THE DIY EVALUATION GUIDE

Professor Rob Coe and Stuart Kime

Durham University

Camilla Nevill and Robbie Coleman

Education Endowment Foundation

January 2013

INTRODUCTION

WHAT IS THE DIY EVALUATION GUIDE?

The DIY Evaluation Guide is an accessible resource for teachers which introduces the key principles of educational evaluation and provides guidance on how to conduct small-scale evaluations in schools.

The guide explains the importance of "Do It Yourself" evaluation. It outlines a range of options open to teachers who want to improve the way they evaluate new interventions or strategies and provides practical advice on designing and carrying out evaluations.

WHY IS DIY EVALUATION USEFUL?

DIY evaluation is useful for three reasons:

- **It indicates whether or not an intervention is effective.** Without evaluation, it is impossible to know whether an intervention is having a positive impact on learning. It can be tempting to implement plausible-sounding strategies which, in reality, don't benefit students.
- **Evaluation saves teachers time.** It is often easier to start doing something new than it is to stop doing something which has been running for a number of years. The evaluation strategies in this guide show how data can be used to obtain an estimate of the impact of a particular intervention. Teachers can then put their time and effort into the most effective things and avoid pursuing approaches that do not work.
- **Evaluation guides future action.** By investing a little time in carefully recording what is being done and by measuring its outcome, it is easy to identify improvements for the future.

HOW SHOULD THE DIY EVALUATION GUIDE BE USED?

It is important to make a distinction between DIY evaluation, which can be undertaken by teachers and take place in a single school or class, and other forms of evaluation such as randomised controlled trials, which will usually take place across large groups of schools and be led by full-time researchers.[1] **Both forms of evaluation are useful, but they serve different purposes.**

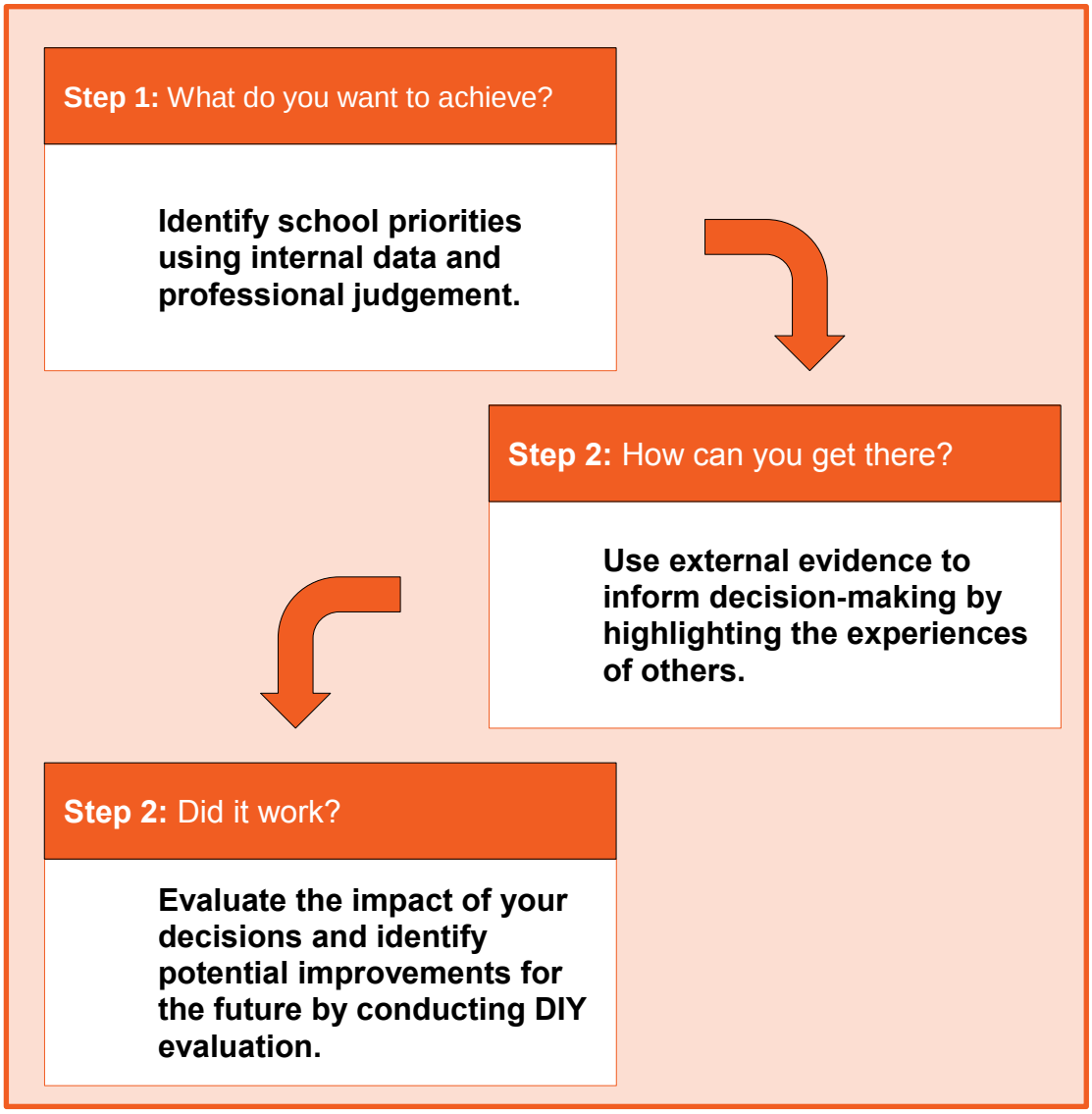
Large-scale evaluations provide the most robust estimate of an intervention's average effectiveness. This kind of evidence can inform teachers' decision-making by highlighting the average experience of others and by picking out the common features which appear to lead to the highest impact. However, the average effect of an approach will not always match its impact in a given school. An approach may be more effective in some contexts than in others. DIY evaluation is essential in order to determine whether or not an intervention is having the impact that was hoped for.

