RCTs and QEDs: Reflection note from shared learning workshop

Background

This note draws on learning from a capacity-building project in which NatCen worked with several organisations that run programmes aimed at helping young people to enter the workplace. This reflection note was created following a workshop, which was part of a suite of learning outputs to help support these organisations.

This note is intended to provide a starting point for thinking about experimental and quasi-experimental methods and their use in evaluations. It includes some considerations, hints and tips built up from our experience of using these methods within evaluation work (https://natcen.ac.uk/centres/centre-evaluation). We have provided references at the end for looking at this topic in more depth.

Measuring impact

A key aim of impact evaluation is to measure the effect of a programme on an outcome. For example, what is the effect of attending a maths workshop on a young person's functional skills maths score (FSQ)?

In this example, you may observe that many people who attend the maths workshop go on to pass FSQ maths:

However, there could be many different explanations for people going on to pass maths – not just attending the workshop.

Some of these alternative explanations may be measurable or "observable" to the evaluation team, for example, demographic variables (age, gender, social class, ethnicity, highest education level), or factors such as prior maths attainment or interest in maths.

There will often also be "unobservable" factors. They may be unobservable because they are more difficult to measure (e.g. parental involvement) or because data about them do not exist (at least not data that you can use in your evaluation).

These alternative explanations or factors are known as **confounders** if they potentially influence both intervention *participation* and *outcomes*. Common confounders include socioeconomic status, intrinsic motivation, age, ethnicity, risk of

experiencing crime, NEET status (not being in employment, education or training) and prior academic attainment.

Revisiting our example, the confounder of 'prior ability in maths' might affect how likely the person is to attend the workshop. It also shows how the confounder of prior ability might affect how likely the person is to do another activity such as "practice sums". This in turn will affect whether they pass FSQ maths (the outcome) regardless of whether they attended the workshop or not.
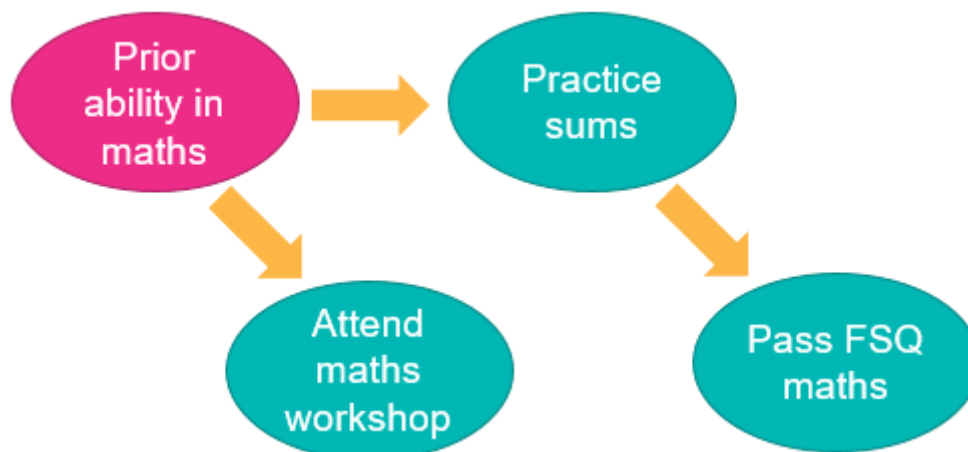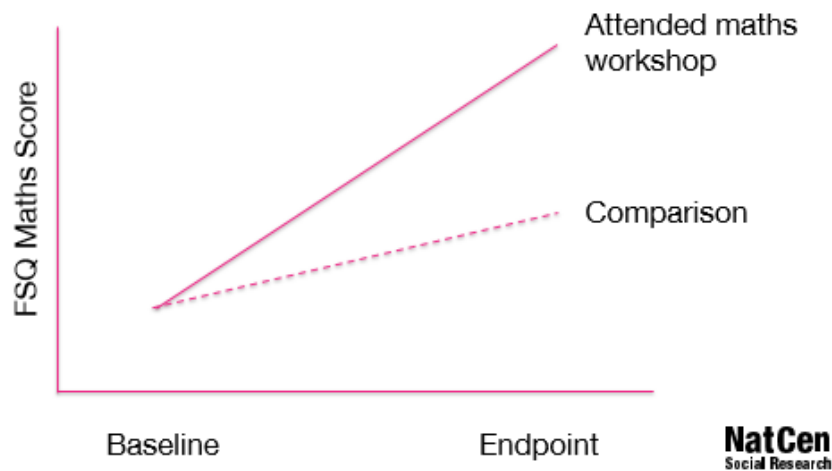

Another way to understand whether attending the maths workshop helped young people pass FSQ maths would be to look at their maths scores before and after attending the workshop. However, if you take a group of people with relatively extreme levels on an outcome measure (e.g. low maths scores), it is common for them to return (or "regress") towards average (or mean) levels over time, regardless of an intervention's impact. This is known as **regression to the mean** and can lead to natural variation in the data being misinterpreted as impact due to a programme.

In the context of evaluations, this means that people who appear to benefit from support from the programme…

…may have improved anyway

…may have improved for reasons other than the support.

A **comparison group** can help you to estimate what would have happened in the absence of the programme.

Figure showing FSQ Maths Score against time (Baseline to Endpoint) for two groups: "Attended maths workshop" (solid line) and "Comparison" (dashed line). NatCen Social Research.

Randomised Controlled Trials and Quasi-Experimental Designs

Not all comparison groups are equally useful and choosing the most appropriate comparison group greatly enhances the strength on the causal claims you can make. This section introduces two common methods for estimating the impact of an intervention using a comparison group:

- Quasi-experimental designs (QEDs)
- Randomised Controlled Trials (RCTs)

QEDs use statistical methods to attempt to construct a comparison group based on what can be measured. This comparison group must comprise people who:

a) Have not received the intervention.

b) Are as similar as possible to the people who have received the intervention.

