EngineeringUK

Evaluation approaches and causality

Evaluation shared practice guide

Summary

STEM engagement activities can play an important role in inspiring young people to pursue STEM educational pathways and encouraging them to consider a career in engineering. Evaluating your activities is important for ensuring you are making the best use of your available resources and achieving the greatest possible impact.

In planning for your evaluation, you will need to make decisions about the type of data you will need to collect and the research methods you will need to use to enable you to measure performance against your programme's objectives. It is important to be aware that **the data and methods you choose will shape the insights you will be able to draw from your evaluation**. The following are aspects that you should consider when planning and designing your evaluation:

- **Certain methods are better able to demonstrate a causal effect than others**. Experimental designs, such as Randomised Control Trials, are the gold standard for establishing causality.
- However, **not every STEM outreach provider will want to strive for more complex evaluation designs or aim to demonstrate causality.**
- **The evaluation approach that is most appropriate for you will depend on a variety of factors**, including how your activity or programme is delivered, the resources and time you have available, as well as various ethical, legal and practical considerations.
- Regardless of the approach you decide to use, it is important to take the time to **regularly review how effective your evaluation is for your purposes and consider ways to embed learning into programme development.** You may want to involve young people or other key stakeholders to gain further insight in this process.

This guide provides an overview of various approaches to data collection and analysis that you might want to consider in your evaluation, including information on their suitability for demonstrating 'causality', or 'causal effect'. This guide is not intended to be prescriptive, nor does it provide an exhaustive list of evaluation approaches. The methods described below should be considered as examples that you may want to draw upon. Examples of evaluation methods in practice are included in a table at the end of this document.

Why evaluate your STEM outreach activities?

Evaluating your STEM outreach activity or programme can allow you to:

- Assess whether your activity meets its overall objectives;
- Identify strengths and weaknesses in your activity, including which aspects are most and least beneficial for participants;
- Inform decisions on how to improve the future delivery of your activity;
- Make informed decisions about the future allocation of resources;
- Demonstrate to funders and other stakeholders how effective you have been in engaging and inspiring young people.

A successful approach to evaluation involves a well thought-out design, tailored to suit the nature of the activity, and adopting good practice methodologies. During the planning stages, it will be important to explicitly define the purpose of the evaluation and clarify the main questions you will need it to answer; these will help shape your approach, since different methods are useful for different purposes. The Rainbow Framework, developed by Better Evaluation, can be a helpful tool to use in the planning stages of your evaluation (see further reading).

Beyond evaluating your STEM outreach activities, it is also important to consider ways to **embed learning from your evaluation** into your programme development. Reviewing your research findings can improve the future delivery of your STEM outreach activity. Additionally, periodically reflecting on and reviewing your evaluation approach can also be helpful to ensure it remains appropriate for your STEM outreach activity.

Further reading: Better Evaluation - 'Rainbow Framework'

What do we mean by 'causality' or 'causal effect'?

Often when evaluating STEM engagement activities, providers will want to determine the extent to which participation results in – or causes – a shift in young people's views, attitudes or behaviour. An example would be determing the extent to which particpating in a STEM careers fair or robotics coding programme causes young people to become more interested in pursuing a STEM career or to become better at coding. The aim of the evaluation, in this case, is to determine the **causal effect** of the activity on a particular desired outcome.

When analysing and interpreting your evaluation data, it is important to distinguish between **correlation and causation**. Just because two things are correlated – or associated – does not necessarily mean that one causes the other. Take sunglass and ice cream, for example – sales of both will increase in summer (their sales are therefore correlated) but it would be a mistake to assume that sunglasses sales cause an increase ice cream sales – or vice versa. Instead, both of these things are in fact being driven by a confounding influence: the presence of sun! Similarly, your evaluation could show that participation in your STEM engagement activity is positively correlated with an interest in STEM, but this relationship could be spurious – it may be driven by a 'confounder', it may simply be coincidental, or it could even imply reverse causation[1].

[1] Reverse causation refers to situations of inverted cause and effect – for example, an interest in STEM increases young people's likelihood of participating in a STEM engagement activity, rather than the other way around.