A useful comparison might be made with health. In the healthcare sector, large-scale trials are used to determine whether or not drugs should be licensed and to provide doctors with the information they require to make prescriptions. Then, the doctor (and patient) will determine whether or not the treatment is having the desired effect, and if it isn't, they will adjust the treatment or try another approach.

Thus, large-scale evaluations and smaller DIY evaluations are complimentary, not substitutes. Box 1 shows how the DIY Evaluation Guide could be used in conjunction with external evidence to evaluate the impact of the Pupil Premium (though the same principles could be applied to any decision).

[1] For example, the EEF is funding a number of large-scale trials that you can read about here: http://educationendowmentfoundation.org.uk/projects.

Information Box 1. Using DIY evaluation to evaluate your Pupil Premium spending



Step 1: What do you want to achieve?
Determine your priorities using internal data and professional judgement. What are your ambitions for improvement? In which particular areas do pupils need more support?

Step 2: How can you get there?
Having identified what your goal is, education research can be useful in providing information about what has worked elsewhere, therefore highlighting the important features of implementing a particular approach. The *Sutton Trust-EEF Teaching and Learning Toolkit* is one resource which provides summaries of education research for this purpose.

Step 3: Did it work and should you continue?
Once a decision to adopt a new strategy has been made, it can be evaluated using the information in the *DIY Evaluation Guide*. The information from this evaluation can be used to inform decisions in future years and shared with other schools.

THE STAGES OF DIY EVALUATION

Table 1, below, sets out the eight steps which comprise a good DIY evaluation. We have split these steps into three stages which are reflected in the sections of this document: *Preparation*, *Implementation*, and *Analysis and Reporting*.

Table 1: The Stages of DIY Evaluation

Step	Description	Page
Stage 1: Preparation		
1. Frame your evaluation question	This is the question that your evaluation will set out to answer.	**4**
2. Decide your measure	This is what you will use to assess whether an approach has been successful. For example, standardised reading, writing, maths or science tests.	**6**
3. Decide your comparison group	This is to understand what would have happened to pupils if you did not implement the new approach. For example, you could compare with pupils in the same or a different class.	**9**
Stage 2: Implementation		
4. Conduct a pre-test	This is to understand pupils' starting point on the outcome measure or form the groups in matched designs. Pupils in your intervention and comparison groups should be starting from the same point.	**17**
5. Deliver the intervention	Deliver the intervention as planned and record exactly what happened. You should ensure that your comparison group does not receive the intervention.	**17**
6. Conduct a post-test	This is to understand the impact of the intervention on the outcome measure. The post-test should be implemented at the same time with both the intervention and comparison groups.	**18**
Stage 3: Analysis and Reporting		
7. Analysis and interpretation	Record the results in a spreadsheet and then calculate the effect on attainment.	**20**
8. Report the results	It is important to report the results clearly, for example using a PowerPoint presentation.	**21**

1. PREPARATION

This section describes the steps you need to go through to prepare for your evaluation including deciding your research questions, planning your outcome measure and establishing your comparison group.

1.1 FRAME YOUR EVALUATION QUESTION

The first step to conducting a DIY evaluation is to identify the question which you are going to investigate as clearly as possible. Though it may sound obvious, without this step, good evaluation is almost impossible.

A number of types of question are possible. A simple question might be whether a particular intervention boosts attainment or not (for example, does having a mentor boost performance in GCSE English?). Alternatively, you could investigate whether a particular form of an intervention works better than another (is weekly mentoring more effective than monthly mentoring?).

DIY Evaluation Glossary: Intervention
Any programme, policy or practice we wish to evaluate. Interventions might be targeted (e.g. small group tuition or a catch-up reading programme) or more general (e.g. a new way of marking books).

Once you have found an approach which you believe may work in your school, a good way to frame the question you wish to answer is to fill out the blanks in the following sentence 10 times:[2]

I would really like to know if _____ [*intervention*] would have an impact on _____ [*outcome*] in our school.

It is critical that a good evaluation question encompasses the following three elements:

- **CHOICE**: choice to be evaluated;
- **OUTCOME**: outcome that will be measured;
- **CONTEXT**: people to be measured and context.

Case Study 1: Framing your evaluation question at Parkview

The English department at Parkview School decided it was important to have a consistent marking policy, but was not sure which method of marking would have the biggest effect on reading.

Discussions led to the teachers narrowing down the choice to grading with ticks and crosses and a comment or providing comment-only marking. Opinion was divided about which is better so they formed the following research question:

What impact does using comment-only or graded marking have on Parkview Schools pupils' reading comprehension over one year?

They agreed that one class in each year group would be randomly allocated to each approach and they would assess the results using a standardised reading test at the end of the year. By evaluating the different ways of assessing, the teachers were able to develop the best possible assessment policy for their pupils.

