ILAC Working Paper 13

Randomised Control Trials for the Impact Evaluation of Development Initiatives: A Statistician's Point of View

Carlos Barahona

October 2010

Institutional Learning and Change (ILAC) Initiative - c/o Bioversity International
Via dei Tre Denari 472°, 00057 Maccarese (Fiumicino), Rome, Italy
Tel: (39) 0661181, Fax: (39) 0661979661, email: **ilac@cgiar.org**, URL: **www.cgiar-ilac.org**

The ILAC initiative fosters learning from experience and use of the lessons learned to improve the design and implementation of agricultural research and development programs. The mission of the ILAC initiative is to develop, field test and introduce methods and tools that promote organizational learning and institutional change in CGIAR centers and their partners, and to expand the contributions of agricultural research to the achievement of the Millennium Development Goals.

Citation: Barahona, C. 2010. *Randomised Control Trials for the Impact Evaluation of Development Initiatives: A Statistician's Point of View*. ILAC Working Paper 13, Rome, Italy: Institutional Learning and Change Initiative.

Table of Contents

Randomised Control Trials for the Impact Evaluation of Development Initiatives: A Statistician's Point of View

Carlos Barahona[a]
[a] *Deputy Director, Statistical Services Centre, University of Reading, UK, c.e.barahona@reading.ac.uk*

1. Introduction

This paper contains the technical and practical reflections of a statistician on the use of Randomised Control Trial designs (RCT) for evaluating the impact of development initiatives. It is divided into three parts. The first part discusses RCTs in impact evaluation, their origin, how they have developed and the debate that has been generated in the evaluation circles. The second part examines difficult issues faced in applying RCT designs to the impact evaluation of development initiatives, to what extent this type of design can be applied rigorously, the validity of the assumptions underlying RCT designs in this context, and the opportunities and constraints inherent in their adoption. The third part discusses the some of the ethical issues raised by RCTs, the need to establish ethical standards for studies about development options and the need for an open mind in the selection of research methods and tools.

2. Randomised Control Trials in Impact Evaluation

Over the past decade Randomised Control Trial (RCT) designs have been proposed and used for impact evaluation studies, sometimes as a requirement. Some proponents of RCTs have even suggested that RCTs set a methodological 'gold standard' that ensures rigour and scientific validity. This section outlines the arguments put forward by RCT proponents and the debate that has been generated. A good overview of RCT for programme evaluation was presented by Clinton *et al*. (2006) in a report written for the Congressional Research Service of the US government. The report provides an excellent description of what RCTs are and what they do:

> *"An RCT attempts to estimate a program's impact on an outcome of interest. An outcome of interest is something, oftentimes a public policy goal, that one or more stakeholders care about (e.g., unemployment rate, which many actors might like to be lower). An impact is an estimated measurement of how an intervention affected the outcome of interest, compared to what would have happened without the interventions. A simple RCT randomly assigns some subjects to one or more treatment groups (also sometimes called experimental or intervention groups) and others to a control group. The treatment group participates in the program being evaluated and the control group does not. After the treatment group experiences the intervention, an RCT compares what happens to the two groups by measuring the difference between the two groups on the outcome of interest. This difference is considered an estimate of the program's impact."*

In the USA the use of RCTs has been proposed by several players, including the Coalition for Evidence-Based Policy and the Institute of Education Sciences. In a paper entitled 'Bringing

Evidence-Driven Progress to Education: A Recommended Strategy for the US Department of Education' (2002), the Coalition for Evidence-Based Policy stated that:

> *"(the Coalition's) primary agenda is to work with key policymakers in the federal agencies and Congress to incorporate an 'evidence-based approach', as follows, into social and economic programs:*
>
> *Allocation of government resources – Government funding or other benefits should be allocated to activities that either (i) have been shown to be effective through rigorous empirical studies, or (ii) as a condition of funding (or other benefit), will undergo an independent, rigorous evaluation.*
>
> *Government R&D investment – The government should undertake, or otherwise advance, a systematic R&D effort to identify and/or develop effective government interventions. This R&D effort should use rigorous study designs, such as the randomized controlled trial."*

In November 2003, the US Department of Education proposed giving priority to:

> *"… program projects proposing an evaluation plan that is based on rigorous scientifically based research methods to assess the effectiveness of a particular intervention. The Secretary intends that this priority will allow program participants and the Department to determine whether the project produces meaningful effects on student achievement or teacher performance.*
>
> *Evaluation methods using an experimental design are best for determining project effectiveness. Thus, the project should use an experimental design under which participants — e.g., students, teachers, classrooms, or schools — are randomly assigned to participate in the project activities being evaluated or to a control group that does not participate in the project activities being evaluated. If random assignment is not feasible, the project may use a quasi-experimental design with carefully matched comparison conditions." (Federal Register, 2003)*

This position is similar to that taken by important players in the international sphere. The introduction of RCTs into evaluation work is, to a large extent, due to the desire to introduce 'rigour', as seen by some disciplines, into evaluation practice. White (2009) discusses how part of this drive came from international institutions such as the World Bank, the Network of Networks for Impact Evaluation (NONIE) and the Development Assistance Committee (DAC) who were looking for 'more and better' impact evaluation. The Abdul Latif Jameel Poverty Action Lab (J-PAL) (http://www.povertyactionlab.org/) is a strong and enthusiastic promoter of the use of RCTs for generating information for policymaking. For example, on its website J-PAL argues that: "To shape good policy, we must understand what causes the problems we are trying to cure and which cures work. Randomized trials are central to generating this knowledge." (http://www.povertyactionlab.org/research/rand.php)

These and other RCT proponents suggest that this approach sets a methodological 'gold standard' for impact evaluation work. This view tends to be based on the ability of RCTs to deal with bias and on the claim that RCTs offer the possibility of attributing impact to the intervention under evaluation.