The influence of confounders

A STEM engagement provider might find that 80% of young people who took part in their activity scored highly in a maths test, compared to 50% of those who did not take part. They might then assume that the 'effect' of their activity is a 30 percentage point increase in maths ability. However, there may be other confounders that have an impact on the relationship between participation and maths ability. For example, participants could have been students from better performing schools. In this case, studying at a high-ability school affects both the likelihood of participation as well as the participants' ability in maths. A provider who wants to show that participating in their activity will improve young people's scores in maths test will need to take account of the confounding influence of schools in order to produce unbiased insights from their evaluation.

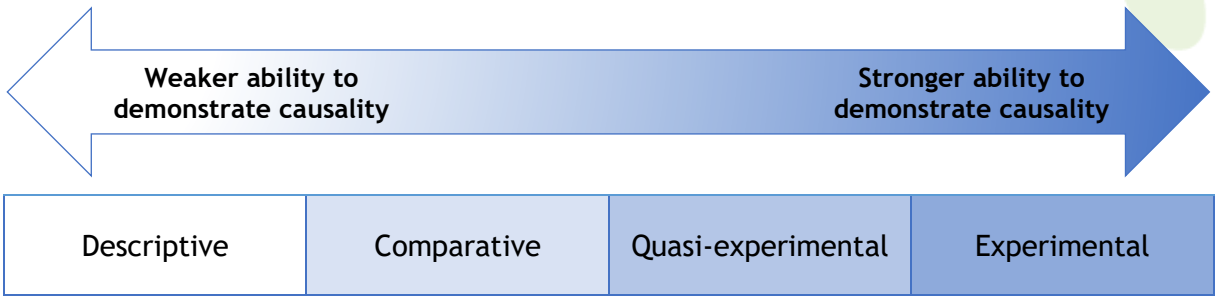
Further reading:
Australian Bureau of Statistics 'Statistical Language – correlation and causation'

Types of evaluation

When considering which evaluation approach to use, it is important to keep in mind that certain methods are better able to demonstrate causality than others. **Experimental designs are the gold standard for establishing causality**. In the natural sciences, experiments take place in a laboratory, where researchers are able to control for, or hold constant, all other potential confounders; that way they can be sure that any effect identified can be attributed solely to the 'treatment'.

In social research, outside of the laboratory setting, it is rarely feasible to adopt an experimental design. However, it is possible to attempt to 'mimic' some of the properties of an experiment, as well as to attempt to minimise the influence of confounders, as the next best thing.

The sections below provide light-touch summaries of different types of evaluation, ranging from examples of basic, descriptive approaches, to more complex, experimental designs. As a general rule of thumb, the closer you closer you get to an experimental design, the better able you will be to demonstrate causality. However, it is important to note that not every STEM outreach provider will want to strive for experimental or quasi-experimental evaluation designs. In fact, the evaluation approach that is most appropriate for you will depend on a variety of factors. These include, but are not limited to, the nature of your activity (e.g. the way in which your activity or programme is delivered) and the resources you have available (e.g. funding, time, staff), as well as a host of ethical, legal and practical considerations.



As you design your evaluation with your goals and objectives in mind, it is also key to assess the skills and expertise within your team. Note that the more complex methods of analysis mentioned in this guide may require statistical methods training of staff or perhaps necessitate outsourcing the evaluation to an experienced research agency.

Descriptive

A very simple and straightforward approach to the evaluation of your STEM outreach activity could involve administering a questionnaire or feedback form either at the event or soon after it takes place. Below are some examples of topic areas you may wish to survey young people on.

To understand what young people feel they gained from the experience, you could ask:

- Did they enjoy taking part in your STEM outreach activity?
- Do they feel that the event was too long, too short, or about the right amount of time?
- How likely would they be to recommend the activity to friends?
- Which aspects of the activity did they enjoy the most?

To measure how your activity shaped young people's views of STEM, you could ask:

- Do they feel that the activity inspired them to work in science or engineering in the future?
- Did taking part make them feel that a career in science or engineering would be interesting?
- How much do they know about what people working in science, technology do?