[2] This idea is taken from Michael Quinn Patton's 'Utilization-focused Evaluation' (1997), Thousand Oaks, California: SAGE.

The following are further examples of well-framed evaluation questions:

'What impact does the Stoneley School's new Year 7 reading support group have on reading support group students' reading abilities?'

'What impact do three different types of effective feedback have on the writing performance Year 6 pupils in Newgate School?'

Information Box 2: Tips on framing evaluation questions

- Once you have a shortlist of questions try to make them as specific as possible. Vague questions are a barrier to successful evaluation.
- Be clear about the intended outcome and how you want to measure it.
- Be clear about what approaches you want to test and the changes that need to be made in order to deliver them properly.
- Be clear about the group of pupils you want your findings to apply to. Approaches that work with one group may not work with another.
- Involve colleagues in the discussion.

1.2 Decide your Measures

Once a question has been defined, the next step is to determine the measure against which success will be judged. Although many different outcomes are important in education, this guide focuses on measuring academic attainment. In this section we discuss different ways of measuring attainment.

Deciding your outcome measure

When deciding your outcome measures there are three main sources from which you can choose:

A. **National assessments**

B. **Standardised tests from reputable suppliers**

C. **Design your own**

It is important not to over-burden children with too much testing. Often, you will want to use tests your school already uses as outcome measures. However, when you need a new test it is important to understand the alternatives, and the advantages and disadvantages of other options. For example, by designing your own test, you can tailor your assessment to exactly what you want to measure, but it might not be as reliable as a test designed by a reputable supplier. A full list of advantages and disadvantages is outlined in Table 2 and a list of reputable suppliers is given in Box 4.

DIY Evaluation Glossary: Outcome measure
The test or exam which provides the data used to analyse and intervention's effect. It is important only to measure outcomes which are of value and not simply those which are easiest to measure.

Above all you are trying to select a measure which is both *valid* (meaning that it measures what it claims to measure) and *reliable* (meaning that it is consistent over time and context). Box 3 outlines a set of quality criteria you should apply to any outcome measure. If you cannot answer these questions satisfactorily then you should probably not use the test.

Deciding your pre-test measure

Your pre-test measure should be as closely aligned to, or predictive of, your outcome measure as possible. In most cases it will make sense to use the same assessment as the pre-test. However, in some cases this is not possible. For example, if the intervention lasts a whole year, an assessment that would be suitable at the end of the year may be too hard at the beginning, especially at younger ages. In these cases, you will need to use something that is similar, but easier.

Information Box 3: Tips for determining the quality of an assessment

The following are a set of questions to ask of any test or assessment that you want to use. Any provider of high-quality standardised tests should have made sure that their tests pass these quality criteria.

1. Can you define clearly what the assessment measures are?
2. Ask other teachers to look at the assessment content, required responses and marking:
    a. Do they think it looks appropriate?
    b. Could anything other than what the assessment is supposed to measure influence the outcomes? (e.g. Reading comprehension influencing the results of a science test).
    c. Does the assessment cover the full range of measures, in terms of content and level? Are there any gaps? Is it too easy or hard for some?
3. Does the assessment predict what it should? Does the assessment correlate with other measures of attainment or your own assessments?
4. Are the outcomes reliable? If the assessment was repeated or marked by someone else, would you get the same results?

You can find more detail on the quality criteria that the EEF uses for judging assessments here: http://educationendowmentfoundation.org.uk/uploads/pdf/EEF_testing_criteria.pdf.

If your intervention is short then you might want to use a different assessment from the outcome measure to prevent "practice effects", e.g.; where students remember answers from previous tests. To avoid these effects, you could use old SAT papers as the pre-test and this year's exam as the post-test. Alternatively, many standardised tests provided by suppliers have multiple forms.

Information Box 4: Providers of standardised tests of attainment

The following are a list of organisations that provide high-quality standardised tests of attainment that have been used by the EEF:

- The Centre for Evaluation and Monitoring at Durham University
- GL Assessments
- Hodder Education
- Pearson Education

Each provider has a large bank of tests, some of which might be more reliable or useful than others. You should consider the reliability and validity of the test (see Box 3) and whether it is aimed at the right age and level, as well as the practical factors, for example, are the tests on paper or digital and how long will the test take?

Table 2: Sources of assessments of attainment

Explanation	Strengths	Limitations
A. National Assessment		
Use old or current national test papers (e.g. SAT or GCSE papers) as the pre-test and post-tests.	• The best predictor of actual performance is national tests. • Cheap. • Good practice for pupils.	• May not be as reliable as some external tests provided by reputable suppliers. • May not be tailored to the needs of your pupils or to the focus of the intervention.
B. Standardised tests		
There are many providers of high-quality standardised tests of attainment. See Boxes 3 and 4 for a list of providers and some of the issues to consider.	• Are likely to be reliable and valid and should be able to provide information on the criteria. • Often standardised using national populations so you can compare your children's attainment to national norms. • Can be highly predictive of performance in national tests and some may provide predictions as well as actual scores. • Digital tests can provide instant results and rich data on individual children.	• May be expensive as many providers are commercial organisations. • May not be aligned with the curriculum, or the specific area in which you are interested.
C Design your own		
Design your own assessment completely from scratch or combine sections of other assessments you have used in the past.	• You can tailor the assessment to suit your own needs including the subject area being assessed, and the age and ability range of your pupils. • Cheap. • Your school may already have home-made tests that it uses.	• Home-made tests may not be as reliable as external tests provided by reputable suppliers which will have been thoroughly piloted. • Home-made tests will mean you cannot compare to national norms. • Ideally, you will need to design more than one version of the test in order to account for practice effects.

