

International Initiative for Impact Evaluation



WORKING PAPER 9

An introduction to the use of randomized control trials to evaluate development interventions

Howard White
February 2011

About 3ie

The International Initiative for Impact Evaluation (3ie) works to improve the lives of people in the developing world by supporting the production and use of evidence on what works, when, why and for how much. 3ie is a new initiative that responds to demands for better evidence, and will enhance development effectiveness by promoting better informed policies. 3ie finances high-quality impact evaluations and campaigns to inform better programme and policy design in developing countries.

3ie Working Paper series covers both conceptual issues related to impact evaluation and findings from specific studies or systematic reviews. The views in the paper are those of the author, and cannot be taken to represent the views of 3ie, its members or any of its funders.

This Working Paper was written by Howard White, 3ie.

© 3ie, 2011

Contacts

International Initiative for Impact Evaluation
c/o Global Development Network
Post Box No. 7510
Vasant Kunj P.O.
New Delhi – 110070, India
Tel: +91-11-2613-9494/6885
www.3ieimpact.org

AN INTRODUCTION TO THE USE OF RANDOMIZED CONTROL TRIALS TO EVALUATE DEVELOPMENT INTERVENTIONS

Howard White*

International Initiative for Impact Evaluation

Email: hwhite@3ieimpact.org

* The author thanks Scott Rozelle and Bill Savedoff for comments on an earlier version of this paper. The usual disclaimer applies. Email: hwhite@3ieimpact.org.

Introduction

Driven by the focus on results, impact evaluation has become a prominent part of the development agenda in the last decade. Much of the discussion of results has focused on outcome monitoring, such as the attention devoted to tracking the Millennium Development Goals. Whilst useful, outcome monitoring cannot tell us the impact of an intervention, and so cannot be used to make an assessment of the contribution an agency has made to development.

But there has been growing use, notably amongst economists and political scientists, of a range of approaches which do directly tackle this question of what difference an intervention has made. Prominent amongst these approaches are experimental designs, or randomized control trials (RCTs).

The purpose of this paper is to provide a short, non-technical introduction to RCTs. More technical treatments are available from Bloom (2006) and Duflo *et al.* (2006). The paper deals briefly with what is meant by impact evaluation, before moving onto the problem of selection bias and how it can be dealt with through experimental and quasi-experimental designs. Practical concerns in designing and implementing a RCT are then discussed before moving on to some of the criticisms which are commonly made of this approach.

What is impact evaluation?

Within the development community many think of 'impact' as meaning long-run effects. This usage is the Development Assistance Committee's definition and is embodied in many versions of the log-frame. However, as I have discussed elsewhere (White, 2010) it is not at all what I mean by impact evaluation. Impact evaluation in my usage refers to looking at what difference a programme made: did it improve lives, save lives even? Impact evaluation is a 'with versus without' analysis: what happened with the programme (a factual record) compared to what would have happened in the absence of the programme (which requires a counterfactual, either implicit or explicit).

Another name for impact evaluation is attribution analysis. We want to attribute some part of observed changes to the policy, programme or project being evaluated. Again, many in the donor community mean something different by attribution. They mean attribution to their agency. I am not concerned here with that issue. I am interested in attribution to a specific intervention, regardless of who funds it. Impact evaluation is about development effectiveness not aid effectiveness. Having said that, impact evaluation of programmes supported by donor funds either directly (project aid) or indirectly (programme aid) should clearly play an important role in addressing the issue of that agency's contribution to development.

So, where is the counterfactual to come from? The answer depends on the nature of the intervention. For 'large n' interventions, in which the intervention is delivered to many units (like households, schools, clinics, firms, villages or districts) then statistical analysis is the most appropriate means of constructing a counterfactual. Specifically, the counterfactual is constructed by identifying a comparison group, which is similar in all

respects to those receiving the intervention, except that it does not receive the intervention. Then the differences in the indicators of interest (usually outcome-level indicators) are compared in the project and control groups after the intervention, called an *ex-post* single difference design. It is preferable to have data on the indicator from before the intervention also, that is a baseline survey, so a double difference impact estimate can be calculated. The double difference is the change over time in the difference in the value of the indicator between the two groups, or, equivalently, the difference in the change.

