go bad

Writing a good research question is harder than it sounds. It can be helpful to ask someone else to read your question and to challenge you about whether all of the elements are really specific. There are several common pitfalls in writing research questions for impact evaluations, all of which can be resolved with time, thought and some help from a critical friend!

### The outcome is poorly defined

What impact does weekly mentoring with a Year 12 pupil for six months have on the attainment of high prior-attaining Year 9 pupil premium pupils?

In this example, 'attainment' is not a well-defined outcome. Keep asking yourself "what do I mean by [outcome]?" until you can't provide any more detail. For example:

- "What do I mean by attainment?" "End of year attainment."
- "What do I mean by end of year attainment?" "End of year exam scores."
- "What do I mean by end of year exam scores?" "End of year exam scores for core subjects."
- "What do I mean by end of year exam scores for core subjects?" "Average score in English, maths and science end of year exams."

The innovation, duration and participants could also be poorly defined, although this is less common. You can use the question stem "what do I mean by ...?" to help specify these elements as well.

### More than one outcome is included

What impact does the whole school Shining Stars reward system, delivered over one academic year, have on instances of persistent low level disruption, internal exclusions and fixed term exclusions of pupils in Key Stage 2?

In this example, three outcomes are included in the same question. It would be easier to plan your evaluation, and to address your research questions when you come to write your conclusions, if this was split into three research questions.

### The question is too broad to be answerable

Can a better use of TAs for two terms help our pupils access learning, make better progress and reach their potential?

Sometimes a question is so broad that it would be impossible to answer. In this example the innovation is unclear: there are lots of ways schools can improve the deployment of TAs, and this question doesn't tell us which will be used. The outcomes, as currently described, are poorly defined and one of the outcomes is impossible to measure (there is no objective way of determining whether a pupil has reached their potential).

Your research question needs to name the specific outcome you would like the innovation to improve and for whom you would like these outcomes to improve. If you aren't sure about the outcome or people of interest, go back to step 1; if you develop a new understanding of the problem you should also repeat step 2. You need to state the specific innovation you plan to evaluate and the amount of time you will deliver the innovation for. If you aren't sure about your innovation or the best duration of delivery, repeat step 2.

### The question makes assumptions about the impact of the innovation

What impact does the improved understanding of A Tale of Two Cities from the use of a knowledge organiser for one unit of work have on performance in exam questions about A Tale of Two Cities for Year 13 English Literature pupils?

This question assumes that knowledge organisers will improve understanding of A Tale of Two Cities, but we don't yet know whether this is true. It is important to write a neutral question, and to be open to the possibility that your evaluation may find that the innovation didn't improve outcomes as you'd hoped.

## 4. Plan a robust impact evaluation

You now need to plan an evaluation that will answer your research question. It is important to make decisions about the sample, recruitment, design, measures, innovation, process evaluation and data analysis before starting the evaluation so the research is scientifically valid; making decisions as you go along is likely to lead to biased results. You can use our evaluation planning template (see Appendix A) to structure your planning. There is no set order in which you need to address the eight subheadings in this section: it is likely you will return to and revise sections as your plan develops.

Involve other stakeholders in your decision making, and write your plans clearly so you and everyone else involved understands what will happen. Create copies of research plans to be agreed and signed by everyone involved. It is also important to ensure you have buy-in from the leadership teams of all schools involved in the evaluation.

It is good practice to publish your plan before you start your evaluation. Having a clear description of what you planned to do will allow you to identify any changes or compromises you had to make as you went along. This will ensure that any changes you make to your plan are transparent, which will help you and others to draw informed conclusions about your findings. You could publish your plan on your school website or the website of any partner organisations you are carrying out the evaluation with.

### a) Sample and recruiting the participants

Decide who will participate in your evaluation. The *sample* should be representative of the group referred to in your research question. Decide on any required characteristics of *participants* in your sample (eg, year group or age, current attainment, home language, gender etc.); these required characteristics are called the *inclusion criteria*.

#### Checklist:

- Have you identified the inclusion criteria for your participants?
- Have you decided how you will recruit participants?
- Have you decided what consent will be needed and how you will gain consent?

Decide how you will recruit participants to your study. If you plan to recruit pupils beyond your school, decide how and when you will approach other teachers/schools to share the opportunity to participate. Plan how you will ensure you have buy-in from the school(s) leadership team(s), and when you will conduct meetings with staff from the schools involved to be sure that everyone is clear about what is going to happen.

Decide whether you need consent for pupils to participate, who would give this consent and what sort of consent should be sought. For pupils under the age of 16, parental

consent should be obtained before the evaluation begins. Pupil consent should also be sought where appropriate, and definitely when participants are 16 or older. Participant consent should also always be obtained before conducting an interview, regardless of age. If consent would not be requested for pupils to participate in the innovation as part of normal classroom practice, it may only be necessary to ask for consent for pupil data to be included in data analysis (ie, all eligible pupils take part in the innovation, but parents can decide whether their child's data will

be included in the data analysis). If the innovation is a change in practice that a teacher might routinely implement and is not very obtrusive, opt-out consent (ie, parents respond if they do *not* want their child to participate in the evaluation) rather than opt-in consent can increase participation rates. Whatever form of consent you choose, it is your responsibility to ensure that procedures and assessments are not potentially harmful; the [British Educational Research Association ethical guidelines](#) are a useful guide to ethics in education research.

#### *b) Assignment to condition*

Pupils usually make progress over time; you want to find out whether your innovation leads to improved progress compared to current practice. In order to establish the impact of your innovation, it is important to compare outcomes for pupils who experience the new practice with similar pupils who do not. This protects against the risk of typical progress (eg, pupils' understanding of algebra typically increases during Year 9) or general patterns (eg, a year-on-year national, regional or school-wide improvement in understanding of algebra among Year 9 pupils) being misinterpreted as an effect caused by the innovation (eg, concluding that the introduction of Active Algebra improved the algebraic understanding of Year 9 pupils). Pupils' non-academic competencies also naturally change over time (for example, school engagement typically declines with age) so allowing for comparison to pupils who did not experience the innovation is also important for non-academic measures.

*Condition* refers to the group participants are in: they may be in the *intervention group* (sometimes called the experimental group) or the *control group* (sometimes called the comparison group). The intervention group will experience the innovation you are evaluating, while the control group will experience whatever their teacher or school would usually deliver. The intervention and control group should be as similar as possible to each other at the beginning of your study and neither group should have previously experienced the innovation. Pupils in both groups should meet the inclusion criteria.