The drive for RCTs has been significant, but the methodological and statistical arguments put forward by proponents, valid under certain conditions, need careful assessment to determine whether RCTs are an appropriate tool in each case. What is clear, however, is that RCT proponents have generated a strong reaction from many institutions, scientists and people in the evaluation community. In the US, the American Evaluation Association (AEA) issued the

following statement in response to the US Department of Education's proposed prioritising of RCTs: "We believe the proposed priority manifests fundamental misunderstandings about (1) the types of studies capable of determining causality, (2) the methods capable of achieving scientific rigor, and (3) the types of studies that support policy and program decisions. We would like to help avoid the political, ethical, and financial disaster that could well attend implementation of the proposed priority."

This response, in turn, split opinion in the AEA. The *Journal of Multidisciplinary Evaluation* has become a forum for authors to present their considered opinions about the role of RCTs in evaluation.

In September 2007, the European Evaluation Society (EES) felt the need to issue a statement about "the importance of a methodologically diverse approach to impact evaluation – specifically with respect to development aid and in development interventions", pointing out that "the EES supports multi-method approaches to impact evaluation and does not consider any single method such as RCTs as first choice or as the 'gold standard.'" (http://www.europeanevaluation.org/news?newsId=1969406)

In the international debate there are many other players who have also reacted to the launching of RCT designs as the 'gold standard' methodology, not all of them organised or with a strong voice. Their counter-arguments reflect a wide range of positions associated with their particular professional disciplines, but in general the message is that other methodologies for impact evaluation are also valid and can be rigorous.

There are many professionals and organisations working in international development and poverty programmes who find it difficult to engage in this debate, but are at the receiving end of the decisions made by funding agencies about the methodologies that should be used for impact evaluation. They often feel under pressure to use or accept the use of RCTs, or quasi-experimental designs, for evaluation work. One of the problems they face is that the RCT proponents often present their arguments using statistical jargon that is difficult to understand by anyone without a mathematical/statistical background. In addition, these proponents have appropriated such terms as 'rigour', 'scientifically valid' and 'objective assessment', making the debate extremely difficult for people who are not experts in evaluation methodology or statistics.

The following sections of this paper describe the origins of RCTs, their general strengths and limitations, and the practical, ethical and quality issues involved in applying RCTs in development contexts. Although this has been done before by other authors, I would like to encourage a more informed debate about the use of RCTs. Above all, I am keen to encourage the selection and use of methodologies and tools that are appropriate for tackling the research questions of interest, taking into account the context under which the research is being conducted.

3. Origin of Randomised Control Trials (RCTs)

The term 'Randomised Control Trial' has its origins in clinical research. Although proposals for the use of controls and some forms of randomisation were made as early as the 18th century, the theoretical and methodological groundwork was carried out in the first half of the 20th century, and the wider adoption of RCTs for clinical studies by governments and the private sector did not start until the 1950s. Clinical experiments have their pedigree in the experimental designs used for agricultural experiments in the 1920s by R. Fisher. A classic textbook on clinical trials, with a detailed description of RCT methodology for clinical applications, is Clinical Trials (1983), by Stuart J Pocock, a book I would recommend to

anyone interested in understanding the strengths and constraints of RCTs and the conditions under which they developed. Clinical trials are "planned experiments which involve patients and are designed to elucidate the most appropriate treatment for future patients given a medical condition" (Pocock, 1983). The use of RCTs is common in Phase III trials when a treatment is known to be reasonably effective and needs to be compared with current practice. Earlier stages in clinical research involving safety and effectiveness (Phases I and II) do not use RCT designs. Using RCT methodology is seen as the most reliable way of establishing the differential effect between accepted treatments and new ones.

Some of the main characteristics of clinical experiments conducted using RCT methodology are as follows:

i. The main objective of an RCT is to compare the size of the effect of two or more treatments for a specific condition.

ii. The clinical trial is not conducted with the purpose of finding out if a treatment has a beneficial effect on the condition of interest. In other words, the RCT is not set up to determine a causal link between the treatments and any improvement in the patient's condition. This link is known from previous phases in clinical research.

iii. The way treatment outcomes are measured is explicitly defined as part of the process of designing the trial and is known to be a reliable indicator of changes in the condition of the patient.

iv. Treatments are applied to a very special type of experimental unit: human subjects. Patients are considered eligible for the trial if they present the specific condition for which the treatment has been developed.

v. Patients are recruited into the trial through their clinicians. Clinicians who are willing to collaborate with the trial screen patients for eligibility using a detailed protocol, and then propose them for inclusion.

vi. There is a requirement to inform patients about the trial objectives and procedures. In general, patients are expected to give informed and explicit consent for their inclusion in the trial. Depending on country-specific legislation and the nature of the condition under treatment, the level of informed consent varies.

vii. The allocation of treatments to patients is done at random in order to obtain an unbiased evaluation of the effect of new treatments, hence the term ‘randomised‘.

viii. Patients in the trial are not allowed to receive any treatment for a condition other than the one that the trial has assigned to them. Patients who deviate from the trial treatments would be dropped from the trial.