So, the next question is how to identify a suitable comparison group?

The problem of selection bias

The problem of selection bias arises because programme participants are not a random sample of the population as a whole. Rather those in the programme are selected through both programme placement and self-selection. Programme placement refers to the fact that the implementing agency targets the intervention at specific sub-populations such as female-headed households, small businesses, children at risk, schools in poor districts and so on. Self-selection occurs since people are rarely coerced to take part in development programmes. They do so voluntarily, and those choosing to participate may have different characteristics compared to those who do not do so.

Problems occur if the factors affecting whether someone participates in a programme or not are correlated with the outcomes of interest, since those participating would do better (or worse) than others regardless of the intervention. Hence if there is such a correlation, then a "naïve impact estimate", which compares average outcomes for programme beneficiaries with those for a sample of non-beneficiaries (the comparison group), will yield a biased estimate of the impact, called selection bias. The following examples illustrate this point.

An example of selection bias from programme placement is a project to improve school quality through a school investment fund for which only schools in the poorest districts are eligible to apply. Schools in poorer areas tend to have pupils whose parents are poorer and less educated, making them less able to afford complementary school supplies and, on average, less likely to want to ensure that their children attend school. Moreover these children live in housing which is not conducive to studying since it is over-crowded and poorly lit. Hence learning outcomes in the schools targeted by the project will be lower than those in non-project schools. Starting with all the disadvantages listed here, learning outcomes in project schools may still be lower than those in non-project schools even after the intervention. Hence a naïve comparison of a random sample of project schools with a random sample of non-project schools would show a negative impact of the project on learning outcomes. But this is a biased impact estimate: we have not compared like with like. To get an accurate estimate of the project impact, we have to compare the schools in the project to a set of schools in similarly poor catchment areas.

As an example of selection bias from programme placement, consider a community-driven development intervention, such as a social fund. Communities make a proposal

to the district administration for funds, to be managed by a community level committee, to undertake a project such as building or renovating the school or clinic, or building a feeder road or a bridge. Proponents of these projects argue that the experience of working together on the project will build social cohesion, or social capital. Hence beneficiary communities will be better placed to undertake local development activities on their own initiative as a result of the initial project. However, which communities will apply for the fund, given that they have to demonstrate a community-based selection process and mobilize the community to take part in construction of the project infrastructure? It is precisely communities that already have a high-level of social capital who are likely to successfully apply for the project. Hence a naïve comparison of social capital in project and non-project communities may well show social capital to be higher in the former, but not as a result of the project, but because having social capital makes selection into the programme more likely (see World Bank, 2002 and 2005, and Vajja and White, 2008 for further discussion). Again we are not comparing like with like.