Case Study 2: Pre-tests and outcome measures at Notre Dame RC Girls School

Teachers at Notre Dame Roman Catholic Girls' Secondary School in Southwark wanted to run a summer school targeted at its incoming Year 7 pupils eligible for free school meals. They wanted to work out the most effective possible form the school could take so decided to evaluate two alternatives (an in-house school and a school run by an external company) against one another.

The teachers collaborated with colleagues at Notre Dame's feeder primaries and used pre-tests produced by the National Foundation for Educational Research to allocate pupils between the two summer schools. The tests were conducted in the Spring term of Year 6.

The students were tested again in the Autumn term and the staff used this data, in addition to attendance data, staff observations and student survey results to determine which of the schools to use in future years.

1.3 DECIDE YOUR COMPARISON GROUP

After determining the outcome and pre-test measure, comes the most important step in DIY evaluation: establishing a comparison group or 'control' in order to understand the impact of the approach you are testing. The key factor in this decision is whether or not you are in control of who gets the intervention. If you are, then we recommend that random allocation (Option A below) should be used. If not, then we recommend matched control groups (Option B below).

DIY Evaluation Glossary: Comparison group
Sometimes called a "control group", this group does not receive the intervention, and as a result allows us to estimate what would have happened in the absence of the intervention. The comparison group may either continue with "business as usual" or receive an alternative intervention.

Children will almost always progress as time passes, but you want to find out whether a new way of teaching means that they progress faster than usual. This means delivering the new approach to one group and then comparing to a group which is not receiving the new approach, or is instead receiving 'business as usual'. The comparison between a group receiving the intervention and one continuing as normal, allows us to estimate what would have happened without the intervention, a concept known as "the counterfactual". We can never truly know exactly what would have happened otherwise, but a control group proves the best possible estimate.

Some teachers may feel it is difficult to justify giving something to some students while withholding it from others, and this feeling is a common barrier that prevents teachers from evaluating. Yet, you could also say that it is unethical to deliver an approach you are not sure works to students without testing it first. Box 5 discusses these ethical issues in more depth.

In this section we explain two types of comparison group (also known as "evaluation designs"):

- Random allocation (Option A)
- Matched control group (Option B)

Ideally, the pupils in your comparison group should be as similar as possible to those in your intervention group in terms of their prior attainment and other characteristics that might affect performance (e.g., they should be from the same school and year group, with a similar mix of demographics).

Each of the methods for establishing a comparison group has strengths and limitations. Often there is a trade-off between the robustness of the design and the practicalities of delivering it, for example. However, each of the designs is much better than no evaluation at all.

Information Box 5: Addressing ethical issues

Teachers may feel that it is unfair to give a new intervention to some students but not to others. However, prior to evaluating we cannot know for sure whether something is actually valuable. Ethical evaluations start from a position of *'equipoise'*; where we do not know for certain what works best. If we are uncertain which of two options works best, it can be argued that it is unethical not to try and establish which is more effective, particularly if the intervention is going to be repeated or rolled-out more widely.

It is also often possible to evaluate interventions which avoid withholding them from pupils. For example, in many cases evaluation is about trying out different ways of delivering an intervention to see which works best. There is also a variety of other designs in which all participants get some form of intervention, described in more detail in the section below.

It is important to explain that an evaluation of a new possible method is taking place, not an evaluation of a better method as if we knew it was better, we'd be doing it already. In addition, it is a general principle of research that participants should give informed consent to take part. However, schools routinely innovate, try out new approaches and informally evaluate them all the time. You should use your own judgement and usual process when it comes to deciding whether to gain consent for children to take part in either the intervention or the testing.

Evaluation Design "A": Random allocation

Random allocation is the most robust way to establish a comparison group. If the choice about whether participants (or participating classes, teachers, etc.) are in the treatment or control group is decided purely by chance, this guarantees that any initial differences between the groups result only from chance. If the groups are large enough and the random allocation has been done properly, this method makes it extremely likely that the groups will be equivalent with respect to every possible characteristic. Therefore, the only difference between the two groups' progress at the end of the trial will be the impact of the intervention.

Without random allocation there are likely to be systematic differences between the groups. For example, one may be taught by a different teacher, or if the intervention is optional then only the most enthusiastic students or teachers might take it up. When this happens it is impossible to say whether it was these differences or the intervention that made the difference.

There is a good explanation of why randomisation is important in *Test, Learn Adapt: Developing Public Policy with Randomised Controlled Trials* published by the Cabinet Office.[3]

There are a number of different versions of random allocation to consider:

- **Business as usual:** the control group is taught "normally', i.e. as they were before the intervention (which may or may not be an improvement) was proposed (see Case Study 3).

- **Alternative treatment:** When you want to compare two competing approaches you can allocate pupils or teachers randomly to each group. If both approaches are plausible, it might be fairer to allocate randomly than letting people choose which one they do. If you let them

[3] http://www.cabinetoffice.gov.uk/sites/default/files/resources/TLA-1906126.pdf.

choose one might be more popular which could make delivery problematic, and there are likely to be systematic differences in who chooses which version (for example one might be favoured by better teachers or higher-ability pupils).