Clinical research is carefully regulated. Since clinical trials are about experimentation on human subjects, legislation exists in most countries specifying the conditions under which such research can be conducted. RCTs are also subject to the approval of research ethic committees that assess the trial before it is conducted. Researchers are required to develop a full protocol for each trial, detailing the objectives, patient selection criteria, treatments, methods of evaluation, experimental design, method of patient registration and randomisation, procedures for obtaining patient consent, procedures for monitoring trial progress, data handling systems, analysis plan and details of administrative responsibilities. No serious clinical trial is conducted without a full protocol and ethical approval. The quality of a clinical trial can be assessed only by careful examination of the protocol and the extent to which the implementation of the trial adhered to it.

4. Why is Randomisation so Important in RCTs?

Randomisation of treatments among experimental units was proposed by Fisher (1935) for the following purposes:

i. To guard against any use of judgement or systematic arrangements leading to one treatment being disadvantaged (i.e., to avoid bias)

ii. To provide a basis for the standard methods of statistical analysis, such as significance tests.

Fisher was dealing with agricultural experiments where the experimental unit was often a plot of land. To date, these remain the reasons for the randomisation of treatments among experimental units in this type of trial.

In clinical trials, the randomisation process requires more care than in the designs proposed by Fisher for agricultural research. The fact that a patient is aware of the treatment he/she is receiving could affect that patient's response to the treatment. Furthermore, if a clinician is aware of which treatment is being assigned to a patient, the temptation to influence the allocation of perceived ‘better treatments’ to patients who are seen to be in greater need of a new treatment or, conversely, to protect vulnerable patients from treatments that have not been fully tested, can introduce serious biases into the trial. This led to the implementation of ‘double-blind’ trials. In these trials neither the patient nor the clinician is aware of the specific treatment that the patient is receiving. Only the researchers in charge of the trial have access to this information, further reducing the opportunities for introducing bias. Another benefit of randomisation is that it helps to reduce the risk of different effects due to the experimental environment. Because of the random allocation of treatments, it is unlikely that systematic patterns associated with non-experimental factors will affect any of the treatments. It is reasonable to expect that any environmental effects would be averaged out among the patients allocated to each treatment group.

5. Selecting Treatments for a Trial

5.1. Control Treatment and Control Group

In clinical trials, one of the treatments under comparison is regarded as a 'control treatment'. Normally, this refers to the best standard treatment available for the condition and therefore the benefits of any new treatment must be compared against this standard. Only in exceptional cases can the control treatment be defined ‘no treatment’. This happens when no standard treatment is available and it always depends on the ethical implications of ‘no treatment’ being considered acceptable. The group of patients receiving the control treatment is referred to as a 'control group'.

The control group can be distinguished from the treated group(s) only by the treatment it receives. In standard RCTs, this is achieved by recruiting the experimental subjects into the trial and then randomly allocating a treatment to each of them only when they have been accepted. Because experimental subjects are recruited using a unique set of rules, and because treatments are allocated at random, it is expected that the control and treatment groups would be similar except in terms of their response to the different treatments. This is what makes it possible to assess the differential effect of the new treatment(s) with respect to the control. It allows an estimate to be made of the difference in the size of the average effect between the treatments under comparison. The random allocation of treatments also allows an estimation of the precision of this difference to be made and statistical hypothesis tests to be used.

The use of a control treatment, involving a control group, allows the researchers to answer the following question: Is there a big enough difference between the average effect of the new treatment and the control, in favour of the new treatment, to persuade us to prefer the new treatment over the standard one?

In clinical trials, the main purpose of the use of a control group is to satisfy the need to demonstrate a significant difference between treatments and not the need to determine whether or not the treatment is effective. Given the level of scrutiny that clinical trials have prior to their implementation, no ethics or professional committee would allow the implementation of a trial where the effectiveness of a treatment had not already been demonstrated.

Because of the level of control over external factors that may confound the effect of the treatment imposed by the research process and the care that researchers take in avoiding contamination of treatment and control groups, the data from an RCT can sometimes be said to provide empirical evidence that supports the claim of a causal relationship between treatment and effect. However, this is not always possible, and I would argue that this type of evidence is not always enough, let alone necessary, to establish a causal relationship between stimulus and effect.