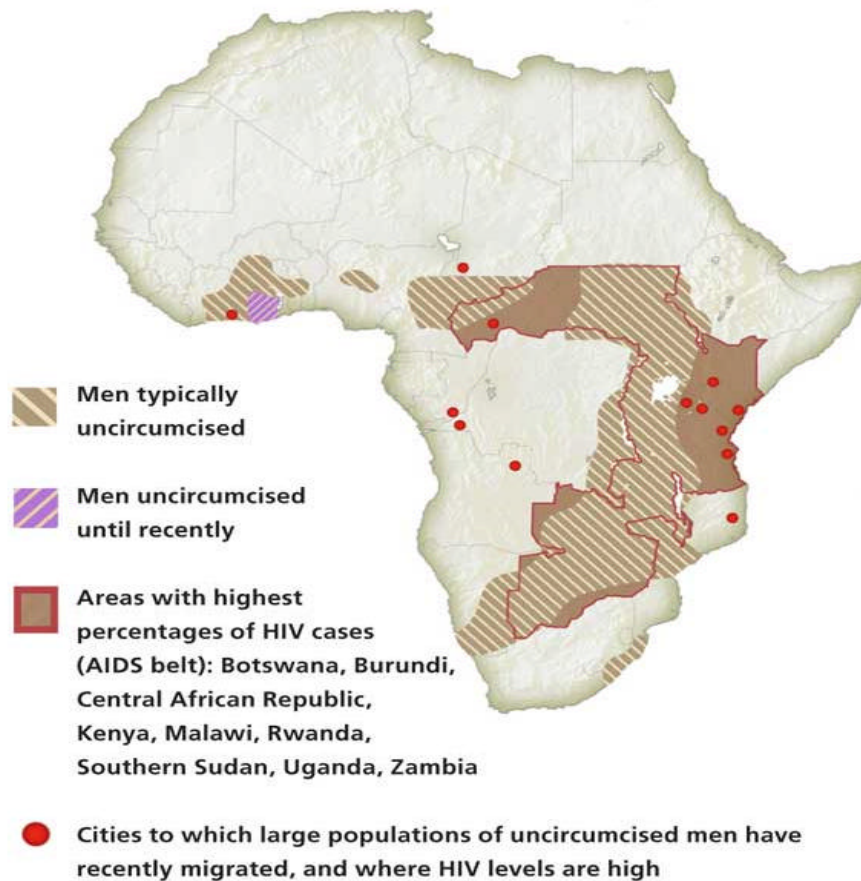
Selection bias matters, as shown here by three examples. Infant mortality in Bangladesh amongst children delivered in hospital is 115 per 1,000 live births (2004 data), compared to just 67 for those children not delivered in hospital (in Bangladesh most deliveries are at home). Does this mean that being delivered at hospital almost doubles the risk of premature death? No. Once again, we are not comparing like with like. Remembering that most deliveries are at home, which children are most likely to be delivered in hospital? It is children for which the mother was identified as having a high risk pregnancy, or for which complications arose during pregnancy so the mother was referred – both cases which are correlated with a higher risk of premature death. An accurate comparison would be with the mortality rate amongst children from high risk pregnancies or deliveries with complications who were born at home. We don't have that figure, but it would certainly be higher than 67 and most likely higher than 115.

Second, a study in Zambia examined whether keeping girls in school helped prevent teenage pregnancy. The author surveyed girls aged 18 both in and out of school, asking if they had experienced a pregnancy. She found higher pregnancy rates for those girls not in school, taking this finding as evidence that keeping girls in school does indeed reduce pregnancy. But this is not a valid conclusion. Why do girls drop out of school? One major reason for doing so is that they get pregnant, and so drop out either because of the stigma attached or simply because they have to look after the child. So the causation is, at least in part, from pregnancy to enrolment, not vice versa.

Finally, take a look at a map of Africa showing male circumcision rates, and impose on that data on HIV/AIDS prevalence (Figure 1). There is a very close correspondence between the two, with the exceptions being cities with large numbers of recent uncircumcised male migrants. One might therefore conclude that male circumcision reduces the chances of contracting HIV/AIDS, and indeed there are medical reasons to believe this may be so. But maybe some third, underlying variable, explains both circumcision and HIV/AIDS prevalence. That is, those who select to get circumcised have special characteristics which make them less likely to contract HIV/AIDS, so a comparison of HIV/AIDS rates between circumcised and uncircumcised men will give a biased estimate of the impact of circumcision on HIV/AIDS prevalence. There is such a

factor, it is being Muslim. Muslim men are circumcised and less likely to engage in risky sexual behaviour exposing themselves to HIV/AIDS, partly as they do not drink alcohol. Again we are not comparing like with like: circumcised men have different characteristics compared to uncircumcised men, and these characteristics affect the outcome of interest.

Figure 1 Male Circumcision and HIV/AIDS prevalence in Africa



Source: *Harvard Public Health Review*, <http://www.hsph.harvard.edu/news/hphr/infectious-diseases/spr08circumcisionmap/index.html> (accessed 21/10/10).

What to do about selection bias

The problem of selection bias is that the group subject to the intervention is systematically different from those not receiving the intervention. As stated above, those participating are not a random sample of the population. One way around this problem is thus random assignment of the programme, often referred to in the impact evaluation literature as the treatment. That is, those who get the treatment are randomly chosen from the eligible population, as is a control group of those who do not receive the treatment. This approach is the randomized control trial, or experimental approach. Note that the randomization is of

who gets to be in the project and who does not. It is not the same as taking a random sample of the project and non-project groups. The latter approach does nothing to address selection bias.