- **Compensation:** Sometimes you might want to compensate the children in the control group by providing them with something different. For example, in Case Study 4, pupils who received peer tutoring in maths, could be compared to children receiving peer tutoring in reading (with the group not receiving that subject providing the control). When using this design, it is important that the alternative does not have an effect on the outcome you are measuring.

- **Waiting-list design:** Here, everyone gets the intervention in the end, but random allocation (a lottery) decides who gets it now and who gets it later. The 'later' group acts as a control in the first phase. This design is particularly appropriate when constraints on personnel or resources mean that not everyone can get it at the same time: making the choice at random is not only fair but allows the impact to be reliably estimated. See Case Study 5.

- **Border-line randomisation:** This can be used when an intervention is intended for the neediest pupils, such as a reading catch-up programme. One group should definitely get the intervention (those behind), another group definitely does not need it (those ahead), but there might be a third group for which you do not know whether giving them the intervention is the best use of resources. Pupils in this borderline group can be allocated at random and their results are compared.

Case Study 3: Random allocation in action at Higher Sandford

Does mentoring Year 8 underachievers at Higher Sandford, a large comprehensive school improve achievement in English, Maths and Science as measured by the CEM's MidYIS standardised tests?

In 2003, the Department for Education and Skills (now the Department for Education) promoted mentoring for Year 8 underachievers. Having seen that the evidence for the success of mentoring is mixed, the Year 8 coordinator set out to test the effect that their programme had on achievement. The mentoring programme involved a one-to-one session once a fortnight.

Evaluation
Twenty students were identified as underachieving in at least two core subjects at the end of Year 7. These students completed the MidYIS test of English, Maths and Science performance at the start of the year. Students were then grouped by sex, ranked by performance and paired with someone of the same sex with a similar score. One from each pair was then randomly allocated to receive the mentoring using a random number generator. Consent was gained from parents for those receiving the intervention. At the end of the year the 20 pupils completed the MidYIS tests again and the two groups' progress were compared.

Results
An analysis of progress between the baseline and post-test measures showed no significant improvement as a result of the mentoring. Qualitative interviews with staff and pupils suggested that a mentoring programme had the potential to stigmatise underachievers. The mentoring programme was stopped.

The results were reported in the Year 8 coordinator's MA thesis.

Case Study 4: random allocation in the Fife Peer Tutoring project

Although it is on a larger scale than this guide is designed to support, the Fife Peer Tutoring study offers a good example of how randomised trials have been used to evaluate the impact of a strategy that the Teaching and Learning Toolkit suggests has high potential. Reading about such studies is a good way to develop your understanding of them, and to help design your own.

In this study, schools and classes were randomly allocated to either the English or Maths interventions. Sometimes older children tutored younger children across year groups; sometimes pupils tutored others of the same age within year groups.

All children received something, and the design is efficient as one intervention group acts as the control for the other, and vice versa.

Case Study 5: a waiting-list design in action [fictional]

Having read a research summary on the efficacy of small group tuition, a group of four English teachers designed an evaluation of proving small group tuition in writing skills to a group of Year 10 pupils. The intervention was designed to last for 14 weeks. One of the issues the teachers faced was the demand on staff time – they simply couldn't provide the teacher time to accommodate 82 Year 10 students in the scheme simultaneously. As a result, the teachers decided to create a waiting-list design in which half of the students were randomly allocated to receive tuition in Term 1 and the other half were allocated to Term 2. The teachers measured the progress of all students at the beginning and end of Term 1, including those who were scheduled to receive tuition in Term 2. These students acted as the comparison group, and allowed the teachers to accurately estimate the impact of the programme.

By using this design, the teachers gave access to small group tuition to all Year 10 students in a way that was manageable with the resources available. An additional advantage of the design was that by the time the control group received the small-group tuition, their teachers had become more adept at running the sessions.

Table 3: Strengths and limitations of random allocation

Explanation	Strengths	Limitations
This means randomly allocating who gets the approach being tested. This could be done by flipping a coin or using a random number generator in Excel. Pupils, classes or schools can be randomly allocated.	• With a large enough number we can be sure there are no differences between groups. • Random allocation provides the best possible design to understand the true effect of any approach we are testing separate from other factors. • It can be a fair and simple way of deciding who gets a new intervention when you have not got enough resource to provide it for everyone (e.g., additional one to one support or a school club)	• It can be impractical to randomly allocate pupils (e.g., if you are teaching a new approach to a whole classes). Using a waiting-list or compensation approach may allay some concerns (see Case Study 5).

You should use this design if:

- You want the best possible estimate of whether something works;
- You are testing an approach where it is practical to allocate randomly who gets it (e.g., one to one support or an evening club) and you are able to control the allocation;
- You are able to justify withholding the approach being tested from students in the control group, or can offer a waiting-list or alternative treatment.

Evaluation Design "B": Matched control groups

All else being equal, random allocation is the best approach to establishing a comparison group, but where this isn't possible, a good alternative is to use a matched comparison group. In this approach, a control group of pupils are identified who are the same as those receiving the new approach on important characteristics such as prior attainment, demographics and school. The matched group might be current pupils not receiving the intervention, or previous year groups if you have historic data using the same or similar tests.