The importance of contextual data

Beyond questions about the activity or programme, it is often helpful to also ask for some socio-demographic information as a part of the evaluation, such as your participants' age, gender, ethnicity, or social background. This would allow you to analyse the data collected to assess, for example, whether there are any differences in responses by gender (e.g. do boys report that they enjoy your activity more often than girls do?), or age (e.g. do younger participants engage with the topics covered in your activity as much as older participants do?).

Collecting and analysing demographic information can be helpful for all evaluation methods mentioned in this guide. In doing so however, it is important you consult with your organisation's Data Protection Officer and/or seek the necessary legal advice to ensure your approach is GDPR compliant.

Producing statistical summaries of the data generated from these questionnaires can help you to understand how successful your activity has been. Examples of common **descriptive statistics** used for evaluations include measures of frequency (e.g. '80% of young people said they enjoyed the activity') and measures of central tendency (e.g. 'the mean, or average, age of young people taking part in the activity was 14').

Data of this kind cannot, however, facilitate a causal interpretation. For example, we could interpret the results as '70% of young people who attended the activity said they would be interested in a career in engineering', but we could not say '70% of young people said they would be interested in a career in engineering due to taking part in the activity'. In other words, with this approach we can describe patterns in the data, but we cannot explain why they have occurred.

Further reading:
EngineeringUK – 'Evaluating using surveys'
EngineeringUK – 'Evaluations with young people'

Comparative

A next step closer to being able to demonstrate the causal effect of your activity could be to adopt a **comparative approach**. A **pre-test/post-test evaluation design,** for example, involves asking young people a set of the same questions both before they take part in your activity (pre-test) and after they complete it (post-test). Results from the pre-test survey can then be used as a baseline against which you can compare the findings from the post-test survey.

A pre-test/post-test design

Before taking part in your STEM careers fair, you might find that just 50% of young people answer 'yes' when asked about whether they would consider a career in engineering. After taking part, you might find that this proportion has increased, with 70% of the same group of pupils responding 'yes' when asked the same question. Using this method limits the possibility of potential confounders biasing your results. For example, since you would be asking the same set of young people before and after your activity, you would not have to worry about differences in the average performance of schools between them. However, a number of unmeasured confounders could still be contributing to the increase in young people considering a career in engineering.

Another comparative approach you might want to consider involves comparing responses to the same set of questions between young people that do and do not participate in your STEM outreach activity. To use this approach, you will need access to results from a group of non-attendees who have been asked the same set of questions as your evaluation sample (i.e. young people who participate in your STEM outreach). It is also important that the sample of non-attendees is representative of the population of interest, and is comparable in some key respects to the sample of attendees. For example, if your evaluation sample only consists of young people aged 11 to 16 in the UK, you would want to compare responses with a sample of young people of similar age in the UK. This can be a helpful way to view how responses from your participants are different to those from the general population.

Further reading:
Harvard University 'Tip sheet on question wording'
Harvard University 'Overview of cognitive testing and questionnaire evaluation'

Quasi-experimental

A further step towards being able to establish causality would be to adopt a quasi-experimental evaluation approach. These methods attempt to 'mimic' certain properties of an experimental design.

Quasi-experimental designs rely on researchers being able to identify or construct **'treatment'** groups (e.g. young people that take part in your STEM engagement activity) and **'control'** groups (e.g. young people that do not take part in your STEM engagement activity) – with the two being as similar as possible with respect to a pre-specified set of characteristics. Data from the treatment group tell us about the outcomes for those that took part in the activity, while data from the control group tell us about what the outcomes hypothetically would have been, had the activity not taken place (known as the 'counterfactual').

Quasi-experimental approaches, as opposed to experimental approaches, are appropriate when it is not possible to randomly assign participants to treatment and control groups before the activity takes place. Examples of quasi-experimental approaches include:

- **Propensity score matching (PSM)** – This approach uses data from treatment and control groups, including information related to your outcome of interest (e.g. knowledge of engineering) and information on young people's individual characteristics (e.g. age, gender, ethnicity). After your STEM outreach activity has taken place, PSM would be used to 'match' individuals with similar characteristics

(so that like-with-like comparisons could be drawn) and then calculate the average difference in the outcome of interest between the treatment and control groups. Any difference identified (usually referred to as a 'treatment effect') could then be attributed to participation in your activity.