*Assignment to condition* refers to the way in which participants are allocated to different conditions. You should assign participants to condition at the level at which the innovation is delivered. For example, if it is delivered to individuals (eg, a one-to-one tutoring programme), then you can assign individuals to control and intervention groups. If the innovation is delivered to whole classes (eg, a new maths curriculum) then it is better to assign classes to control and intervention groups. If the innovation is a whole-school approach (eg, a school-wide healthy eating programme), then it is best to assign schools.

#### **Checklist:**

- Have you decided how you will allocate pupils to the intervention and control groups?
- Have you decided how you will check that groups are well-matched before the innovation starts?

*Random assignment* of participants to the control or intervention group is preferable to somebody deciding who will be in each group as it removes a possible source of bias from the evaluation. *Randomised control trials* randomly assign participants to groups. You can randomly assign participants to intervention and control groups by using a random number generator (there is one in Excel and many are available online), by flipping a coin or by pulling names out of a hat. Once you have

randomly assigned participants to groups you need to check whether the intervention and control groups are *well-matched* on characteristics that are important for your *population* and innovation. For example, if your research questions asked about the impact of a mental health programme on the self-reported happiness of a) all Year 11 pupils and b) Year 11 pupils in receipt of free school meals, you would need to check that Year 11 pupils allocated to the intervention and control groups had similar levels of happiness before the innovation started and that the groups contained similar proportions of pupils in receipt of free school meals. If the groups are not well-matched, you can re-randomise the groups until you are satisfied they are well-matched.

It is sometimes not possible to randomise participants to groups for pragmatic reasons (for example, timetabling constraints may mean it is only feasible for one of the participating classes to carry out the innovation). Experiments which do not have random allocation to groups are called *quasi-experiments*. If allocation is non-random it is particularly important that the groups are very well-matched before the evaluation starts in relation to the outcome(s) you are measuring and any relevant demographic characteristics.

### **Box 3: Control groups.**

Some people object to using control groups as they feel it is not fair for the control group to not receive the innovation. However, at the moment you don't know whether the innovation works better than what you're currently doing – that's why you're going to carry out an evaluation. It is only unethical to withhold the innovation if you already know that it has a positive impact. 80% of trials carried out by the Education Endowment Foundation show no evidence of effectiveness, so genuinely "better" interventions are actually quite rare.

If you feel uncomfortable about using control groups, remember:

- The control group will be a *business as usual control group*. This means they will be taught using the approach currently used at your school. It is important to teach the control group the content delivered by the innovation – not only would not teaching them be unethical, it also wouldn't help you to understand whether the innovation is more effective than current practices.
- If you find that the innovation is effective you can use the innovation with the control group after you have completed the evaluation (you could see the control group as a *waiting list control group*). This means the control group gets the best of both worlds: if the innovation is more effective than current practice they will be able to benefit from it; and if current practice is more effective they (and future pupils) will not have wasted their time with the less effective innovation.

### c) Innovation

Describe the innovation clearly so that people who are involved in implementation understand exactly what they are expected to do, and external people who read about the evaluation understand what the innovation involves. Clearly outline:

- What the innovation aims to teach (eg, subject(s), topics, skills, etc.).
- What participants will experience, including:
  - the length of time over which the innovation will be delivered
  - the number, duration, frequency and timing of innovation delivery
  - the types of activities participants will engage in
  - any expectations beyond the direct delivery of the innovation (eg, homework).
- The level at which the innovation will be delivered (eg, to individuals, groups, classes or whole schools).
- Who will deliver the innovation.
- Any training that will be given to those delivering the innovation, and any support which will be provided while the innovation is being delivered. Consider what will happen if members of staff leave during the innovation and how new staff would be trained.
- Any resources available to support the delivery of the innovation, such as manuals or lesson plans (having clear directions from common resources can help to increase the consistency with which the innovation is delivered).
- Any other information which is pertinent to your innovation.

You should also outline expectations for control group activity: this may be 'business as usual' (ie, whatever the teacher or school would usually provide), an alternate intervention or a variation of the innovation. Consider how you will avoid *contamination of the control group*, which occurs when the control group experiences some elements of the innovation.

As the person running the evaluation, you need to carefully consider your own involvement in delivering the innovation. This is especially important if you also developed the approach. Ideally you shouldn't teach either the intervention or control group: if someone else teaches the intervention group you will get a better idea of how easily other people can understand and deliver the innovation, while not teaching the control group reduces the risk of contamination (i.e. accidentally using elements of the innovation with the control group).

#### Checklist:

- Have you clearly described the innovation?
- Have you decided how those delivering the innovation will be trained, and what follow on support they will be given?
- Have you identified the resources that need to be collated or created, and when these will be bought, gathered or made?
- Do you know what the control group will do? How will you make sure no-one working with the control group learns how to deliver the approach or gains access to innovation resources?

#### d) Outcome measures

It is important to carry out objective *outcome measures* if you wish to establish the impact of your innovation. People tend to believe that things in which they have invested time and effort have a positive impact, even when the data shows this is not the case, so relying on participant views is unlikely to give a reliable picture of the impact of your innovation. You should therefore use objective measures of the outcome named in your research question(s). If possible these should be externally written (i.e. the test should not be written by anyone involved in developing the innovation or running the evaluation) *standardised measures* as these measures are more reliable and valid and less biased than measures made for the purpose of the evaluation. However, we know that the lack of freely available standardised measures is a problem for teacher-led evaluations, meaning it may be necessary to create your own outcome measure to use in the evaluation.

##### Checklist:

- Have you selected an outcome measure?
- Have you identified dates for the pre- and post-test and agreed who will administer the tests?
- Have you agreed who will mark pre- and post-tests and decided how you will increase the reliability of marking?

You should carry out the same measures with your intervention and control groups to compare outcomes for pupils who experienced the innovation with those who did not. You need to administer an outcome measure before the innovation starts (a *pre-test*) to establish pupil starting points and to determine how similar the control and intervention groups are. You will also need to administer an outcome measure after the innovation has ended (a *post-test*) to establish how much progress pupils in the intervention and control groups have made.

You may also choose to test participants a number of weeks or months after the innovation has finished (a *delayed post-test*) to establish whether any impact is sustained over time.

Both pre- and post-tests should not be too onerous for pupils or teachers; if possible use a common measure that pupils would carry out anyway (eg, typical end-of-year assessments). Measures must be fair to the control group; this means you should not assess content only taught to the intervention group or create tests from innovation resources or materials which will be familiar to the intervention group but not the control group. Where possible, pre-tests, post-tests and delayed post-tests should be comparable (eg, *equivalent versions of standardised measures*). The Education Endowment Foundation's [DIY Evaluation Guide](#) has information about the different types of outcome measures (national assessments, standardised tests and self-made measures). The EEF's [SPECTRUM database](#) lists measures of non-academic skills.