The main limitation of this approach is that there are likely to be differences between the two groups that cannot be controlled for, such as school or class. However, if you use matching you can at least control for known differences. Box 6 gives a step-by-step example of matching.

Information Box 6: A step-by-step guide to matching pupils

A teacher in a primary school wishes to understand the impact of a new reading scheme on Year 4 reading ability over the year and decides to compare this year with last year's cohort.

Step 1: Ensure both groups have taken the same baseline test
This year's results will be compared with last year's on the same reading test, since children are routinely tested at the start and end of every year. The test taken at the end of Y3 is the pre-test for both groups.

Step 2: Rank on pre-test
All the scores on this test for both groups are pasted into adjacent columns in a spreadsheet with current Year 4s coloured blue and the previous Year 4s coloured red. All the scores from both groups are sorted in order.

Step 3: Identify matches
The teacher matches each current Year 4 pupil's test score (blue) with a close score for a pupil in last year's group (red). She decides beforehand that pupils must be within two marks to be a match. By doing this, pairs of matched pupils are created.

Step 4: Check the groups are balanced
The teacher checks that the pairs are balanced. The number of pairs where blue is higher than red should be is the same as the number the other way round. Some students' scores do not match any from the previous so the teacher excludes these pupils from the analysis. The teacher also checks that the pairs are balanced on other characteristics, so they are removed from the evaluation. Importantly, the teacher ends up with two balanced groups, each with 16 children in, creating 16 matched pairs.

Step 5: Compare outcomes
The intervention is implemented over the next year, followed by the post-test. She is then able to look at the average scores for the intervention and matched control group, and observe the average differences in score between the two groups of matched pupils. The teacher compares the average scores and calculates the effect on attainment (see *Stage 3*).

Table 4: Strengths and limitations of using a matched control group

Explanation	Strengths	Limitations
Pupils who receive the new approach are matched to comparison pupils on characteristics which may affect the outcome, e.g. prior attainment, demographics and year group. You will need data from the comparison group using the same measure you use on your intervention group, e.g. we could match a current Year 4 reading group with last year's Year 4 students, or to similar pupils in in another school, as long as we have the same reading test data with which to rank and pair them.	• Often more convenient than random allocation. • Schools often have data available for the matched control group (e.g., last year's test results). • Can be more practical as you can avoid withholding the intervention from children in the same class, year or school. • Using several years' historical data in the match will strengthen the evaluation.	• The reliability of our findings will be less strong than if random allocation had been used. • It is likely that there will be some bias in the results. It is impossible to control for all the characteristics of the matched control group—they may be taught by different teachers, in a different school or there may be other differences we can not account for.

You should use this design if:

- You are unable to use random allocation;
- You are able to establish a suitable local control group;
- You have access to good data on the initial characteristics of the both groups;
- You have access to the data on the same pre-test and post-test measures for both groups;
- You are prepared to accept that the findings will be unreliable, but can provide indicative evidence and foundations for further evaluation.

Case Study 6: Matched control group in action [*fictional*]

What effect does an assessment for learning strategy have on student achievement in Year 7 English classes on a Shakespeare reading test?

A Head of English saw that there was good evidence that the traffic light assessment for learning approach might improve results. Recognising that providing high-quality feedback is not always easy the Head of English decided there was merit in trialling the approach.

Evaluation

The Head of English was unable to establish a comparison within his own school (e.g. with half the students in Year 7 taught differently), so asked another local school if their students could provide the comparison (by giving the same pre- and post-tests). Both schools gave the sample Shakespeare reading paper to students at the start of the summer term. Marking was standardised and teachers sent each other electronic scripts to blind mark. In total 90 students were in the intervention group and 90 students were in the control group.

Results

The results from both schools were recorded in a spreadsheet, pupils were matched, and the differences averaged for the matched pupils in each school and compared. There was a small significant positive effect. When considering the findings the Head of English realised that there may be differences that could not be accounted for, but on balance decided it was worth continuing the strategy and using a randomised design to test it more rigorously in Year 8.

2. IMPLEMENTATION

Once you have completed planning, this section describes the steps you need to go through to deliver your evaluation including the pre-test, intervention and the post-test.

2.1 CONDUCT THE PRE-TEST

A pre-test helps us establish where pupils start from and enables us to create a good comparison group. Pupils in your intervention and comparison groups should be starting from the same point.

You should conduct your pre-test:

- Before you start implementing the approach you are testing;
- At the same time (or time of year if you are using a prior year as a comparison) for both the comparison and treatment group;
- At a time when as many of the pupils as possible will be there to ensure you have the largest possible sample for the analysis; and,
- Before you randomly allocate or decide groups, if you are going to use random allocation (this reduces the chance of bias influencing your baseline results).

You could also collect additional information on pupils that might have an effect on outcome, such as sex, free school meal status, as it will be interesting to look at this in the analysis.

2.2 IMPLEMENT THE INTERVENTION

Prior to implementing the intervention, it is useful to write down exactly what you intend to do (e.g. how long will the intervention be delivered for?; how many times a week will it take place?; what training or preparation will teachers receive?). This will ensure that if the intervention is successful then you will know exactly what it was you did to make it work. Without this step we could end up making claims about the impact (or lack of impact) of something that was not actually implemented.