5.2. Selection of Treatments

The selection of treatments (i.e., the treatments other than the control treatment) to be tested in an RCT is an important topic. In clinical trials it is rarely discussed because it is assumed that when a treatment gets to the stage of being tested in an RCT, its safety and effectiveness would have already been established. No researcher or clinician, either in government or the private sector, would consider submitting a proposal for a trial for the consideration by a research or ethical committee unless there was enough evidence to demonstrate that the new treatment was safe, effective and, at least in least some respects, expected to be better than the current best practice.

6. Using RCTs to Determine What Works and What Does Not

There are many reasons for conducting an impact evaluation and many views about what an impact evaluation is intended to do. For example, in the NONIE Guidance on Impact Evaluation, Leeuw and Vaessen (2010) write that the "*important reasons for doing impact evaluations are to:*

- *Provide evidence on 'what works and what doesn't'*
- *Measure impacts and relate the changes in dependent variables to developmental policies and programmes*
- *Produce information that is relevant from an accountability perspective*
- *Benefit from individual and organizational learning."*

It should be evident that, in order to achieve these goals, a broad range of methodologies is needed and the evaluator's toolbox should contain more than the RCT design. However, it has been argued that the RCT approach is the best way of finding out 'what works and what doesn't'. RCT designs have been shown to be effective in enabling an estimation to be made of the difference in the average effects of the treatment and the control, under very stringent conditions that isolate the experiment from contamination by influential factors other than the

treatment(s) and the control. I see two problems, however, when attempting to use RCTs for the impact evaluation of development initiatives:

i. The need to limit the indicator(s) of impact to quantities that can be measured with enough accuracy

ii. For the methodology to work, the level of control needs to be good enough to render any contamination effect negligible

Within the context of development interventions, the solution to the first problem depends on the choice of the impact indicators. This is a minor problem when an indicator can be measured accurately (e.g., when looking at percentages of people/households with access to a particular service, provided that 'access' is defined unambiguously). The problem becomes more difficult when the indicator is prone to significant errors of measurement (e.g., indicators such as income or poverty level). I am referring here to non-sampling, non-measurable errors that are likely to have a significant impact on the accuracy of the method used to measure the indicator. These errors carry the non-measurable (or at least very difficult to measure) risk of rendering useless any precise comparison of the mean effects of treatments.

The second problem – contamination – is more worrying. This problem is not new to the impact evaluation community. Mohr (1995) mentions two types of contamination. The first is contamination in the delivery of treatments (where the treatment is not administered properly), which is not uncommon in development interventions due to the complex physical, social and institutional environment in which the interventions take place. The second one is the contamination that "takes the form of incomplete or improper environmental controls". In development interventions, this relates to the difficulty of isolating experimental units from other interventions that produce changes in the impact indicator(s). This is particularly true where interventions occur in environments in which multiple actors are working towards promoting development with overlapping or parallel interventions that generate known and unknown impacts over the complex set of indicators associated with development. The result is that the difference between treatment and control cannot be said to be due only to the intervention of interest, thus eliminating the main characteristic that makes the RCT design so useful under 'controlled conditions'.

In some development interventions, contamination also occurs when units assigned to a control intervention are able to 'learn' from those units in the treatment interventions. Trying to prevent this type of contamination can reduce the comparability of the control and treatment interventions.

To some extent, statistics could help to mitigate this problem. It could be argued that if these contaminating factors were measured, their contaminating effect could be accounted for in the data analysis. Although this is true, it also takes us back to the problem of measurement, and requires careful measurement of detectable confounding factors. Scriven (2008) points out that: "These factors have four important properties: they are potentially fatal flaws in an RCT, they have often ruined very expensive RCTs designed and run by very well trained researchers, their effects can virtually never be factored out ex post facto, and they require specially trained observers almost constantly watching for their emergence, continuing presence and magnitude - observers who have to be empowered to act quickly to stem the swift haemorrhaging of validity." The problem remains unsolvable for those contaminating factors of which the evaluator is unaware.

Deaton (2010) discusses the use of RCTs in development economics to "accumulate credible knowledge of what works" and argues that "experiments have no special ability to produce more credible knowledge than other methods, and that actual experiments are frequently

subject to practical problems that undermine any claims to statistical or epistemic superiority." It is encouraging to see that, as a result of the increased use of RCTs in development evaluation, a more systematic and evidence-based analysis is beginning to emerge on the pitfalls and weaknesses of RCT and quasi-experimental approaches. Donaldson and Christie (2008), in their book *What Counts as Credible Evidence in Applied Research*, argue that what is accepted as credible evidence depends on the question, the context, the assumptions we are prepared to make, the theory of evaluation we are prepared to accept and the resources available. The problem faced by many individuals and organisations wanting to find out what works is that powerful players seem to have bought the argument that one methodological approach is suitable for all situations. This has generated in many quarters what appears to be an ill-informed and dogmatic approach to the request for and assessment of evaluation proposals, with the consequent impact on contract awards. Perhaps Donaldson and Christie's call for more reflective assessment about what we accept as credible evidence is something we need to consider more seriously.

7. Using RCT Designs to Investigate Casuality

What has been claimed for RCTs is that they establish inferential causality between intervention and impact. This is achieved by comparing mean effects and randomisation, which helps to reduce selection bias.