You will need to make decisions about:

- Which measures you will use.
- When measures will be administered.
- How measures will be administered and by whom (measures should be carried out in the same conditions for all participants and, if possible, they should be administered by someone who is not otherwise involved in the evaluation).
- How tests will be marked and by whom (if possible, the marker should not be able to identify individual pupils or which condition each pupil is in).
- How you will ensure tests are marked reliably (e.g. use of a mark scheme, moderation).

If the outcome in your research question relates to pupil or teacher behaviours, it may be appropriate to carry out a *structured observation* as your pre- and post-test measure. This involves an observer recording when a behaviour is carried out by a pre-defined group of people (this may be a single person, a selected group of participants or all participants) over a set amount of time. It is really important to very clearly define the target behaviour and to train the observer to reliably identify the behaviour; using an *observation schedule* can help with this.

#### Box 4: Structured observations

Structured observations are a systematic way of recording behaviour. They can be a useful way to check whether the innovation is being delivered as intended, or they can be used as an outcome measure if you're interested in changes in behaviour.

Before carrying out a structured observation, you will need to define the behaviour you are interested in very clearly. If you are using the observation as an outcome measure you should write down what does and what does not count as the behaviour. And if you are carrying out an observation as part of the process evaluation you should list the features of the innovation that you would expect to be present. These descriptions should be detailed and specific enough that someone else could use the list to identify the behaviour or innovation characteristics in the same way you would. You can make an *observation schedule* by turning the list into a form to be filled out during the observation.

Example 1 is an observation schedule that could be used as part of the process evaluation:

Example 1: fidelity check for retrieval practice lesson starters

Date: \_\_\_\_\_ Time: \_\_\_\_\_ Class: \_\_\_\_\_ Teacher: \_\_\_\_\_

Feature of retrieval practice lesson starters	Tick if observed
Pupils start retrieval practice activity as soon as they enter the classroom.	
Pupils carry out an activity which requires retrieving information from their memories. This might include, but is not limited to, using flashcards, using retrieval practice software, testing a peer, creating mind maps or recreating revision notes.	
The activity lasts between 5 and 10 minutes.	

The following should not be present in this innovation. Circle any that are observed:

- Pupils re-reading notes.
- Pupils using textbooks or previous written work to make revision materials.
- Pupils using textbooks, previous written work, revision materials or online resources to find answers to the questions.
- The activity lasts less than 5 minutes.
- The activity lasts longer than 10 minutes.

Example 2 could be used as an outcome measure for a discrete behaviour:

Example 2: behavioural observation of pupils raising their hand to answer a question

Date: \_\_\_\_\_ Time: \_\_\_\_\_ Class: \_\_\_\_\_ Teacher: \_\_\_\_\_

Pupils being observed: \_\_\_\_\_

What counts as a pupil raising their hand:

- Hand above the level of their chin
- Hand not in contact with head
- Hand up when the teacher has just asked a question

What doesn't count as a pupil raising their hand:

- Hand below the level of their chin
- Hand touching face or hair
- Hand raised at a time the teacher has not asked a question
- If the pupil raises and lowers their hand multiple times in response to a single question, count this only once.

	Tally (one tally each time each pupil raises their hand in response to a question from the teacher)
Pupil A	
Pupil B	

Counting how many times a behaviour happens is useful for discrete behaviours. But if you are interested in a behaviour that can last variable amounts of time, breaking the observation down into smaller chunks of time and recording the behaviours that occur in each time period can be a more realistic measure. You can then report the proportion of the observation in which the behaviour was observed. For example:

Example 3: behavioural observation of on and off-task behaviour

Date: \_\_\_\_\_ Time: \_\_\_\_\_ Class: \_\_\_\_\_ Teacher: \_\_\_\_\_

Pupil being observed: \_\_\_\_\_

What counts as on-task behaviour

- Looking at the teacher when teacher is talking to the class
- Participating in class discussion
- Working on written work
- Carrying out practical work as instructed

What counts as off-task behaviour

- Moving around the classroom when this is not part of the lesson
- Making comments to the teacher or class about topics unrelated to the lesson
- Talking to peers about topics unrelated to the lesson
- Looking around the classroom or out of the window for more than 30 seconds
- Staring in one direction (not at the teacher or person speaking) for more than 30 seconds

Minute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
On task															
Off task															

It may not be feasible to observe all pupils in a class and record their behaviour accurately, in which case you should select a sample of the class to observe. If your research question refers to pupils with specific characteristics, these are the pupils who should be observed. If you are interested in the behaviour of all pupils, or too many pupils meet the inclusion criteria to be able to record their behaviour accurately, select a representative sample from the class who will be observed. Make sure that pupils being observed in the control class are selected using the same criteria as those being observed in the intervention class. If you do wish to observe all pupils in the class you could film the lesson and complete the observation schedule later.

Whoever you observe and however you manage the observation, if you are using the observation as an outcome measure the same pupils should be observed in the pre- and post-test.

#### e) Process evaluation

A **process evaluation** provides information about how the innovation was delivered. It is useful to carry out a process evaluation alongside the outcome evaluation so that you can understand if the innovation was delivered as intended (**implementation fidelity**), and to obtain teacher and pupil views of the innovation and their perceptions of things that went well or could be improved (participant views). This will help you to understand the results of your outcome evaluation.

#### Fidelity

To determine if the innovation is delivered as intended you should consider conducting observations of the innovation being implemented. **Structured observations** are a useful process evaluation tool as they are easier to summarise and analyse, and more likely to be objective, than unstructured observations. Identify the key features of your innovation and create an **observation schedule** to use during unobtrusive observations; there is an example of an observation schedule for a fidelity check in Box 4. To check that the control group is not receiving elements of the innovation you could conduct observations using your observation schedule with the control group as well. Usually observations are conducted after participating staff and pupils have had some experience with the innovation.

#### Checklist:

- Have you decided how you will check if the innovation is being delivered as expected?
- Will you also check that the control group are not receiving any elements of the innovation?
- If you plan to carry out observations, have you decided who will carry them out and when? Have you created an observation schedule?
- If you plan to use other fidelity checks, have you decided who will carry them out, when and what criteria they will use?

You could also check planning or pupil books to see whether the innovation has been delivered. Sometimes those delivering the innovation are asked to keep an implementation diary or to complete a checklist for each session delivered; before deciding to use this approach, consider how much additional time this will take those delivering the innovation.

### Participant views

It is useful to ask those who delivered or experienced the innovation what they thought about the approach. As well as asking everyone involved for their opinions of the innovation, it can be helpful to ask people who delivered the innovation about any challenges they experienced in delivering the approach and any factors which supported successful implementation. Their responses can help you think about how you could make the innovation easier to deliver or improve training and support if you choose to use the approach again.

