Randomised controlled trials – gold standard or fool's gold? The role of experimental methods in voluntary sector impact assessment

Sally Cupitt, Head of NCVO Charities Evaluation Services, May 2015

1. Introduction

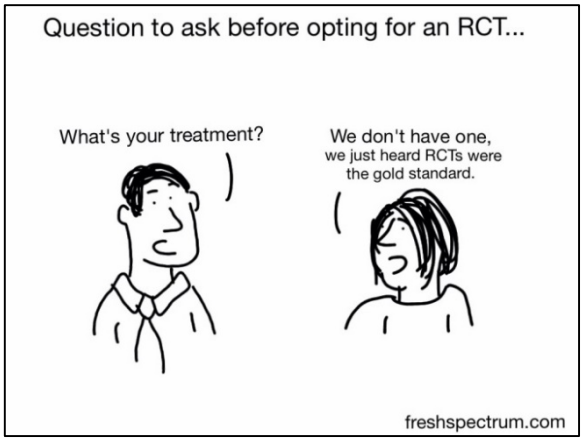
Demonstrating impact is a thorny subject for voluntary and community sector organisations. How can we show that our intended changes have occurred, and that these can be directly linked to our own interventions?

In medical settings, randomised controlled trials (RCTs) are a common way of investigating whether or not a treatment works. RCTs have also been used to evaluate the impact of educational or social care interventions, looking at everything from microfinance schemes to the effect of uniforms on school attendance.[1] This article discusses the use of RCTs in a voluntary sector setting.

Our sector has come a long way in relation to outcomes and impact assessment. Twenty-five years ago, Charities Evaluation Services' evaluation consultants frequently met people who simply saw no benefit in evaluation at all, let alone assessing change. That argument is now (mostly) won, and the debate has moved on to issues of raising standards of assessment and evidence, and appropriate methodologies for doing so.

Although some charities are still making first practical steps towards an outcomes and impact focus, other voluntary sector organisations have really flown, developing sophisticated and powerful approaches. They are also becoming more receptive to technical approaches like Social Return on Investment and RCTs.

The main advantage of RCTs is that they allow you to investigate the effect of an intervention while eliminating some common forms of bias. They're often described as the 'gold standard' of scientific research. However, RCTs aren't without their problems, especially when they're used to test complex interventions



Reproduced with permission of freshspectrum.com

[1] Goldacre (2011) 'How can you tell if a policy is working? Run a trial' www.theguardian.com/commentisfree/2011/may/14/bad-science-ben-goldacre-randomised-trials

in complex social situations.[2] This means there are some particular challenges when designing RCTs for use in the voluntary sector, which we'll discuss in this article.

2. What is a randomised controlled trial?

2.1. Basic concepts

To help explain RCTs, it's useful to understand some basic concepts first:

- **Impact**. 'Impact' can be defined in a range of ways, but in the context of RCTs, impact is usually defined as the change that has occurred minus what would have happened anyway, without the intervention. The change can include harmful effects as well as positive ones.
- **Intervention**. By this we mean the thing being tested. In medical trials it might be a pill or injection, but it could be anything – for example training or therapy.
- **Bias**. Bias is a mistake made in the way research is carried out that affects the strength of its findings. Bias can cause researchers to over- or underestimate the effects of an intervention. For example, if outcomes data is only collected using an online survey, people without access to the internet are excluded. Internet users may be different in lots of ways to non-internet users, and these differences would lead to bias in the study's findings.
- **Control group**. The 'controlled' part of a randomised controlled trial means that people who get the intervention are compared with a second group, called the control group, who don't get the intervention. Designing a good RCT means minimising differences between the intervention group and the control group. Randomisation is the main tool used to achieve this in an RCT.
- **Randomisation.** Random allocation is where people are chosen at random to either get the intervention, or not. Imagine an experiment where people from two towns were given a different intervention. Any differences between people from the two towns could lead to bias in the results. For example, if people from one town were older on average, it might be difficult to separate the effects of age from the effects of the interventions being studied. However, by randomly allocating people from both towns to the two interventions, we would average out age differences – and any others – allowing us to be more confident that any effect we observed had happened because of our intervention.