- **Difference-in-difference designs (DiD) (or double difference method)** – This approach uses data from treatment and control groups, including information on young people's individual characteristics and information related to your outcome of interest (e.g. knowledge of engineering), before and after the activity takes place. You might expect that knowledge of engineering increased among both the treatment and control groups due to young people's participation in other, unrelated activities, such as participation in their school's STEM club. But we can still look at whether this increase (i.e. the difference in pre- and post-test knowledge) is larger among those that participated in your activity than those who didn't; this way, you can disentangle the effect of your activity from the effect of attending the school's STEM club.

In general, DiD methods provide more reliable estimates of impact than matching techniques that involve comparisons of between treatment and control groups at just a single point in time, like PSM. However, DiD methods also tend to present more logistical challenges, since data collection is required at multiple time points and it is usually necessary to link young people's pre- and post-test responses (unless analysis is done at an aggregate level – differences between schools, for example).

Sampling

As selecting a control group is a key aspect of both quasi-experimental and experimental research, a solid understanding of the necessary **sampling process is fundamental** to appropriately carry out these methods. The decisions you make related to sampling for your evaluation will have implications on the extent to which you will be able to interpret your findings as being generalisable to a wider population. Sampling theory is beyond the scope of this guide, however, the resources included as further reading can offer an introduction to these concepts. Regardless, it is important to consult an expert to ensure the design, implementation, analysis and interpretation of the evaluation data and results are appropriate and correct for your evaluation.

Further reading:
UNICEF 'Quasi-experimental design and methods'
European Commission 'Quasi-experimental methods: propensity score matching and difference in differences'
Better Evaluation 'Sample'

Experimental

Using experimental designs in evaluation allows providers to reliably demonstrate the causal impact of their STEM outreach activities. Experimental methods can help to eliminate some common forms of bias which might lead to an over- or under-estimation of your activity's effectiveness.

Randomised Control Trials (RCT), as mentioned above, tend to be considered the 'gold standard' in experimental evaluation. This approach consists of randomly allocating

individuals to treatment and control groups and measuring their outcomes both before and after the activity takes place. As with quasi-experimental evaluation, the control group is used to represent the counterfactual. However, as the allocation in an RCT is random, both known and unknown factors that could affect the outcome can be controlled for, thereby limiting bias, and ensuring that the differences between groups can be considered as robust, valid estimates.

The random allocation to treatment and control groups can be done either at an individual level (e.g. pupils) or at a cluster level (e.g. schools). In case of the latter, this approach is called **cluster randomised control trial**. You may want to consider using this approach in contexts where there is a risk of potential 'contamination' across the treatment and control groups. For example, if conducting an RCT with young people who all attend the same school, there may be a risk that pupils in the treatment group mix or talk with pupils from the control group, possibly sharing information related to the treatment (i.e. your activity). In this case, the control group can no longer be used as a good comparison for the evaluation, as these young people may have been exposed to some extent to the treatment. The random allocation of treatment and control groups at a school level could minimise the risk of 'contamination' across groups. You might also consider a cluster randomised control trial if your STEM outreach activity is implemented at a school level. In this case, randomising by individual pupil may not be logistically feasible for you, and the unit you may be able to randomise instead is the school.

While RCTs are considered to be the gold standard in demonstrating the causal effect of a treatment or intervention, they are not without limitations; in particular, they can be extremely resource intensive and they usually require experienced monitoring and evaluation staff to be involved in all stages of planning, execution and analysis. As mentioned, an RCT may not be feasible or appropriate for all providers and may not be suited to all activity types.

Reviewing the evaluation approach

Regardless of the evaluation approach you decide to use, it is important to take the time to regularly review how effective it is for your purposes. You might find that changes in resourcing or in the delivery of your activity should prompt a shift in the evaluation design, for example.