You should *survey* key stakeholders to find out about their views of the innovation: *interviews*, *questionnaires* and *focus groups* are common survey methods. Think about the purpose of your survey and allow yourself enough time to write questions which will meet your aims (you can read more about writing a survey in Appendix B). It can be helpful to include some closed questions that are quick and easy to analyse, as well as open questions which take longer to analyse but will add depth to your understanding of participant views. Questionnaires allow you to gather and analyse information from large numbers of people relatively quickly and easily, while interviews and focus groups can be a flexible way to gather in-depth feedback from a smaller number of people. Be aware that focus groups are very difficult to run well; when run by someone without expertise, it is common for group dynamics or strong personalities to influence responses.

Questionnaires and interviews are usually conducted towards the end of the evaluation so that you avoid novelty effects (people liking or disliking something purely because it is new) or learning curve effects (participants developing their understanding of the innovation over time). Consider asking someone who wasn't involved in developing the innovation or running the evaluation to conduct face-to-face interviews; participants may give more honest responses to someone who isn't involved in the project.

### f) Analyses

#### Outcomes

Describe how you will analyse the *quantitative* (numerical) data from your outcome measures. Planning for data analysis before starting the evaluation may seem like planning overkill, but it is an important step if you want to ensure your analysis is reliable, robust and unbiased. You can read more about the dangers of unplanned analyses in this [blog post](#).

You will probably want to calculate the average score on the pre-test, the post-test and, if you plan to carry one out, the delayed post-test for both the intervention and control groups; you

#### Checklist:

- Have you decided whose opinions you will gather?
- Have you decided how you will gather opinions (interviews, questionnaires, focus group or a combination)?
- Have you written your survey questions?
- Have you decided how questionnaires will be administered and/or who will conduct interviews and when these will take place?

may also like to calculate the average progress made by both groups. Decide which sort of average you will calculate for all of the measures you will use:

- The *mean* is the average that is found by adding up the numbers and dividing by the number of numbers. This can be used to find the average of data where the distance between scores is constant. You can use the mean for test scores, time taken to complete a test or activity or the number of times a behaviour is observed.
- The *median* is the value above which half of the figures lie. This can be used to find the average of data where the distance between scores is not constant, such as responses to rating scales. Many self-report measures of social and emotional competencies and experiences use rating scales, so the median is the appropriate average to calculate for these measures.

You may also like to calculate an *effect size*. Effect sizes are a measure of the difference in outcomes of the intervention group and control group. You could calculate an effect size using post-test scores (which would tell you whether the intervention group or the control group scored higher on the post-test) or for progress (which would tell you whether the intervention group or control group made more progress); you should state whether you will use post-test data or progress to calculate the effect size. If you are using both an immediate and a delayed post-test you should also state what comparison(s) you will use to calculate an effect size.

#### Checklist:

- Have you decided what averages you will calculate?
- Have you decided what averages you will compare? Will you calculate an effect size for these comparisons?
- If you plan to analyse data for specific subgroups of pupils, have you specified how these groups will be defined and what analyses you will carry out?

You do not necessarily need to conduct fancy high-level statistics, unless of course that is your cup of tea. As your evaluation is likely to be small scale, there is a strong likelihood that *tests of statistical significance* will show that the results were due to chance. The smaller the number of participants the more likely that tests of statistical significance will find that results were due to chance, rather than as a result of the innovation.

If you are going to carry out analyses of impact for any subgroups of pupils (eg, pupils with specific prior attainment, gender, indicators of financial deprivation, etc.) decide what analyses you will carry out on which subgroups. It is important to decide this in advance because the more analyses you carry out, the more likely you are to stumble across results that appear to suggest a strong impact, but which have probably occurred by chance (reporting only the positive results from numerous analyses is known as “cherry picking”).

For each subgroup, state:

- how the group will be defined (eg, if you plan to analyse results for high prior attaining pupils, what measure of attainment will you use to identify these pupils? What cut-off will you use to define high attainment on this measure?), and

- the data analysis you plan to carry out: what measures will you analyse for each subgroup? Will you calculate the mean or median score? Will you calculate an effect size? If so, will you compare post-test scores or progress?

#### Checklist:

- Have you decided how you will analyse quantitative and qualitative data from the process evaluation?

#### Process evaluation

Describe how you will analyse data gathered for the process analysis. When you have *quantitative data* (eg, data from observation schedules or responses to closed questions or scaling) you need to decide whether it is most appropriate to calculate the *mean*, *median* or *mode* of the responses (the mode is the value that most

commonly occurs in a group of scores. This is used as an average for non-numerical, categorical data, such as the element of the innovation that pupils reported enjoying the most).

Qualitative responses to simple open questions can often be counted. For example, if analysing responses to the question 'which of the books did you enjoy the most?' you could count each time each book is named; this will tell you which books were most and least popular. For more complex *qualitative data* such as answers to more complex questions (e.g. why do you feel pupils chose not to attend the after school sessions?) you could conduct a *thematic analysis*. During a thematic analysis you examine the responses given by participants to identify themes that point to similar conclusions; it can also be interesting to look for contradictory responses from different participants.

#### g) Timeline

Create a detailed timeline or *Gantt chart* for your project plan, including milestones and who is responsible for achieving all actions. Include this in your agreements with leadership team(s), teachers and others involved in the evaluation.

#### h) Budget

It is sensible to create a budget during the planning stage. This will help ensure that you have sufficient funding to complete your evaluation and allow you to arrange access to funding at appropriate times during the evaluation.

## 5. Carry out the evaluation as planned

Carry out your evaluation as described in your plan. Remember to complete elements of the process evaluation, such as carrying out lesson observations, as well as administering outcomes measures and delivering the innovation.

Have clear, consistent and frequent communication with everyone involved throughout the evaluation, and address any issues as soon as they arise. You should also closely monitor your budget during the evaluation.

While you should make every effort to follow your plan, things might not happen as you intended for all sorts of reasons. That's OK! Just remember to make a note of anything which deviated from your original plan and explain clearly what changed and why when you share your findings.

## 6. Analyse your data

Once you have collected your data, you need to analyse your results. This includes analysing data from outcome measures and information gathered as part of the process analysis.

### *a) Outcome measures*

Carry out the data analysis described in the analysis section of your plan. To maintain confidentiality, allocate each pupil an ID number to replace their name. Keep a record of the pupils' names and ID numbers securely in a separate location.