[2] See Concato, Shah and Horwitz (2000), 'Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs'. *New England Journal of Medicine*; Osrin et al. Ethical challenges in cluster randomized controlled trials: experiences from public health interventions in Africa and Asia. *Bulletin of the World Health Organization* 2009;87:772-779. doi: 10.2471/BLT.08.051060

2.2. How an RCT works

RCTs are a type of experimental research. Initially used primarily in the medical field, more recently RCTs have grown in popularity as a way of testing social programmes, as a form of impact evaluation.

RCTs work by comparing the outcomes for people who received an intervention with those who did not. Done well, RCTs can provide strong evidence of whether an intervention caused an outcome or not. However, it requires some technical skill to carry out an RCT, and we do not set out to describe one in full here – rather we give a flavour of the key aspects of an RCT.

At the outset, the intervention in question needs to be described, and often a theory of change is developed. Having then identified the samples, people (or sometimes groups) are usually randomly allocated (often by an unbiased third party) to one of two groups:

1. The intervention group (those who get the intervention; sometimes called the treatment group) and
2. The control group (those who don't – sometimes called the non-treatment group).

An RCT helps to construct what's called a 'counterfactual' – this assesses what would have happened to clients in the absence of any intervention. Results in the intervention group can then be compared with those in the control group and the difference examined.

The two groups must be constructed to be as similar as possible, so that the only difference between them is the intervention. Then, any difference between them should be due to the intervention itself. The fact that they are randomly allocated is a central part of the methodology; this helps reduce the likelihood of difference between the two groups, and therefore reduces potential bias.

The randomisation can be done 'blind' to further reduce potential bias. This is where participants and practitioners do not know who is allocated to which group. Clearly this is easier in a medical setting, where placebo drugs can be given instead of the real drugs, than in a social intervention. If you're testing a training programme, for example, you can't realistically hide it from participants whether they receive the training or not. In this case blinding is more likely to apply to the people doing the data analysis – they may not know who received what intervention.

After the intervention, the outcomes are measured in each group; often they are measured pre-intervention too, to give a baseline. While you can do RCTs without a baseline, for complex social programmes baselines are increasingly important, not least as they make subgroup analysis easier – you can assess change against the different start points of various subgroups.

Often there will be some small differences between your intervention and control group, just by chance. You might hear a difference between groups described as 'statistically significant'. That means the difference is so large it's unlikely to have happened by change alone, and is therefore probably an effect of the intervention.

2.3. Quasi-experimental designs

If the evaluation doesn't use random allocation, but just compares groups selected non-randomly (for example, a school may compare results across two different classes) it is called 'quasi-experimental', and is usually considered less robust as a way of proving what caused a change. This is because the two groups are more likely to differ in important ways, for example the outcomes might be better for people who have chosen to receive the intervention as they are more motivated to change.

There are a number of different quasi-experimental designs. Evaluators can choose a comparison group which resembles as closely as possible the group receiving the intervention, such as another group of young people going through a programme, of the same age group and gender distribution. Sometimes, statistical procedures can be used to adjust the comparison group to remove some of the differences and make it more like the intervention group. One of the most familiar quasi-experimental designs is a pre-post study, where data on outcomes is collected from clients at the beginning and end of the intervention, and the results compared. While a very useful form of data collection, this is considered one of the least robust quasi-experimental methodologies. It can be strengthened if data is collected at a number of points throughout the intervention.

3. Why are RCTs today's news?

When evaluation was first developed in the UK it initially followed quite a scientific model, favouring measurement and objectivity. However, in the 1980s and 1990s, many evaluators favoured more qualitative approaches, focusing on description and valuing multiple perspectives. This approach fitted well with the ethos of the voluntary sector, and was well suited to developing a culture of self-evaluation as such methods required less technical expertise.

More recently, there has been a resurgence of interest in scientific and technical models of evaluation, including RCTs. There is some concern that methods used by voluntary sector organisations are not sufficiently robust to make causal links between outcomes and the intervention - that is, to show that changes resulted from their own activities and services.

Our colleagues in international development have also been exploring experimental methodologies as a way to deal with the problem of causality. For example the Poverty Action Lab provides examples on its website of a considerable number of

RCTs of development aid programmes, often extensive and significant, but evaluating fairly straightforward interventions, like offering mosquito nets to reduce malaria.[3] We found little readily available evidence that UK international aid charities were carrying out many RCTs.