Under a well-designed and carefully conducted RCT, where contamination is known to be negligible, issuing inferential causality statements is not difficult because the experimenter will control everything else apart from the treatment; any bias would have been removed by the randomisation process, and the variable of analysis is known to be a reliable indicator of the effect of the treatment. This requires the experiment to be conducted under very special conditions.

However, as argued earlier, the difficulties in controlling the conditions under which development interventions are carried out are such that the ability of RCTs to provide a standard (‘gold’ or otherwise) that justifies causality statements is left on shaky ground.

As a statistician, trained originally in biometrics, I am surprised at how frequently the RCT approach is recommended as a way of determining whether the recorded impact can be attributed to the intervention. As discussed earlier in relation to the origin of RCTs, they enable an estimation to be made of the size of the mean difference between treatment(s) and control, and not whether the treatment has an effect. In the field where RCTs were first widely used – clinical research – no serious researcher would consider conducting an RCT using a treatment for which there was no evidence of effectiveness and safety. No research committee would consider a proposal for an RCT without evidence that the new treatment(s) had known advantages over the standard treatment (control) and, more importantly, no ethics committee would approve the implementation of such a trial.

There is abundant work on the theory of causation and on methodologies for conducting enquiries about causation. Menzies (2008) presents a good review on the counterfactual theories of causation. Scriven (2008) suggests alternative methods for causal analysis, and also provides a critical view of the use of RCTs for this purpose.

8. Control Groups as Counterfactuals

Proponents of RCTs for impact evaluation often claim that having a control group is the best

option to set a counterfactual. White (2009) points out that "having no comparison group is not the same as having no counterfactual" and describes his approach to impact evaluation as trying to understand the underlying causal chain, starting with the outcomes and impacts in order to identify the range of factors that influence these factors. In a similar vein, Scriven (2008) proposes that “critical observation is a valid methodology for establishing causal effects.” Both these authors describe what I consider to be a sensible approach to inferential causation.

At best, a control group provides a reference for determining the difference in impact, but if this control is to serve as a counterfactual this implies the use of a control treatment that in fact is ‘no treatment’. Pocock (1983) discusses the use of ‘no treatment’ as a control in clinical research and describes how, when no standard treatment is available, the use of an intervention known to have no effect – the placebo – is required. He points out that “one great danger in having control patients who are completely without treatment is that one cannot decipher whether any response improvement in the treated group is genuinely due to therapy or due to the act of being treated in some way." The problem for impact evaluation is even greater. The ethical issues that accompany ‘no intervention’, the practical impossibility of a placebo and the natural reaction of a group that knows that it has been excluded from the intervention, make the rigorous use of control groups for experiments in development initiatives extremely difficult.

The extent to which a control group tells us "what would have happened in the absence of the intervention" would vary depending on the conditions under which the experiment is conducted. Under ideally controlled conditions, the comparison with the control is probably equivalent to a counterfactual, but under the imperfect conditions in which most development interventions work, what would have happened in the absence of the intervention is not necessarily what happened to the control group. A common situation is that if the intervention under assessment had not taken place, a different intervention would have happened, often because of the activities of different agents promoting development. The way in which development agencies balance the use of their resources depending on a range of factors (e.g., their knowledge or perception of what is going on in specific locations, the history of their interventions, their local contacts, the presence and activity of other agencies) makes the use of control groups as counterfactuals problematic. The argument is even weaker where a development intervention contributes to but does not cover all the activities that an implementing agency carries out in a particular area or with a specific population. In these cases, the possibility of establishing a counterfactual using a control group is almost non-existent, and in the context of development interventions the assumption that 'if my intervention had not happened nothing else would have happened' is unrealistic and arrogant.

In those cases where all these issues can be solved satisfactorily, one further issue remains – the requirement that an RCT design ensures that no intervention takes place in certain experimental units. This raises ethical issues that need to be seriously considered.

9. When to Ask if the Intervention Works or Not

A question that needs to be answered by those promoting the use of RCT design for the evaluation of development interventions is this: Why is anyone asking whether the intervention works at the time when it is being implemented on a medium to large scale by agencies whose function is to promote development? At this stage in the implementation of a development initiative, the causality should be well established and understood (as in clinical research where safety and efficacy has been reasonably established before the RCT phase). If this is not the case, the ethical aspects of the development initiative (not just of the evaluation)

need to be questioned.

I have recently been involved with some projects that the Research Into Use (RIU) programme 12 is funding in Asia. There, I have often been asked if the use of control groups, and RCT design in particular, should be part of the Monitoring and Impact Learning (MIL) process. According to the RIU website, the programme "builds on the DFID Renewable Natural Resources Research Strategy [RNRRS][1] 1995 to 2005, which funded research on crops, livestock, fisheries, forestry, post-harvest issues and natural resource management. Much of this research has a great deal of unfulfilled potential to impact on poverty. RIU aims to realize that potential and to learn lessons that can be incorporated into future research for development." The main RIU outputs[2] are described thus:

- RIU is enhancing access to research outputs to greatly benefit the poor. Output 1: Significant use of RNRRS and other past research results increased
- RIU is gathering concrete evidence of what works and why. Output 2: Research-into-use evidence generated
- RIU is working to embed innovation for the poor in development agendas. Output 3: Research-into-use lessons on policies and practices gathered and shared