It is possible that you may not have a complete data set for all participants who completed the innovation (for example, some participants may have been absent on the day of one of the tests). For each analysis you carry out, only include participants for whom you have a complete set of data. However, you only need to remove the participant from the analysis for measures where they are missing data. For example, you may have evaluated the impact of delivering additional PE lessons on pupils' weight and time taken to run 800m. If a pupil completed both measures at the pre-test and their weight was measured at the post-test but you did not time them running 800m at the post test, you should include this pupil's data in data analysis for weight but not for time to run 800m. You should report how many participants were included in each calculation; this is usually reported as *n* in results tables.

Alternatively, you may have a complete set of data for a participant who decided not to complete the innovation. This might happen if the innovation is voluntary or takes place outside normal school hours. In this case, you should include data in the analysis for all participants based on the condition to which they were initially allocated. This is because participants who drop out of the innovation are likely to have different characteristics to those who complete the innovation. If your intervention and control groups were equivalent before the innovation started it is likely there are also some participants with these characteristics in the control group, but they will not be identified as they cannot drop out of the innovation. It is therefore a fairer comparison to carry out the analysis based on *intention to treat*.

If you planned to calculate an effect size, the formula you need to use is:

$$\text{effect size (ES)} = \frac{(\text{mean of the intervention group} - \text{mean of the control group})}{\text{standard deviation (SD) of the whole sample}}$$

If you are calculating an effect size based on post-test scores, you should use the mean post-test score for the intervention group, the mean post-test score for the control group and the standard deviation of post-test scores for the whole sample. If you are calculating an effect size based on progress, you should use the mean progress (ie. the post-test score minus the pre-test score) made by the intervention group, the mean progress made by the control group and the standard deviation of progress for the whole sample.

#### Box 5: Steps for computing an effect size in Excel

1. Enter outcome data for each participant into a spreadsheet, with data from intervention and control participants in different columns. To make it simple and clear, you only need data from the measure for which you are calculating an effect size in this sheet. Delete the rows for any participants who do not have a complete set of data for this measure.
2. Compute the mean for the intervention group (in Excel this is called 'AVERAGE' and can be found under 'AutoSum' in 'Formulas').
3. Compute the mean for the control group.
4. Compute the **standard deviation** for the whole sample (in Excel select 'STDEV.S' from 'Statistical' under 'More Options' in 'Formulas', highlight the outcome data for all intervention and control group pupils and click OK).
5. Subtract the average of the control group from the average of the intervention group, and divide that number by the standard deviation of the whole group. That figure is the effect size.

You may have decided in the planning stage to examine the results of pre-identified subgroups (eg, by gender, attainment, eligibility for free school meals etc.) to determine if the innovation has a greater impact with certain groups of pupils. You should carry out the data analysis for these subgroups as described in the analysis section of your plan.

#### b) Process evaluation analyses

Analyse the data gathered from your process evaluation as described in your plan.

## 7. Describe your findings

State your results: you do not need to explain them at this stage, just describe them. You should consider information from both the outcome and process analysis when describing your results. Make sure that you include all the information you have as clearly as you can; tables, graphs or other infographics are often a concise way to report outcome data.

## 8. Identify the limitations of your evaluation

All evaluations have limitations, and it is important to consider the constraints and methodological limitations of your evaluation when drawing conclusions. There are many factors which could mean that something other than the innovation could have caused the effects you measured (these are called threats to *internal validity*, and you can read more about them [here](#)), such as:

- changes other than the innovation in some or all of the participating schools/classes
- lack of similarity between the intervention and control groups before the innovation started
- lack of similarity between the teaching delivered to intervention and control groups beyond the innovation
- drop-out from one or both of the groups influencing findings
- the control group receiving elements of the innovation
- the control group feeling different or being perceived differently because they are not receiving the innovation, leading to changes in control group behaviour or treatment.

Think about any potential sources of bias in your evaluation. Bias is especially likely to be an issue if the person who developed the innovation was strongly involved in delivering or evaluating the approach. If you believe any element of the evaluation may have introduced bias, you can have less confidence in the results obtained.

You should also consider the settings to which findings can be generalised (the *external validity* of the findings); based on the scale of your study and the characteristics of the schools involved, think about what sort of schools might find similar results.

## 9. Draw conclusions

Use your results alongside your understanding of the limitations of your research to draw conclusions. These conclusions should answer your research question(s). Consider whether your findings are similar to or different from the existing evidence base and why this might be the case.

### Box 6: Interpreting outcome data and drawing conclusions

*Effect sizes* are a way of comparing outcomes for the intervention and control groups (the outcome could be post-test scores or progress). By considering the difference in outcome for the two groups in relation to the standard deviation of the whole group, effect sizes give an indication of the scale of the difference. A positive effect size shows that the intervention group had better outcomes than the control group, and a negative effect size shows that the control group had better outcomes than the intervention group. The larger the effect size, the greater the difference in outcomes for the two groups.

There are no general rules for what constitutes a 'good' effect size. While some categories exist, these were not designed to be used in isolation. Effect sizes are influenced by characteristics of the evaluation including homogeneity of the sample, the outcome being measured, the reliability of the test, and many others. And the practical implications of effect sizes will depend on characteristics of the innovation such as cost and potential scalability.

You should use all of the data you have gathered from the outcome and process evaluation when drawing conclusions. Look at the effect size, mean pre- and post-test scores and mean progress to understand the patterns of progress made by both groups over the course of the evaluation: results tables and graphs can make it easier to see these patterns. Consider this outcome data alongside the limitations of the study, your understanding of the context in which the evaluation took place and participant views gathered as part of the process evaluation to draw conclusions about the potential evidence of promise of your innovation in your setting.

If you would like to know more about what effect sizes mean, [this paper \(Kraft, 2019\)](#) on interpreting the effect sizes of education interventions is a good starting point.

## 10. Decide on your next steps

Consider how you will use your findings and think about the things other people (including other teachers and researchers) should know about your findings. Decide how you will share your findings with others. This is also a good time to reflect on what you have learnt from carrying out the evaluation.

### *a) Implications for practice*

Decide:

- what (if anything) you will do differently in your classroom as a result of these findings
- what (if anything) you will do differently in your school as a result of these findings
- recommendations you would make to other teachers or school leaders on the basis of these findings.

Results from either the outcome or process evaluation may lead you to conclude that you should make changes to the approach before delivering it again. If you plan to make major changes, you will need to carry out an evaluation of the updated innovation before you can be confident that it will work effectively.