The government holds evidence from RCTs as the ideal standard in impact evaluation,[4] for example for Social Impact Bonds; the government is also bringing together evidence from RCTs in a number of policy fields. In the third sector, the Big Lottery is funding a demonstration programme for youth offending projects evaluating with experimental methods. In London, Project Oracle, an 'evidence hub' for children and young people's projects, puts a high value on experimental methods as part of its evidence standards.

It is worth noting that, as yet, RCTs are still exceptional in the UK voluntary sector, and even within UK-based international aid charities. We found little readily-available evidence of many RCTs being carried out in the UK voluntary sector, and of those we found there had been mixed success. Echoing this, the recent Project Oracle Synthesis Study[5] found only eleven RCTs that had been undertaken in London with young people; the majority were carried out in schools and few were funded by the voluntary sector.

4. Challenges for the voluntary sector

RCTs are a powerful methodology, arguably offering one of the best ways to provide proof – or the nearest we can get to it – as to whether or not something worked. However, there are a number of challenges with the use of RCTs in the voluntary sector:

1. Scale and timescale
2. Technical skill
3. Ethical issues
4. Controlling for differences between the groups
5. The nature of the intervention
6. Generalisation
7. The need for other evaluation approaches alongside RCTs.

[3] The Abdul Latif Jameel Poverty Action Lab (J-PAL) was established as a research centre at the Economics Department at the Massachusetts Institute of Technology, and is now a global network of researchers who use randomised evaluations to test the effectiveness of development programmes and policies aimed at reducing poverty http://www.povertyactionlab.org/

[4] See HM Treasury, Green Book, providing guidance on how to run a value for money evaluation

[5] http://project-oracle.com/uploads/files/Project_Oracle_Synthesis_Study_5-2015_RCTs_HQ.pdf

4.1. Scale and timescale

Costs

RCTS, done well, are not cheap. It's not always clear that the value gained from them is worth the cost, especially when we remember that evaluation evidence is only one of many factors affecting how decisions are made about whether or not to fund interventions. Quasi-experimental methods are often cheaper and may provide sufficient information to make adequate decisions. Also, it is unlikely that RCTs could replace the internal self-evaluation that quality organisations already do – it would be an additional activity on top of that.

Large numbers

Crudely, larger samples give more accurate results, and can show more subtle effects. For many in the voluntary sector, the scale of the intervention – perhaps also constrained by budget – may mean there simply are not sufficient clients to make an RCT viable.

Timescale and timing

As with many impact evaluation methods, timescale may be an issue. You need to have sufficient time for the outcomes to have been achieved, and not all voluntary sector organisations have that within their funded period. Finding that sweet spot for outcomes data collection – not too long after that the intervention that you can no longer find people, and not too soon that outcomes haven't yet had time to appear – is a problem for any form of outcomes evaluation.

RCTs in particular may take some time to produce results, which runs counter to the current interest in real-time evaluation, where evaluation results are fed back to service deliverers on an ongoing basis, enabling them to make changes as soon as possible.

4.2. Technical skill

RCTs are complex – it is highly unlikely that any voluntary sector organisation could undertake an RCT without external help, which may have cost implications. For example, those running an RCT will need an expert statistician in their team.

As a very technical methodology, RCTs may mean there is less stakeholder participation in, and ownership of, evaluation, something we at CES have been promoting since our inception in 1990.

4.3. Ethical issues

A commonly mentioned problem with RCTs – especially within the value-driven voluntary sector – is an ethical one. RCTs may involve denying an intervention to a group of people who need it. They have to need it as much as the intervention group, or they would not be sufficiently similar. Front-line staff in particular sometimes find this difficult.

However, there are some ways round this, for example offering the control group the intervention at a later date. Another argument is that the benefit of the intervention is not yet known, so a benefit is not being denied.[6] It is worth noting that ethical concerns may be particularly strong in situations where people are in crisis or at risk, for example child abuse or domestic violence, although it is likely that both groups in the experiment will be receiving some other basic services.

4.4. Controlling for differences between the groups

A major difficulty with RCTs is trying to control the differences between the intervention and control groups. People are not usually in laboratories – the intervention being studied is likely to be one of many complex variables affecting their lives.