The RNRRS funded research projects to develop pro-poor approaches. Its goals were the "alleviation of poverty, promotion of economic growth and of economic reform, and the mitigation of environmental problems. Achievement of the goals required the research to be wealth creating and/or improving the life for beneficiaries in developing countries" (RIU, 2009). The RNRRS generated many research outputs through studies carried out by researchers based throughout the world, with support from universities, international research institutes and national research programmes in the countries where the studies took place. These results were presented in scientific reports, peer-reviewed scientific journals and conferences. While the research outputs are numerous and useful, the RIU programme acknowledges that "much of that potential remains unrealized, in part because of the difficulties of scaling up the results of research, i.e. multiplying them on a large scale. This is a process about which all involved in development have much to learn" (RIU, 2009). My response to the RIU question about the need for control treatments is in the form of the following questions:

- Do you know if the research outputs that are being up-scaled work?
    - If not, why are you scaling them up? Should you not be trying to test first whether they work?
    - If yes, then what do you want to find out about them?

The answer has been that the RIU programme is working with interventions that the RNRRS developed, tested and are known to work, and that the programme was established to meet the twin challenge of scaling up and learning about scaling up (RIU, 2009). The conclusion is that the process of learning about how scaling up works, what factors enable, contribute to and hinder the work of partnerships between government institutions, NGOs, research organisations, community organisations, households, and individuals participating in the scaling up process, and how the prevailing conditions (policy, natural environment, social capital) affect the process of scaling up needs to be addressed using multiple methodologies. As a statistician, I have not yet found the need to bring any of the experimental design tools

[1] http://webarchive.nationalarchives.gov.uk/+/http://www.dfid.gov.uk/research/renewable-natural-resources.asp

[2] http://www.researchintouse.com/rnrrslegacy/pub_practcomponents.html

out of my toolbox when working with RIU.

10. Randomisation

The general benefits of the randomisation of treatments to experimental units have been clearly stated and, from a statistical point of view, are not under discussion. The fact that a design is said to be randomised, however, does not automatically means that the process has been conducted properly or that the benefits of randomising treatments to experimental units can be realised. Extensive work has been conducted on how randomisation should take place in an experiment, and in most cases it is slightly more complicated than closing your eyes and choosing a random number from a table with the tip of your pencil.

In particular, one of the pillars that support the reputation of RCTs is the fact that they are 'double blind' – neither the experimental unit nor the agents administering the treatment know what treatment is being applied. This eliminates bias.

There are major difficulties, however, in using 'double blind' RCTs for the impact evaluation of development initiatives. Where an initiative is implemented is often determined by where the capacity to implement the intervention exists, where the investment is more likely to yield higher benefits, where the need is considered to be greatest and (a less palatable consideration) where it is politically more convenient to intervene. Although in clinical research the 'where' is also to some extent deliberately decided, the main difference in RCTs in clinical research is that these studies produce a carefully developed protocol for recruiting experimental units within the 'where'.

The protocol for recruitment rigorously establishes which patients can be recruited and which patients cannot. The selection of experimental units in the context of development initiatives would also have to be guided by a properly developed recruitment protocol if the RCT methodology was to be considered rigorous.

If we assume that a protocol is developed and followed rigorously, how should randomisation be carried out? After the experimental unit has been recruited and has agreed to take part in the trial, a treatment should be allocated at random to it. This is where the 'double blind' method is applied. One of the 'blind' elements is that the experimental unit knows it is receiving a treatment, but is totally unaware what the specific treatment is. The second 'blind' refers to the agent applying the treatment being unaware of which specific treatment is being administered. Neither of these levels of blinding happens in the majority of RCTs used for impact evaluation. Blinding the implementing agency is almost impossible and would have undesirable consequences (e.g., making it difficult to engage in a process of institutional, organisational or individual learning, something generally considered to be important in achieving developmental goals). Blinding the experimental units is also difficult because development initiatives do not come in the form of pills that can be made to look the same regardless of their content. It could also be argued that individuals, households, communities and institutions need to be aware participants, if not full partners, in development initiatives, and therefore should not be blinded about its nature.

In pointing out how the randomisation of treatments in development initiatives diverges from the randomisation process in clinical research, it could be said that I am taking a pedantic attitude towards the conduct of RCTs and that in impact evaluation the RCTs need to be adapted to fit into a reality that is more complex than the more controllable set of conditions under which clinical research is conducted. I agree. Given that 'double blinding' is fundamental to RCT designs, however, if we are not 'double blinding' we are not using RCT methodology with rigour. If this is so, we are moving towards seeking appropriate

methodologies that could be used with rigour to answer relevant questions. I would strongly argue for this attitude rather than the inaccurate application of methodologies and their labelling to be able to claim methodological rigour.

11. Requirement of Randomisation for Valid Statistical Inferences

The issue of randomisation remains important when statistical inferences are required. This could be the random allocation of treatments to experimental units when conducting experiments or the random selection of sampling units when estimates are being sought.