Despite this step, often the intervention may not be delivered exactly as you intended. Teachers may change it, select from it, improve it or just fail to do it properly (e.g. the plan might have been to deliver an intervention daily, but in reality it may have been delivered only once a week). It is also useful to record exactly *what was actually* delivered for a number of reasons. In cases where the intervention does not appear to be effective, you will be able to check whether this was because it didn't actually take place as intended. In cases where the intervention is effective, you may have learned something new (e.g. weekly mentoring might be effective as you had hoped daily mentoring would be).

In addition, it is useful to find out about people's perceptions of the intervention and challenges faced in delivering it to understand how it might be improved. This is known as **process evaluation** and is explained in Box 7.

Information Box 7: Process evaluation

This guide is primarily about **impact** evaluation – understanding *whether or not* an intervention has had an impact on attainment. However, in addition to impact evaluation, **process** evaluation can be used alongside to understand *how* the intervention was delivered on the ground, including:

- Was it delivered as intended?
- What are staff and pupils' perceptions of the approach?
- What has worked well and what has worked not so well?

Information from the process evaluation will enable you to understand how the intervention might be improved and whether it is practical to roll it out. There are various kinds of qualitative and quantitative data you could collect in an impact evaluation. For example:

- **Delivery records:** how many sessions were actually delivered and to whom.
- **Observations:** how an intervention is being delivered
- **Interviews with or surveys of pupils, staff and parents:** to understand their perceptions

It is important that any process evaluation does not change the intervention as a result of trying to record it. For example, observing all intervention classes could change the way teachers teach them. Observations should be unobtrusive, done on only a sample and balanced across the treatment and control groups.

You should carefully select just the data that is most relevant. You should also ensure, particularly when collecting qualitative data, that the person doing so is as independent as possible. Anyone who is committed to and has put effort into making something work may find it hard to report neutrally on it.

Finally, this kind of process evaluation data is complementary but not a substitute to good impact evaluation data. It cannot tell you whether something has worked, only how.

2.3 Conduct a post-test

This is to understand the impact of the intervention you are evaluating. It is important to think about the timing of your post-test. You should think about how long it is likely to take for the intervention to have an effect on children's attainment and ensure that you conduct your post-test after this time. You could also conduct one post-test at the end of the intervention and an additional post-test a period of time after that to see whether the effect lasts (e.g. does the impact sustain after one year).

You should ensure you conduct your post-test:

- At the same time (or time of year if you are using a prior year as a comparison) for both the comparison and intervention group; and
- At a time when as many of the pupils as possible will be there to ensure you have a large sample for the analysis.

Information Box 7: Timing of the post-test in action

A group of teachers in a primary school were designing an evaluation of a feedback intervention aimed at improving Year 6 pupils' data handling ability. They had a clear plan for all other steps of the evaluation, were happy with the reliability of the pre- and post-tests they had chosen, but were unsure about the best timing of the post-test.

The post-test could have been given immediately following the intervention (at the end of Term 1), but this would have added another test to an already busy schedule. Consequently, they decided to give the post-test mid-way through the second term; if the intervention had an effect that lasted, it would still be detectable. Consequently, the intervention was implemented as planned in Term 1, but the post-test given just before half-term in Term 2, with all students returning to the same planned curriculum for the duration of this term.

By designing their evaluation in this manner, the teachers were able to mitigate the effects of an additional test in an already busy schedule, while still conducting a well-designed, effective evaluation.

Had the teachers not been concerned about the timing of the post-test, they could have delivered it immediately following the intervention's completion, and then followed it up later with another test. This would have given them an indication of the short-term and longer-term effects.

Finally, when conducting a post-test which involves some teacher judgement you should consider to what extent the assessment outcomes could be influenced by the expectations or desires of the assessor and think about how to ensure the results are 'blinded' (see Box 8).

Information Box 8: Assessment and blinding

If the member of staff who is doing the assessment knows whether pupils are in the intervention or control group bias can creep in. This is not about cheating or lying – the bias is subconscious and inevitable however honest we think we are being. There are two main ways to address this:

- **Eliminate any judgement from the assessment:** Objectively marked items (such as multiple choice tests) do not require judgement, so are less likely to be prone to any expectation effects. You could buy in online or digital tests that include automated marking. However this is not always possible.

- **Ensure the testing is done 'blind':** Where any judgement is involved the outcome assessment should be done by a member of staff who does not know which pupils are which, or the papers should be marked anonymously. This is sometimes called *blinding*. For example, if essays or short answers are to be marked, candidates' names should be concealed. If a teacher might recognise handwriting or other features, then someone who does not know those students should mark it. If different classes (or schools) received different interventions then their scripts should be mixed up before marking.

This may seem like a lot of trouble to go to and in some cases may not be possible. However, there is a lot of evidence that un-blinded judgement-based assessments are biased, often substantially so.

3. Analysis & Reporting

3.1 Analysing and interpreting results

Analysing results

Once you have completed your intervention and testing you should put all of your data into an Excel spreadsheet with columns for the post-test data and a row for each pupil, then calculate an 'effect size'[4] for your intervention.

DIY Evaluation Glossary: Effect size

Effect sizes are quantitative measures of the *size and consistency* of the impact on an outcome, in this case attainment. It is calculated by taking the average difference between two scores and dividing it by the variation in that difference (see Box 9 for a step-by-step guide).

$$\text{Effect size} = \frac{\text{Difference between average post-test scores}}{\text{The variation in that difference as a standard deviation}}$$

Effect sizes can be approximately translated into additional months' progress you might expect pupils to make as a result of a particular approach being used in school, taking average pupil progress over a year is as a benchmark. The progress that an average pupil in a year group of 100 students makes over a year is equivalent to them moving up from 50th place to 16th place, if all the other students had not made any progress.