It's worth emphasising that managing the control group is as important as the intervention group. For example, when denied access to the intervention, some of the control group might try to access a different version of the intervention elsewhere. If they did, this could have a profound effect on the results.

Imagine an RCT looking at whether cookery classes for parents could help reduce childhood obesity. If some parents at a school got the classes and others didn't, they might compare notes at the school gates, and some of the information from the classes could reach the control group. That may be a good thing for the children's diets, but it would undermine the results of the experiment.

Drop outs

With most social programmes, participation is voluntary, which means people can drop out. It is possible that those who drop out are different from those who do not – perhaps they are less motivated to change, for example. This could un-randomise the samples.

While some RCTs only look at outcomes for those who complete the intervention, it is generally considered more robust to also consider outcomes for those who drop out, certainly if you want to use the results to justify replicating the intervention elsewhere.

4.5 The intervention itself

The nature of the intervention – the thing you want to evaluate – is relevant when choosing any evaluation design, including RCTs.

The nature of the intervention

Some types of work lend themselves more readily to an RCT-type approach. Simple, linear interventions, where change isn't affected by many other variables, are often more appealing as subjects for an RCT. By contrast, for a complex multi-layered

[6] This is also called the equipoise principle

initiative, with many actors involved, an evaluation that seeks to look at an intervention's contribution to change rather than its attribution might be preferable.

Controlling the intervention

Imagine we are running an RCT on an intervention that is being offered around the country in different places. What is offered must be exactly the same in each place if they are to be compared. Anyone who has been involved in a social programme over multiple sites knows how hard it is to ensure consistency. For example, practitioners often understandably want to tailor their work according to local need – this could be problematic within an RCT.

Reproduced with permission of freshspectrum.com

The stage of the intervention

When undertaking an RCT the intervention itself must be closely monitored to ensure it is delivered in an agreed, standard way. This means that the stage of the project being evaluated is a relevant consideration. If a project is very embryonic, and the intervention being evaluated has not become sufficiently robust in its delivery methods, it may simply be too early for an RCT.

4.6 Generalisation

A significant issue with RCTs is that it is hard to generalise from them, or from one single RCT study. They can provide excellent evidence that an intervention worked for a particular group of people in a particular context and, as such, they have power in arguing for the continuation of something being done already. They can be limited in terms of saying whether the same intervention would work in a different context with a different group of people. If you're interested in whether an intervention works in very different contexts, or for people who are very different, you might need to run more than one trial.

4.7 The need for other evaluation approaches alongside RCTs

Increasingly, to counter the difficulties with RCTs, researchers are combining RCTs with other forms of evaluation, for example process evaluation, especially for complex programmes.

The useful NFER *How to guide on RCTs*[7] argues that a process evaluation to check the fidelity of the evaluation is crucial before assumptions are made about what the results mean. If the way that the intervention is carried out is not the same, the groups will not be comparable.[8]

The use of programme theory – and having a well-developed theory of change – is also being used to strengthen RCTs and to limit some of their challenges.

5. Summary

RCTs can be powerful, but may not always be an appropriate methodology for assessing the impact of social programmes. And their practical application in the UK voluntary sector is, currently at least, limited. The resources and skills needed to do a quality RCT, combined with ethical issues and the need for additional context and process data to understand the results, can limit their relevance to many voluntary sector organisations. Other methods, while perhaps not being gold standard, may still produce very useful results, and it's important to remember that when choosing evaluation design it's about what methods are appropriate to the evaluation questions and the intervention being studied.[9]

The debate must continue as to the role of quasi-experimental methods in impact evaluation. Further, using qualitative methods in impact assessment to construct a counterfactual needs more exploration. More on this from CES soon, in our upcoming article on the use of qualitative approaches to assess impact.

[7] National Foundation for Educational Research (2014) *How to Run Randomised Controlled Trials (RCTs): An introduction*, Slough, NFER www.nfer.ac.uk
[8] See also Ann Oakley on the RIPPLE study - an RCT of a pupil peer-led sex education project that reported in 2006.
[9] See Stern (2012) *Broadening the range of designs and methods for impact evaluations: Report of a study commissioned by the Department for International Development.* DfID