Impact evaluations tend to ask such questions as: “Has the intended impact been achieved?”, “To what extent has the intervention delivered the intended impacts or outcomes?” and “Have the intended beneficiaries been reached?” From a statistical point of view, these questions are better answered through the use of sampling methods because they refer to the estimation of parameters in a population. One problem with using RCTs for impact evaluation is that experimental designs are designs for the comparison of treatments and not for the estimation of population parameters. Using experimental designs for estimating population parameters is the methodological equivalent of drinking soup with a fork.

If the problem at hand is one of estimation, sampling methods rather than experimental designs are more likely to allow the researcher to construct a study where the estimates about the characteristics of a population of interest (or segment of it) can be obtained with the required level of precision. It is possible that there will be cases where elements of experimental design can be incorporated into early stages in the search for development initiatives that work. It is important to remember that RCT designs are only one type of experimental design. Useful and efficient experiments can be constructed if the structure of the treatments and the structure of the experimental units are carefully considered by the researcher, probably with the advice of a statistician with experience in experimental design. In the context of development interventions, however, the role of such studies is important at pilot stages and at a small scale, when hypotheses about what might or might not work are being formulated. It is only at these stages that it might be justifiable considering some of the ethical risks of experimenting with people.

12. Ethical Issues

In conducting research with people, the need for guidance and adherence to ethical standards is of the utmost importance. Most areas of research involving human subjects have compulsory or voluntary codes of conduct and ethical rules, and many countries have strict processes in place to ensure that ethical standards are met by any research involving human experimental units. There seems to be a gap, however, in research that involves human subjects carried out in the context of international development. We do not have a system of checks and balances that ensures adherence to high ethical standards. This may be because the jurisdiction of research committees does not extend to the areas where some of this research is conducted, or because so far it has not been considered research in the strict sense of the word.

It is research, and issues such as the selection of the method of enquiry, the required level of precision, control of biases, the relevance and need for certain treatments, the size of the sample or the number of replications, the appropriateness of the measurements and tools, and the ethics of the study must be considered.

An important aspect of RCTs is the need to obtain consent. When RCTs are used in clinical research, during the process of recruiting subjects these subjects are informed about a range of factors, such as the trial objectives, the treatments, the risks, the random nature of treatment allocation, the support available to participants, the mechanisms for opting out, and the procedures for providing comments and submitting complaints. This process concludes with each subject agreeing to be included in the trial. When RCTs are proposed for impact evaluation, the issue of consent from participants is not discussed. Telling a group of people that they will be included in an experiment, but not implementing a development intervention that might benefit them, is something that most people working in international development would find difficult. The subjects are normally not blind to the intervention, and therefore even if consent were obtained from all the trial subjects, the behaviour of the experimental units that were allocated the control treatment would be affected to the extent that their role as control groups would no longer be useful. I would argue that it is up to those promoting RCTs for the impact evaluation of development initiatives to provide a satisfactory answer to the issue of consent.

As noted earlier, if a control group is to be formed by a subset of experimental units that receive no treatment, the experimenter must establish a priori and deliberately the fact that these units will not receive any treatment. The problem for people working in development is that by deliberately excluding a group of people from benefiting from the development intervention, they are affecting the lives of the experimental subjects by deliberately prolonging the problems these people are facing. The issue here is with the deliberate action taken and whether it can be justified. It has been argued by RCT proponents that without doing so the quality of the research obtained is not good enough to inform decision making.

Javier Ekboir (pers. comm.) suggests that the RCT proponents' argument should be contrasted with the position of "policymakers who would argue that because of limited resources, rationing cannot be avoided, therefore, some rationing criteria must be chosen". The rationing criteria would not be suitable for forming control groups, because the conditions for excluding certain units from receiving the treatment would tend to make them different from treated units. Is the policymaker position telling us that they are not really interested in treatment-control comparisons? This is probably not true, but is closer to a practical argument where the resources for development need to be prioritised to those who need them.

Although I have earlier questioned the use of such controls, here I would argue that if and when this type of experimentation is considered absolutely essential, it must be as limited in scale and duration as possible. Any experiment that deliberately excludes individuals in need from benefiting from an intervention that has the potential to contribute to meeting their needs must be reviewed at the proposal stage by a competent ethics research committee and monitored throughout in order to ensure that the experimentation process is effective, efficient and as short as possible. The burden of proof to society (and particularly to those affected by the research) that such deliberate exclusion is absolutely necessary must lie with the proposers of the research.

I am aware of discussions about whether to include a 'no intervention' treatment in the evaluation of the impact of safety nets for the ultra poor in developing countries. Some time ago I was involved in a pilot study that provided a benefit to work-constrained and vulnerable individuals in extremely poor communities in Malawi. The pilot did not include 'no intervention' treatment, but we were questioned about its the absence. Our answer was that it would have been unethical to apply such treatment in the experiment, and that the more important question was: "What mechanism for delivery of the benefit was more effective?" The concept of a safety net is well described by its name: when everything else fails, in the absence of this support, the consequences for the individual that falls – and for those around

him/her – are highly undesirable and a safety net is needed to prevent this. At the time I had not read the paper by Smith and Pell (2003), 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials'. In this paper they proposed that "individuals who insist that all interventions need to be validated by a randomised controlled trial need to come down to earth with a bump."