Involving young people or other key stakeholders can be beneficial in reviewing and reflecting on your approach. For example, you may test your survey questions with young people to assess whether the language you use is clear and easily understood by your participants. In terms of your evaluation findings, it is good practice to share these with young people who took part in the evaluation. You may also consider seeking feedback on your approach with subject experts or discussing your findings with peers (e.g. staff at your organisation or other STEM outreach providers).

Further reading:
Wharrad H. and Silcocks P. 'An introduction to experimental design'
Better Evaluation 'Randomised Controlled Trial'
HM Treasury 'Magenta book: central government guidance on evaluation' – see Table 2.3 Choosing an experimental or quasi-experimental approach to impact evaluation

Table 1. Examples of evaluation methods in practice

Descriptive	Comparative	Quasi-experimental	Experimental
The British Science Association carried out an evaluation of their British Science Festival (BSF), a national science event aimed at creating opportunities for people to enjoy science and connect with researchers from various scientific fields. Data for this evaluation was collected on overall attendance and through an Audience Questionnaire disseminated to attendees. This form included questions asking participants for feedback on the events they took part in, as well as demographic questions. Overall, a total of 104 events were evaluated, and 3261 feedback forms from event attendees were received. Key findings include: • 15,260 was the total attendance • 92% of attendees rated events as either excellent or good • 48% of attendees were between the ages of 16 and 34	EngineeringUK carries out an annual survey of young people, their parents and STEM secondary school teachers which asks about knowledge, perceptions and desirability of STEM education and careers – the Engineering Brand Monitor (EBM). There are several questions contained in the EBM which use identical wording as the evaluation questionnaires administered to attendees of The Big Bang UK Fair, Robotics Challenge and Energy Quest activities. This allows EngineeringUK to compare, for example, the level of knowledge about engineering careers expressed by attendees of these activities with the level of knowledge in the general population of young people the same age. Over recent years, comparative analyses using EBM and evaluation data have consistently shown that young people taking part in these STEM engagement activities have more positive perceptions and knowledge of engineering than their counterparts across the UK who have not taken part in these activities.	The Department for Digital, Culture, Media and Sport commissioned an evaluation of the National Citizen Service (NCS), an initiative aimed encouraging personal and social development among 16 to 17 year olds from different backgrounds. In 2018, over 100,000 young people participated in the NCS. Data was collected through a baseline survey (at the start of the programme) and a follow-up survey (three months after) with NCS participants and non-participants. PSM was used to match NCS participants to non-participants whose profiles were as closely aligned as possible. This method allowed researchers to control for factors (e.g. demographic data, differences in attitudes or behaviours prior to NCS) that might affect responses, and to more confidently attribute differences between groups to NCS participation. DiD analysis was then used to calculate differences in outcomes between NCS participants and non-participants, based on data both from before and after the initiative took place. Researchers found that the programme had a positive impact on measures related to social mobility and social engagement. However, results were mixed in terms of the impact the programme had on participants' wellbeing, loneliness, and overall social cohesion.	Education and Employers conducted an RCT pilot study to understand how employer encounters can change young people's attitudes and improve their educational attainment. Around 650 Year 11 pupils from five schools across England took part in the study. Using a random allocation, schools divided young people in tutor groups into either a treatment or a control group. The 307 pupils in the treatment group took part in three additional career talks, beyond the typical career activities provided by their school. Year 11 students who took part in the study (i.e. in both the treatment and control groups) were asked to complete a baseline survey (at the beginning of the academic year) as well as another survey at the end of the year. Data on these pupils' predicted GCSE grades were collected, and later compared with their actual exam results. Researchers found that, compared to the control group, students who took part in the extra career talks showed improvements on their self-efficacy, attitudes towards school being useful and confidence in pursuing career aspirations. Findings also showed that pupils who were less engaged and lower achievers benefitted the most from the treatment.
British Science Association 'British Science Festival evaluation report',2018.	EngineeringUK 'Engineering Brand Monitor', 2019.	The Department for Digital, Culture, Media and Sport 'National Citizen Service evaluation report', 2020.	Education and Employers 'Motivated to achieve', 2019.