RCTs are widespread in the medical arena. Here, the environment can be fully controlled.

Photo: WHO/Eduardo Soteras Jalil

RANDOMISED CONTROLLED TRIALS – THE GOLD STANDARD?

Although randomised controlled trials are seeing widespread use, they have also been in for some criticism. Our author shows some of the snags that the method may meet with and recommends that context and appropriateness be given more consideration in designing evaluations.

By Maren Duvendack

Randomised controlled trials (RCTs) have recently grown in popularity. The basic idea is simple. In a randomised study, individuals are randomly assigned to so-called treatment and control groups, whereby both groups must be drawn from individuals whom the programme has yet to serve, so that the impact of an entire programme can be evaluated.

This random assignment to either treatment or control groups ensures that potential outcomes are not contaminated by self-selection into treatment. Self-selection refers to individuals selecting themselves into participating in particular programmes, e.g. they may self-select into microfinance programmes because they are particularly entrepreneurial or have certain risk attitudes and/or business skills. If randomisation is successful, it is assumed that individuals in treatment and control groups are equivalent in terms of observable and unobservable characteristics, with the exception of the treatment status. As a result of this, the differences we observe in the outcomes of each of these individuals are understood to be the effect of the programme.

THE CRUCIAL ASPECT OF CAUSALITY

Hype surrounding RCTs has led policy-makers, funders and researchers to believe that randomisation is the only method that convincingly establishes causality. However, for RCTs to convincingly establish causality, they need to be implemented properly. In other words, we have to be convinced that individuals have been truly randomly allocated to treatment and control groups; only then will we have succeeded in constructing an accurate counterfactual scenario (i.e. what would have happened in the absence of a programme). At the same time, we must be able to check for self-selection bias without having to resort to sophisticated econometric techniques that require particular technical expertise.

RCTs may be an attractive methodological option but they are not free from challenges, which can be of technical, ethical and/or practical nature. In academic circles, the chorus of critical voices has become louder arguing that there are threats to the internal and external validity of RCTs. For instance, how much can we really trust the causal claims of RCTs, and how generalisable are their results to other situations and/or individuals? Let us now look at some of these threats before examining potential alternatives to RCTs.

THE CHALLENGES OF CONDUCTING SUCCESSFUL RCTS

Successfully implementing RCTs is not an easy task, mainly due to technical challenges such as ensuring double-blinding, avoiding pseudo-random methods, addressing attrition and considering behavioural changes caused by the experiment itself such as Hawthorne and John Henry effects which may affect the results in positive as well as negative ways (as explained below). Furthermore, spill-over effects cannot be fully ruled out, and ethical and practical challenges need to be considered. We will now investigate some of these challenges in more depth and start with the key feature of RCTs, which is double-blinding.

Evaluation expert Michael Scriven, among others, stresses that double-blinding is one of the prerequisites for a robust RCT. Double-blinding implies that individuals participating in the RCT and researchers executing the RCT do not know who is receiving a particular treatment or not. The rationale for striving to achieve double-blinding is to avoid biased research outcomes caused by the placebo effect. In the medical arena, where RCTs are well established, double-blinding can be ensured by running RCTs in laboratories, where the environment can be fully controlled, but the case is different for studies in the area of the social sciences and international development in particular. For example, RCTs evaluating the impact of education, social services or microfinance programmes are usually not even single-blinded but essentially 'zero-blinded'. In other words, individuals usually discover whether they belong to treatment or control groups, which undermines the notion of double-blindedness.

Another challenge is the prevalence of pseudo-random methods which often occurs during the process of assigning individuals to treatment and control groups. It pays to investigate how exactly individuals were assigned to their respective groups; was the underlying process truly random? For example, the evaluation of the Girl's Education Challenge in Mozambique, funded by the UK's Department for International Development (DFID), claimed to be a RCT but upon further investigation and discussions with the evaluators, it became apparent that some non-random elements had crept into the allocation of individuals to treatment and control groups through challenges encountered during fieldwork. This can obviously have serious consequences for the reliability of the estimates obtained from RCTs, and it is not unusual for studies not to describe their randomisation process accurately, or in much depth.