### *b) Implications for further evaluation*

Lots of questions that were beyond the scope of your research question are bound to have occurred to you while you were carrying out the evaluation and analysing your results. These might relate to comments from other stakeholders, your observations of how participants engaged with the innovation or from patterns you noticed in the data. They may be questions about how the approach would work in different contexts, subjects or age groups. These may be things you plan to look into in another research project, or they may be interesting questions that you don't currently have the capacity to investigate yourself. Consider:

- any further research you would like to carry out
- recommendations for researchers working in the field.

### *c) Communicate your findings*

You should share your findings so that others can benefit from your evaluation. It is important to share your findings, even if results are negative or if you did not complete all elements of the evaluation. Sharing negative results will reduce the chance of people trying things which are unlikely to work or potentially wasting their time evaluating a similar approach. If the process evaluation showed that teachers stopped delivering an innovation because they didn't like it, people who are interested in the approach might consider how the innovation, or the implementation plan, could be improved to increase the chance of delivery. And if you stopped the evaluation because of unforeseen negative outcomes of the innovation on participants, it is particularly important that you share your findings to prevent others from using a similar approach with similarly negative consequences.

We would recommend writing a report and publishing it. You may be able to publish the report on your school website, or on the website of partners who collaborated in the evaluation. Teaching publications, such as the TES, are very interested in the results of experiments run by teachers and so you may like to approach them to see if they would like to publish your findings.

Our suggestions for writing up your evaluation can be found in our guidance [writing up your evaluation report](#). You could also share your findings by speaking at conferences and events or presenting at local network meetings.

#### *d) Reflections*

After you have completed your evaluation, it is useful to reflect on the process. You should consider what went well, what you have learned and what you would do differently in future evaluations. You might like to consider the following questions:

#### **Carrying out the evaluation**

- What went well in your evaluation? What challenges did you face?
- Was your initial budget realistic? Did you spend more, or less, on any aspects of the evaluation than anticipated?
- Did you stick to your original timeline? Which elements of your evaluation took more or less time than anticipated? Are there any aspects to which you would allocate more time in future evaluations?
- What will you do differently in future evaluations?
- What additional support would have been beneficial? Where could you find this support if you carry out further evaluations?

#### **Teaching practice**

- Will you use the innovation in your practice? If so, who will you use it with?
- How will you monitor the impact of new teaching approaches in the future?

#### **Next steps**

- What do you plan to do now as a result of this evaluation project?
- Will you apply for funding to evaluate your innovation further?
- Would you be interested in carrying out further evaluations of innovative practice?
- Will the experience of carrying out an innovation evaluation change anything in your practice more generally?

## References

British Educational Research Association, Ethical Guidelines for Educational Research  
<https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018>

Community Tool Box, Logic Model resources  
<https://ctb.ku.edu/en/table-of-contents/overview/models-for-community-health-and-development/logic-model-development/main>

Education Endowment Foundation, DIY Evaluation Guide  
<https://educationendowmentfoundation.org.uk/tools/diy-guide>

Education Endowment Foundation, Putting evidence to work: a school's guide to implementation  
[https://educationendowmentfoundation.org.uk/public/files/Publications/Implementation/EEF\\_Implementation\\_Guidance\\_Report\\_2019.pdf](https://educationendowmentfoundation.org.uk/public/files/Publications/Implementation/EEF_Implementation_Guidance_Report_2019.pdf)

Education Endowment Foundation, Spectrum database  
<https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/measuring-essential-skills/spectrum-database/>

Institute for Effective Education, Engaging with Evidence Guide  
<https://the-iee.org.uk/what-we-do/engaging-with-evidence/>

Institute for Effective Education, Unplanned Analyses, Data Dredging and Cherry Picking  
<https://the-iee.org.uk/2020/10/15/unplanned-analyses-data-dredging-and-cherry-picking/>

Institute for Effective Education, Writing up your innovation evaluation report guidance  
<https://the-iee.org.uk/wp-content/uploads/2017/03/Writing-up-your-innovation-evaluation-report.pdf>

Kraft, Interpreting Effect Sizes in Educational Interventions  
<https://scholar.harvard.edu/mkraft/publications/interpreting-effect-sizes-education-interventions>

Research Methods Knowledge Base, External Validity  
<https://conjointly.com/kb/external-validity/>

Research Methods Knowledge Base, Internal Validity  
<https://conjointly.com/kb/internal-validity/>

The Center for Theory of Change  
<https://www.theoryofchange.org/what-is-theory-of-change/>

## Sources of evidence

Association for Science Education <https://www.ase.org.uk/>

Campbell Collaboration <https://campbellcollaboration.org>

Chartered College of Teaching <https://chartered.college>

Deans for Impact <https://deansforimpact.org>

Early Intervention Foundation [www.eif.org.uk](http://www.eif.org.uk)

Education Endowment Foundation Early Years Toolkit  
<https://educationendowmentfoundation.org.uk/evidence-summaries/early-years-toolkit/>

Education Endowment Foundation Guidance Reports  
<https://educationendowmentfoundation.org.uk/resources/guidance-reports/>

Education Endowment Foundation Teaching and Learning Toolkit  
<https://educationendowmentfoundation.org.uk/resources/teaching-learning-toolkit>

EPPI-Centre <https://eppi.ioe.ac.uk/cms>

Institute for Education Sciences <https://ies.ed.gov>

Institute for Effective Education Best Evidence Encyclopaedia [www.bestevidence.org.uk](http://www.bestevidence.org.uk)

Institute for Effective Education Best Evidence in Brief [www.beib.org.uk](http://www.beib.org.uk)

Institute for Effective Education Evidence 4 Impact [www.evidence4impact.org.uk](http://www.evidence4impact.org.uk)

National Centre for Excellence in the Teaching of Mathematics <https://www.ncetm.org.uk/>

National Foundation for Educational Research <https://www.nfer.ac.uk/>

The Nuffield Foundation <https://www.nuffieldfoundation.org/>

Research Schools Network <https://researchschool.org.uk>

Teacher Development Trust <https://tdtrust.org>

The Learning Scientists [www.learningscientists.org](http://www.learningscientists.org)

Wellcome <https://wellcome.org/>

What Works Clearinghouse <https://ies.ed.gov/ncee/wwc/>

## Glossary

**Assignment to condition** – relates to how participants are allocated to different conditions.

**Condition** – refers to the group participants are in (eg. they may be in the intervention or control group).

**Contamination of the control group** (sometimes called diffusion of treatment) – the control group receives some or all elements of the innovation being trialled. This usually means it is impossible to draw conclusions about the impact of the innovation.

**Control group** (sometimes called the comparison group) – the participants who do not receive the innovation.

**Active control group** – the control group receives a different intervention or a different version of the innovation.

**Business as usual control group** – the control group receives whatever teaching and support the school and teacher would usually provide. This is the typical control group used in impact evaluations in schools (no treatment control groups and placebo control groups are rare in educational evaluations).