The advantage of QEDs is that they do not require you to change intervention delivery, but they do require lots of data on relevant characteristics. Identifying data sources and relevant characteristics can be challenging and often requires collaboration with partner organisations. Examples of data sources that could be used for QEDs include administrative data such as the National Pupil Database (for schools-based programmes) or the Longitudinal Education Outcomes (LEO) dataset which connects individuals' education data with their employment, benefits and earnings data.

RCTs, conversely, attempt to control for what we do not (and cannot) know/measure by randomly assigning people to different conditions. This usually means some modification to the intervention. The comparison group(s) might be randomised to:

- Waiting lists – the comparison group will receive the treatment at a later date, after outcomes have been measured across the two groups.

- Other support – the comparison group can be offered different types of support in the absence of the programme.
- "Business as usual" – the comparison group is not offered any programmes or services beyond what would usually be available.

Three common kinds of RCTs include:

a) **Individually randomised** – one study site, individuals randomised either the programme group or the comparison group

b) **Clustered randomised** – multiple sites, sites randomised to either the programme group or the comparison group

c) **Multisite RCTs** – multiple sites, individual randomisation

Random allocation to groups means the statistical analysis can even take account of unknown confounders (for example, intrinsic motivation). However, this often means making quite substantial changes to the way the programme is delivered. For example, you will need to randomly allocate participants to either take part in the programme or not, where you didn't have to do this before. Alternatively, you may need to create a waiting list for people allocated to the comparison group to enrol on the programme later.

It is the evaluator's role to determine, in consultation with other key stakeholders such as the client funding the evaluation, the developer of the programme, and the delivery organisations, the specific RCT/QED design.

What to evaluate

It is uncommon for an RCT/QED to evaluate an organisation's entire programme/offering. Instead, they might evaluate:

- A specific programme or element of the programme
- An innovation in delivery
- Elements of communication, such as evaluating how effective your recruitment materials and processes are

Whatever is being evaluated, organisations need to be willing to receive negative/non-significant results. If it was certain that an RCT would lead to the desired outcome, then there would be no point in running an RCT. The implications of such negative or null findings are obviously greater if it is an evaluation of the whole programme.

Sample size

There is no one-size-fits-all sample size for an RCT/QED. Studies must be 'powered' to detect effects, generally speaking, this power comes from the size of the sample. The sample size necessary for a sufficiently powered evaluation depends on:

- How many in the comparison group are similar enough to those in the intervention group at baseline

- Expected attrition (dropout/loss to follow-up)
- The size of effect you want a good chance of detecting – a smaller true effect means more participants are needed to spot it

Ethical practice

There are often ethical concerns regarding evaluations, especially RCTs. RCTs can be seen as randomly withholding an intervention or service from young people who are in need. An RCT might be ethically justified in a number of circumstances, such as when:

- **You are not taking something away but only have limited resources to allocate** – e.g. new funding could offer laptops and internet access for 50% of young people, you may argue it is fair to allocate these randomly.
- **You have a waitlist** – you are not withholding a service from those randomly allocated to the comparison group, they will receive the service later.
- **You are adding something to the existing service** – you could allow all young people to access your standard service, then randomly allocate some to receive an extra service.
- **You have equipoise** – you do not know that a certain approach is better than the other (you wouldn't randomise parachutes, but researchers have randomised parachutists to different ankle supports!)

This is less of a concern for a QED, where the comparison group would not receive the intervention, regardless. A QED does not change who received an intervention.

Readiness for an RCT/QED

In most cases, an impact evaluation (e.g. a QED or RCT) will be carried out by an independent evaluator. A What Works Centre such as YFF can pair up organisations with evaluators.

An evaluation will often include the following steps:

1. **Design**
   a. Develop/confirm Theory of Change
   b. Agree any modifications to programme delivery

2. **Data collection**
   a. Primary data and/or administrative data collected/requested
   b. Qualitative data collection alongside quantitative

3. **Analysis**
   a. Analysis as per published plan – estimates added-value
   b. Triangulate quantitative and qualitative findings

4. **Reporting**
   a. Findings published regardless of what they show

b. Will be given a quality rating, e.g. based on attrition

While an independent evaluator would typically carry out the evaluation, there are some things that can help you be ready for an RCT/QED:

- If needed, undertake further reading or training so that you understand what an RCT or QED involves and how these methods can provide robust estimates of the effect of a programme on a particular outcome.

- Have a clearly defined programme and Theory of Change

- Define the outcomes that you expect your programme to have an effect on

- Clearly define the eligibility criteria for your programme

- Collect monitoring data on your participants

- Look to see if there are any differences in outcomes before or after the programme. For the reasons listed in this note, this will not show the impact of the programme, but it may show some indicative evidence of promise.

Please refer to our note on 'Preparing for Evaluation' for more ideas on how to prepare for and get the most out of an evaluation.

Useful resources

The Magenta Book (HM Treasury guidance on impact evaluation): https://www.gov.uk/government/publications/the-magenta-book

Better Evaluation (website with evaluation resources): https://www.betterevaluation.org/

What Works Network (website providing information on the What Works centres): https://www.whatworksnetwork.org.uk/what-works-centres/

UNICEF Overview of Impact Evaluation (website and videos on evaluation): https://www.unicef-irc.org/KM/IE/impact_1.php

Causal Inference: The Mixtape (more advanced free online textbook on methods for establishing causality in the social sciences, with Stata and R examples): https://mixtape.scunning.com/