Furthermore, many RCTs do not address the issue of attrition appropriately. Attrition refers to individuals that have been assigned to either treatment or control groups but have then decided not to proceed with the experiment. It is often not clear why those individuals drop out, and this behaviour can have adverse effects on the results of the experiment. It is frequently argued that individuals dropping out would have been worse off than the ones remaining and hence a risk of overstating impact estimates exists, but the opposite can also be true. Drop-outs change the composition of treatment and control groups thereby influencing the results of the experiment since their outcomes cannot be observed. It is possible to track the individuals that drop out, and thereby one can address any side effects of attrition, but this is a costly undertaking. More importantly, all randomised studies should report the level of attrition and compare drop-outs with the individuals that remain in the study to gauge whether there are systematic differences between these two groups – at least in terms of observable characteristics.

Another key challenge affecting the generalisability of RCTs is linked to behavioural changes that can influence treatment and control groups. These behavioural changes are known as *Hawthorne* and *John Henry* effects, with Hawthorne effects referring to behavioural changes in the treatment group while John Henry effects relate to behavioural changes in the control group. For example, individuals in the treatment group might positively change their behaviour for the duration of the study as they feel thankful for receiving treatment and as a response to being observed. The same behavioural changes might apply to members in the control group altering their behaviour positively or negatively.

A final technical challenge we need to understand is related to spill-over effects that can have adverse effects on the impact estimates obtained from a RCT. Spill-over effects refer to individuals in the control groups that are affected by the treatment in physical ways or in the form of price changes, learning or imitation effects. But individuals in the treatment group can also be affected by spill-overs, e.g. changes in migration patterns through being attracted by the treatment can have an effect on the impact of the programme. In the case of Mexico's PROGRESA conditional cash transfer programme, spill-over effects caused by migration were detected, but the good news is that these spill-overs, if significant, can be measured and checked for. For example, the level of treatment exposure within groups can be adjusted to assess the magnitude of potential spill-over effects.

In addition to these technical challenges, potential ethical challenges should not be overlooked. The implementation of RCTs is not always feasible because of ethical considerations, e.g. how can it be justified that certain individuals are assigned to a treatment group while others are excluded from a potentially

Double-blindedness is usually not possible in evaluations such as those on the impact of microfinance programmes.

Photo: Jörg Böthling

beneficial treatment. Many argue, however, that these ethical concerns are not valid considering that if a treatment is proven to be beneficial, it will eventually become available to all individuals in the control group as well.

Finally, there are practical challenges to overcome in the successful implementation of RCTs; extensive co-operation from the programmes that are being evaluated is required. This can be time and cost intensive. Laura E. Bothwell and co-authors argue that RCTs are high-cost and high-value marketing tools and hence value for money will need to be carefully considered before embarking on one, e.g. with regard to what percentage of the overall programme budget should be allocated for conducting evaluations. Are the funds sufficient to conduct a high quality RCT? Is the RCT the appropriate methodological option to answer the questions of interest in relation to its costs? Moreover, for RCTs to work, the environment needs to be rigorously controlled, so that any difference in outcomes between the two groups can be adequately attributed to the impact of the programme. Therefore, applying RCTs is in many cases not desirable or feasible and hence, we need to consider robust alternatives.

> *"We need methodological pluralism and an open-mindedness among researchers and commissioners of evaluation research."*

LET'S THINK ABOUT ALTERNATIVES

There is an increasing role for qualitative methods in impact evaluation such as process tracing and life histories but also for experimental and behavioural games as well as for social network analysis, longitudinal studies and other modelling approaches. It is beyond the scope of this article to discuss these alternatives in depth, but it should be noted that strictly quantitative approaches such as RCTs can easily be replaced and/or complemented with cost-effective alternatives that often focus on gaining a better understanding of the causal mechanisms that underpin a particular programme with the objective to unpack its 'black box'.