**Waiting list control group** – sometimes it is not possible to deliver the innovation to all of the people who meet the inclusion criteria at the same time (eg. you might not have the staffing capacity or have enough rooms available in the school) so the innovation is run over two cohorts. The people selected to be in the second intervention cohort can form a waiting list control group: the innovation will be delivered to them if it proves to be effective for the first cohort.

**Effect size** – a measure of the difference between the outcomes for the participants who experienced the innovation (the intervention group) and those who did not (the control group). In education research this is usually between -1.0 and +1.0. If the number is negative, the control group performed better than the intervention group. If it is positive, the intervention group performed better. The larger the effect size, the larger the difference between the two groups.

**External validity** (sometimes called generalisability) – the extent to which research findings can be applied to other situations and populations. More information about threats to external validity and ways to improve external validity can be seen [here](#).

**Focus group** – a group of participants who are interviewed together. Focus group participants interact to respond to questions or topics from the interviewer; this is in contrast to other forms of group interview where participants are interviewed together but provide individual responses to questions.

**Gantt chart** – a visual representation of a timeline which shows the tasks to be completed during certain time periods as horizontal bars.

**Hypothesis** (plural **hypotheses**) – a prediction, based on what is already known, that can be tested to see if it is true.

**Impact evaluation** – a type of experiment which considers the impact of an innovation by carrying out simultaneous outcome and process evaluations.

**Implementation fidelity** (sometimes referred to as treatment integrity) – refers to whether the innovation was delivered as intended.

**Inclusion criteria** (sometimes referred to as eligibility criteria) – the characteristics prospective participants must have in order to be included in an evaluation. This may include attainment, demographic characteristics or previous experiences. If your innovation is targeting a specific group of pupils (eg, those with low prior attainment, poor attendance etc.), the inclusion criteria will include having been identified within this group.

**Exclusion criteria** – characteristics prospective participants must **not** have if they are to be included in the evaluation. This may include having experienced the innovation in the past or engaging in a similar innovation during the evaluation period as well as demographic, attainment or behavioural characteristics.

**Intention to treat analysis** – participants' data is analysed in the condition to which they were initially allocated, whether or not they participated in all or any innovation activities.

**Internal validity** – the extent to which we can be confident that the innovation has caused any effect identified in the evaluation. There are many threats to the internal validity of an evaluation, and you can read more about them [here](#).

**Intervention group** (sometimes called the treatment group or experimental group) – the participants who experience the innovation.

**Interview** – a conversation in which the interviewer asks questions to elicit information from the person being interviewed.

**Mean** – the average that is computed by adding up the numbers and dividing by the number of numbers. This can be used to find the average of data where the distance between scores is constant (eg, height, speed, test score).

**Median** – this is the value above which half of the figures lie. This can be used to find the average of data where the distance between scores is not constant (eg, ratings on a scale of 1–10, ranking of pupils' performance on a test).

**Mixed-method research** – studies which use both qualitative and quantitative methods.

**Mode** – the value that most commonly occurs in a group of scores. This is used as an average for non-numerical, categorical data (eg, participant reports of their favourite element of an innovation).

**Outcome evaluation** – this is the part of the study which considers whether or not the innovation has an impact on pupil achievement or other outcomes.

**Outcome measure** – a test that can be used to measure performance.

**Participant** – an individual who takes part in an evaluation.

**Population** – all individuals who could have formed part of the sample (ie, all people who meet the inclusion criteria).

**Post-test** – an outcome measure administered after participants have completed the innovation.

**Delayed post-test** – an outcome measure administered weeks or months after participants have completed the innovation.

**Pre-test** (sometimes called a baseline measure) – an outcome measure administered before participants begin the innovation.

**Process evaluation** – the part of the evaluation that collects and analyses data to determine whether the innovation was delivered as intended (implementation fidelity) and what participants thought of the innovation. It generally involves carrying out observations of the innovation in action and conducting interviews or questionnaires with staff and pupils who experienced the innovation.

**Qualitative data** – non-numerical data. This information is usually (but not always) in the form of words (eg, written or spoken answers to open questions, narrative reports of observations).

**Qualitative research** – studies for which the results are expressed in narrative interpretations of the data.

**Quasi-experiment** – experiments in which the intervention and control group are not allocated randomly.

**Quantitative data** – data in the form of numbers, which is generally analysed using statistical methods.

**Quantitative research** – studies which report their findings using numerical data.

**Questionnaire** – a series of written questions completed by research participants.

**Random assignment** – a method of assigning research participants to intervention or control groups merely by chance.

**Randomised controlled trial (RCT)** – a research design in which participants are randomly assigned to receive the innovation or not. It increases the likelihood that the two groups are equivalent at the beginning, increasing the chance that differences between the groups' outcomes are due to the innovation.

**Reliability** – considers whether the results of an investigation (or an individual measure) are stable and consistent and so would lead to the same results with the same sample on different occasions.

**Sample** – the group of people who take part in an evaluation. The individuals in the sample are called participants.

**Standardised measures** – outcome measures which are administered and scored in a standardised way, allowing comparisons in performance to be made across large groups of people. These tests are externally written, and are sold by publishers following extensive testing.

**Equivalent versions of standardised measures** – some standardised measures have equivalent versions of test papers, which measure the same skills using different questions. These can be administered at different times so different but equivalent tests can be taken as pre- and post-tests.

**Standard deviation** – this is a measure of the variability of a group of scores. The smaller the number the less variance there is among the scores.

**Statistical significance** – relates to whether the results are likely to have occurred by chance according to tests of statistical significance.

**Structured observation** (sometimes called a systematic observation) – an observation that uses pre-determined, clearly defined categories to record behaviour or events.

**Observation schedule** – a form completed by the observer during a structured observation.

**Survey** – a method of finding out about peoples' opinions or experiences by asking them questions. Survey methods include carrying out interviews or asking participants to complete a questionnaire.

**Tests of statistical significance** – statistical tests which determine whether findings are likely to have been caused by chance, which usually give results as p-values. P-values fall between 0 and 1, and the smaller the number the less likely that the result occurred by chance. If p is less than 0.05 the result is said to be statistically significant and is unlikely to be due to chance.

**Thematic analysis** – a way of analysing qualitative data which identifies, analyses and records patterns within the data.

**Theory of change** – a model which explains how and why you expect a new initiative to lead to changes in outcomes by outlining the causal links from the innovation to the outcome.

**Well-matched** (sometimes called equivalency) – a term used to consider how similar the intervention and control group are before the innovation begins. As a guide, the two groups' mean scores should be less than 0.25 standard deviations from each other if they are to be seen as well-matched in relation to that measure. If groups are well-matched, they are sometimes referred to as being equivalent.