The conversion we have used corresponds to progress in Key Stage 1. Of course, a typical month's progress at primary school is greater than at the end of secondary school and so as a result this conversion may understate some impacts at secondary level. However, the conversion still gives an indication as to the relative effectiveness of interventions (e.g. it will always show which of two interventions are more effective, even if the months progress conversion is slightly conservative).

As well as calculating the average effect on attainment, you might also want to consider seeing if there are any differences in the effect for different subgroups, such as boys and girls.

Only the post-test results are used in determining the effect size of an intervention, as we are interested in the difference created by the intervention.

[4] A guide to effect sizes can be found here: http://www.cemcentre.org/evidence-based-education/effect-size-resources.

Information Box 9: Step-by-step guide to calculating an effect size

The following is a step by step guide to calculating a simple effect size using two groups (a control group and a treatment group), from their scores on the post-test:

1) Input the scores into a spreadsheet. For example, if there are 30 matched pairs, put the control group scores in cells A1:A30 and the treatment group in B1:B30.

2) Calculate the average post-test score for pupils in the control group. For example, in Excel, type into cell D1 "=AVERAGE(A1:A30)" and press return.

3) Calculate the average post-test score for pupils in the treatment group. Type into cell E1 "=AVERAGE(B1:B30)".

4) Calculate the difference between the two averages (i.e. treatment group average – control group average). In Excel, type into cell F1 "=E1-D1" .

5) Calculate the standard deviation of scores for the control group. In Excel, type in cell D2 "=STDEV(A1:A30)". Alternatively, you can use a pocket calculator or online calculator (e.g. http://easycalculation.com/statistics/standard-deviation.php) for this.

6) Divide the difference in average post-test score - as calculated in (4) - by the standard deviation of the control group- as calculated in (5). In Excel, type into cell F2 "=F1/D2" and press return The figure shown is the effect size.

Table 6: How big are effect sizes?

Months' progress	Effect Size from …	... to	Description
0	-0.01	0.01	Very low or no effect
1	0.02	0.09	Low
2	0.10	0.18	Low
3	0.19	0.26	Moderate
4	0.27	0.35	Moderate
5	0.36	0.44	Moderate
6	0.45	0.52	High
7	0.53	0.61	High
8	0.62	0.69	High
9	0.70	0.78	Very high
10	0.79	0.87	Very high
11	0.88	0.95	Very high
12	0.96	>1.0	Very high

Interpreting results

If you use random allocation and have implemented your evaluation exactly as planned the only difference between the control and treatment groups should be the intervention. Unfortunately, however, this is not always the case and there may be a number of reasons why differences may have occurred. When interpreting your results you will need to consider the other factors that may have brought about the change (or lack of change) that you are seeing. When interpreting your results you should consider that the effect might be due to:

- **The intervention or approach that you are testing:** the effect on attainment may be a direct result of the intervention you are testing.
- **Systematic differences between the groups:** if you are not using random allocation there might be systematic differences between your groups that have brought about the effect. For example, one group of children might be taught by a different, better, teacher, or be in a different school where they are implementing additional interventions which might affect your results.
- **Problems with your evaluation methods:** there are a number of factors regarding your evaluation that might affect your results. You should think about all the steps above, and in particular whether there were any differences in the timing or delivery of your pre- and post-testing that might affect the results. For example, the intervention group test might be done at a different time of day or when more pupils were absent from school.

3.2 Reporting your results

Table 7: Reporting your results

Section	
Title page	• Research question • Authors
Context	• Where and when the study took place • What were the characteristics of the children involved and how were pupils selected to be included
Design	• Explanation of how the comparison group was established • Details of any random allocation or matching • Planned timing of the tests
Intervention	• Details of the intervention and how it was delivered • Details of what the control received if anything (whether 'business of usual' or anything different)
Outcome measure	• Explanation of the measures used to assess impact • How reliable and valid the measures are • How the data was collected • Whether there was any data missing (e.g. due to pupils not turning up to the assessments)
Results	• The effect on attainment • Any analysis of subgroups • Interpretation of the results
Conclusion	• Summary of the study's fit with existing evidence base • Implications and next steps

WHO PUT THE GUIDE TOGETHER?

The *DIY Evaluation Guide* has been produced by Stuart Kime and Professor Rob Coe of Durham University for the Education Endowment Foundation.

Stuart taught English in secondary schools for ten years before starting a full-time PhD in Education in 2011. Stuart's PhD focuses on the use of student evaluations of teaching in secondary schools.

Rob is Director of the Centre for Evaluation and Monitoring at Durham University, which is the largest educational research centre in a UK university. Prior to beginning an academic career Rob taught Mathematics in secondary schools and colleges.

The EEF is an independent grant-making charity dedicated to raising the attainment of disadvantaged pupils in English primary and secondary schools by building and sharing evidence of what is effective to improve learning. Founded by in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust, the EEF was set up with an initial £125m grant from the Department for Education. With investment and fundraising income, the EEF intends to award over £200m to support its aims over the next 15 years.

For more information about the EEF, visit:

www.educationendowmentfoundation.org.uk