Given the challenges outlined above, is the recent enthusiasm for RCTs sustainable? In principle, RCTs have the best chance to meeting a range of evaluation challenges such as controlling for selection bias, constructing robust counterfactual scenarios, etc. However, Elliot Stern and co-authors argue that in 95 per cent of all cases RCTs are not feasible or appropriate. Hence, we maintain that we need to think more seriously about alternative as well as complementary methods to RCTs.

RCTs promise rigour and certainty which may explain why they have become so popular but rigour is not just limited to RCTs. Other disciplines such as law, ecology and others rely on other techniques such as rules of evidence, aerial photographs and satellite imagery to demonstrate causation. There may also be value in exploring relatively inexpensive methods that have been little used in the area of impact evaluation so far such as experimental and behavioural games, social network analysis, agent-based modelling and other simulation approaches. These approaches can often be more powerful than RCTs alone for understanding the underlying causal mechanisms of programmes, and they are particularly useful when faced with small n evaluations (those involving small sample sizes) and/or evaluations of complex interventions in particular in conflict-affected areas where RCTs have serious limitations.

The choice of an evaluation study design, whether to use a RCT, a quasi-experiment, qualitative tools or a mixture thereof, should depend on the objectives of the evaluation, access to financial resources and time horizons. Methodological rigidity will not help us to better understand the effectiveness of development programmes – what we need is methodological pluralism and an open-mindedness among researchers and commissioners of evaluation research to allow the best possible evaluation design given the specific context we find ourselves in. Context and appropriateness of methods matters!

Maren Duvendack has a PhD in development economics from the University of East Anglia (UEA), UK. Her key research areas cover applied micro-econometrics, impact evaluation methods, systematic reviews and meta-analysis, microfinance, replication and reproduction of quantitative analyses as well as research ethics. Duvendack has extensively worked on microfinance impact evaluations in India and Bangladesh. Contact: m.duvendack@uea.ac.uk

For a list of references, see the online version of this article at: **www.rural21.com**

WHAT ABOUT EVALUABILITY?

All considerations regarding the right design of an evaluation aside, one aspect that must not be forgotten is evaluability, i.e. "the extent to which an activity or project can be evaluated in a reliable and credible fashion", as defined by the Development Assistance Committee of the Organisation for Economic Co-operation and Development (OECD-DAC).

While an evaluation aims to judge the merits of a particular intervention, an evaluability assessment occurs before an evaluation. It can support formulating a recommendation on whether an evaluation is worthwhile in terms of its likely benefits, consequences and costs. Also, it can show at which point the evaluation should take place and help decide whether a programme or intervention needs to be modified, whether it should go ahead, or whether it should be stopped. Assessing the evaluability of a measure can prevent wasting valuable time and resources on a premature or inappropriately designed evaluation. And, as a WorldBank Group blog explains, it can "thwart 'evaluitis' and the 'ritualization' of evaluation processes".

The Overseas Development Institute (ODI – UK) authors of the manual "Evaluability Assessment for Impact Evaluation" maintain that three focus areas ought to be covered by an evaluabiltiy assessment:

- the adequacy of the intervention design for what it is trying to achieve,
- the conduciveness of the institutional context to support an appropriate evaluation, and
- the availability and quality of information to be used in the evaluation.

The guide contains a checklist to help evaluators to answer the following key questions:

1. Is it plausible to expect impact? This is where the adequacy of the intervention design is examined. Do stakeholders share an understanding of how the intervention operates? Are there logical links between activities and intended impact?

2. Would an impact evaluation be useful and used? Here, the focus is on stakeholders, demand and purposes. Are there specific needs that the impact assessment will satisfy, and can it be designed to meet needs and expectations?

3. Is it feasible to assess or measure impact? This question refers to data availability and quality. Is it possible to measure the intended impact, given on-the-ground realities and evaluation resources available?

The manual is available for downloading on the ODI website: **www.odi.org**. Useful information on evaluability can also be found on the BetterEvaluation project website: **www.betterevaluation.org** (sri)