Ekboir (pers. comm.) adds that "a related ethical problem is that in any development intervention where a group benefits, another group is hurt. For example, when famers organize to sell their products jointly, they get a better price but middlemen can be bypassed and they lose income. Policies are evaluated on the criterion that the total benefits should outweigh the total losses." On a different point about the use of RTCs for evaluation purposes, Ekboir points out that "the issue is that implementation of policies and programs was not (and still is not) seen as a learning exercises within complex, evolving systems." Traditionally, policymakers defined policies (or programmes) on the basis of what was supposed to work; only in the 1990s with the New Public Administration approach was "definite and quantitative evidence" a requirement (a good analysis of this is provided by Kraemer, 2006). This requirement was abandoned in most developed countries when it was found that it was not possible to develop such evidence, but it is still backed by many donors and multilateral organizations.

It is probably time for more of us to start questioning the ethics of the research, including impact evaluation, that is carried out in development initiatives. This is particularly true in the case of RCTs because the number potential of ethical pitfalls is large. There is also broad agreement that high quality research is necessary if the results are to be useful and that achieving high quality research relies partly on the rigorous use of proven methods and tools. Reducing the demand for the use of rigorous research methods is not an option, but limiting the assessment of rigour in an impact evaluation study to whether a control treatment and the randomisation of treatments have been used is an oversimplification that carries too many risks and distracts attention from fundamental issues.

Different disciplines have set standards and definitions for what is considered 'rigorous' in their methodologies. As a statistician, I am aware of the strengths and limitations of statistical methods and of how many types of enquiry would be deficient if statistical methods were forced upon them. The appropriate selection of methods depends to a large extent on the research questions, and I would argue that the dogmatic imposition of a particular design is unlikely to yield useful results in research in general or in impact evaluation in particular.

References

American Evaluation Association, (2003). Response To US Department of Education Notice of proposed Priority, Federal Register RIN 1890-ZA00, November 4, 2003 'Scientifically Based Evaluation Methods'. http://www.eval.org/doestatement.htm

Clinton, T., Nunes-Neto, B., Williams, E., (2006) Congress and Program Evaluation: An Overview of Randomized Control Trials (RCTs) and Related Issues. Congressional Research Service, Library of Congress.

Coalition for Evidence-Based Policy (2002). Report by the Advisory Board and the Executive Director. Bringing Evidence-Driven Progress to Education: A Recommended Strategy for the US Department of Education".

Deaton, A. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* 48 (June 2010): 424-455.

DFID. Evaluation of DFID Renewable Natural Resources Research Strategy 1995-2005 (EVD659) http://www.dfid.gov.uk/Documents/publications/evaluation/ev659-ch1.pdf. Accessed April 2009.

Donaldson, S., and Christie, C. (2008) What Counts as Credible Evidence in Applied Research and Evaluation Practice? Sage Publications. ISBN 78-1412957076

Duflo E., Glennersterand R. and Kremer, M., (2006). Using Randomization in Development Economics Research: A Toolkit. J-LAB, http://www.povertyactionlab.com/papers/Using%20Randomization%20in%20Development%20Economics.pdf

European Evaluation Society (2007). EES Statement: The importance of a methodologically diverse approach to impact evaluation – specifically with respect to development aid and in development interventions". http://www.europeanevaluation.org/news?newsId=1969406).

Federal Register 68(213), 62445-62447, (November 2003). Scientifically-based evaluation methods. Priority proposal by the US Department of Education. Retrieved from the Federal Register Online via GPO Access May 11th, 2006, http://www.gpoaccess.gov/fr/

Fisher, R.A., (1935) The Design of Experiments. Edinburgh. Oliver and Boyd.

Kraemer, S. 2006. Science and Technology Policy in the United States. New Brunswick, N.J.: Rutgers University Press.

Leeuw, F., Vaessen, J. (2010). Impact evaluations and development. Nonie guidance on impact evaluation. Independent Evaluation Group. World Bank. http://www.worldbank.org/ieg/nonie/

Menzies, P. (2008). Conterfactual Theories of Causation. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/causation-counterfactual/

Mohamed, M.T., and Clifton Jr., C. (2008). Processing Inferential Causal Statements: Theoretical Refinements and the Role of Verb Type. Discourse Processes,45:1,24-51.

Mohr, L.B. (1995). Impact Analysis for Program Evaluation. SAGE Publications Inc.

Pocock, S.J. (1983). Clinical Trials: A Practical Approach. Chichester, Wiley.

RIU. Research Into Use programme website (2009). http://www.researchintouse.com/ . Accessed April 2009.

Scriven, M. (2008). A Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research. *Journal of MultiDisciplinary Evaluation*, Volume 5, Number 9.

Smith, G.C.S. and Pell, J.P. (2003) 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials.' BMJ 2003; 327;1459-1461. Downloaded from bmj.com on 30 April 2009

White, H. (2009). Some Reflections on Current Debates in Impact Evaluation. International Initiative for Impact Evaluation, 3ie, Working Paper 1. http://www.3ieimpact.org