# Appendix A

## Evaluation planning

### Introduction

#### 1) Problem: Describe the problem or issue you want to address.

What challenge(s) do your school(s) have that need to be addressed? Identify a single, clear problem and a hypothesis of what may be causing or sustaining the problem.

#### 2) Existing evidence: What existing research evidence exists?

What does the existing research evidence say about your issue and potential solutions? What solution will you use to address your problem? Outline your theory of change.

#### 3) Research question(s) or hypothesis

What impact does [what practice?] delivered [over how long?] have on [what outcome?] for [whom]?

**4) Method: Describe exactly how the evaluation will be conducted.**

**a) Sample**

Who will participate in your evaluation? What inclusion criteria (if any) must your participants meet? How will you recruit participants, or other teachers/schools, to participate? What is the likely number of participants? What consent will be gathered for them to participate?

**b) Assignment to condition**

How will you allocate participants to the intervention and control groups? How will you check the intervention and control group are well-matched before the innovation begins?

**c) Innovation**

Describe your innovation in detail. What training and support will be given to those delivering the innovation? What resources need to be created, and when will this be done? What will the control group do? How will you make sure the control group does not receive any elements of the innovation?

**d) Outcome measures**

What outcome measures will you use? When and how will they be administered and scored?

**e) Process analysis**

How and when will fidelity of implementation be checked? Who will carry out fidelity checks? Will you also check the control group are not using the innovation? Which participants' views will be gathered? How and when will this be done? Who will be responsible for gathering participant views?

FIDELITY

PARTICIPANT VIEWS

**f) Analyses**

How will you analyse your outcome data? What analyses and comparisons will you make? Will you analyse the data for any sub-groups of participants? If you plan to analyse data for particular subgroups, state which group(s), how the groups will be identified and defined, and what data analysis you will carry out. How will you analyse the quantitative and qualitative data from your process evaluation?

### g) Timeline

You can use this table to create a detailed timeline for your project plan, including milestones and who will be responsible for these. Alternatively you may prefer to create a timeline using a different format, such as a Gantt chart.

Activity	Start and end date	Person responsible for activity

### h) Budget

Either complete the table below or create a spreadsheet with costs and when the costs will occur.

Budget Item	Date money will be needed	Amount
<b>Total Expenditure</b>		

## Appendix B

### Guidelines on writing a survey

Here are some general suggestions and things to think about when writing a survey.

#### *Purpose*

- Be clear about what you want to know from the survey and why. Make sure the survey will help you to answer your research question. If you are writing the survey for the process evaluation element of an impact evaluation, you could write specific process evaluation research questions.
- Consider how you might want to analyse or present your findings and make sure you collect the right data in a way which will allow you to do this.
- It can help to justify to yourself why you are asking each question and how it will contribute to meeting the objectives of your research.

#### *Structure*

- Include brief, clear instructions at the beginning of your survey, and if necessary when types of questions change. Giving examples for any items which might be confusing (e.g. ranking items) can be helpful.
- Order of questions
  - Organise questions in a logical sequence, and group similar questions together.
  - If practical, start with some easy and non-threatening questions.
  - Do not put the most important question at the end of a long questionnaire.
  - Think about whether earlier questions may influence response to later questions (question order effects).
- You need to decide what degree of structure to use in your questions (eg. open or closed questions, or a combination of the two; open questions give more detail, but closed questions are quicker and easier to codify and analyse).
- If you plan to use any closed questions with multiple response choices it can be useful to:
  - Ask a pilot group the question in an open question format first to increase that chance that you include all reasonable responses to the questions. The pilot group should be people who are not participating in the evaluation, but who are similar to your participants (eg. pupils in the same year, teachers of the same subject).
  - Think about whether the order in which you present response options for multiple choice questions might influence people's response (response order effects).

#### *Writing questions*

- Wording
  - Clearly define terms which might mean different things to different people to ensure all items will be interpreted as you intended by all respondents.
  - Avoid jargon.
  - Write out or say acronyms in full the first time they are used.

- Short questions are generally better than long ones.
- Avoid negatively framed questions.
- Only ask about one idea in each question.
- Avoid biased or leading questions.
- Take care with sentence structure to avoid ambiguous questions.
- Awareness of audience
  - If you are using a written questionnaire, consider the approximate reading age of the people who will complete the questionnaire and write questions appropriate to this.
  - Be aware of the likely knowledge of your respondents (especially if they are not school based, e.g. if you are gathering parent views) and take this into account when writing questions.
- If asking questions about behaviour avoid using the term "usual", as it makes questions a bit ambiguous. It can be useful to include a time period in questions about behaviour (e.g. "Last academic year, what approaches did you use when teaching ...?").
- If you decide to use a rating scale to gather responses for any of your items, use no more than 5 points on your scale (if you need to use more than a 5 point scale include a visual aid with the options presented on it - people tend to get confused by rating scales with 6 or more options unless they can see it).
- Where possible; avoid questions which participants may see as threatening. If they are necessary:
  - Reinforce the purpose of the questionnaire.
  - In questionnaires open questions can be best for potentially threatening questions as they allow respondents to qualify or justify their responses.

### *Pilot the survey*

- If possible ask some people who are not participating in the research, but who are representative of the group who will complete the survey, to try out the survey before you use it in your research.
- Ask for feedback on any items they found unclear, any additional questions they felt should be added, whether there were any missing options for multiple choice responses, if multiple choice options overlap or are confusing, questions that they found intrusive or inappropriate etc.
- If it isn't practical to carry out a pilot with a group representative of your participants, ask colleagues, family or friends to read the questionnaire to check whether the wording and meaning is clear.

### *References*

- Mertens M D (2005). *Research and evaluation in education and psychology*. London: Sage.
- Robson C and McCarthan K (2016). *Real world research*. Chichester: John Wiley & Sons.



This guide is produced by the Institute for Effective Education (IEE).

Written by Dr Alicia Shaw.

Designed by Nicola Mann

### **Contact us**

+44 (0)1904 328166 | [info@the-iee.org.uk](mailto:info@the-iee.org.uk)

Ron Cooke Hub, University of York, York YO10 5GE

Twitter: [@IEE\\_York](https://twitter.com/IEE_York) | [the-iee.org.uk/](http://the-iee.org.uk/)

© Institute for Effective Education, 2020

The Institute for Effective Education (IEE) is an independent charity working to improve education for all children by promoting the use of evidence in education policy and practice.

The Institute for Effective Education is a charity registered in England, charity number 1168744

Institute for  
**Effective Education**

Empowering educators with evidence