It is easy to see how randomization solves the problem of selection bias. The bias occurs because of systematic differences between the project and non-project groups. But if these two groups are drawn at random from the same underlying (sub-)population then the average characteristics must be the same. Any differences observed in outcomes must be attributable to the intervention. The two groups are identical except that one group got the intervention and the other did not.

Of course, statistics tells us that the two groups will have similar average characteristics provided we pick a large enough sample. If we just pick two people (or villages, or districts) and assign one to the project group and one to the control, then it is not that likely at all that they will be similar. Table 1 shows the average characteristics of random samples of women selected from the Zambian Demographic and Health Survey (2007). When we just take two women, the project woman lives in town, but the control in rural areas, and the former has a much larger household, older household head and more years of education than the latter. They are not very comparable at all. But it can be seen that these averages get closer as we increase the sample size. Once we are drawing a total sample of 2,000 women, roughly equal proportions live in rural areas (66 and 64 percent respectively), have same number of years of education (5.2 and 5.4) and so on.

Table 1: Average characteristics by different sample sizes (n)

	Rural (%)		Years of education		Number of household members	
	Treatment	Control	Treatment	Control	Treatment	Control
n=2	100	0	12.0	9.0	9.0	5.0
n=20	70	80	6.4	5.8	6.4	6.7
n=50	72	60	5.8	5.3	6.4	6.5
n=200	65	61	6.0	5.0	6.7	6.5
n=2,000	66	64	5.2	5.4	6.5	6.5

	Age of household head (years)		Literate (%)		Earth floor (%)	
	Treatment	Control	Treatment	Control	Treatment	Control
n=2	52	39	100	100	0	0
n=20	39	43	70	80	40	80
n=50	40	46	68	56	49	50
n=200	43	42	69	48	55	58
n=2,000	42	41	59	56	60	64

Source: Calculated from Zambia DHS (2007)

If a randomized control trial is not possible then a large n impact evaluation can instead be based on a quasi-experimental design, which uses statistical means to construct a comparison group, which, like the control group in a RCT, has the same characteristics as the treatment group.

The problem of selection bias is a problem of endogeneity. That is the right-hand side programme participation variable is a function of the outcome, either directly or through some mediating variables. Hence traditional statistical methods of addressing endogeneity, such as instrumental variables can be used to address the problem. These approaches hold other factors constant rather than creating a comparison group with similar characteristics to the treatment group.

An alternative is the approach of propensity score matching (PSM) in which a 'participation equation' is first estimated. This is either a probit with a dichotomous dependent variable, $Y=1$ for those in project, and $Y=0$ if not, or a multinomial logit if there are multiple treatments. The right hand side variables are variables expected to affect programme participation. The fitted values give the propensity score (probability of participating). The comparison group is made by matching treated observations with non-participants with the nearest propensity score, though dropping observations outside the region of common support; i.e. observations in treatment group with a propensity score higher than the score for any untreated observations, or observations in the untreated group with a score lower than any in the treated group.

PSM is preferred to instrumental variables, as the former does not require specification of the functional form of the outcome equation. However, both suffer from a problem of participation determinants which are unobserved or unobservable. Leaving these determinants out causes omitted variable bias. With randomization all characteristics are on average the same between treatment and control, both observed and unobserved.

If these unobserved characteristics do not change over time (time invariant), then panel data, i.e. data from before and after the intervention, can be used to difference them out using a double difference analysis. But if there are time varying unobservables then panel data will not help remove them. However, there is one quasi-experimental approach which can take care of unobservables, regression discontinuity design (RDD).

RDD can be used when there is an eligibility threshold to be admitted into the intervention, such as the poverty line, the score for a business proposal or a landholding threshold. Those households, firms or individuals just either side of the threshold are argued to be the same in terms of both observed and unobserved determinants of participation, so any observed difference in outcome can be attributed to the intervention. The external validity of the approach can be questioned since the impact estimate only applies to those at the threshold.

So, whilst there are alternatives to randomization for large n impact studies, these alternatives can be subject to various criticisms. Moreover, the simplicity of RCT designs makes them easy to present to policy makers: we took two identical groups and applied

intervention X to one and not the other, after which outcome Y has improved by x% more in the treatment group. Hence, where feasible, attempts should usually be made to implement a RCT.

Issues in implementing a RCT

Preparing for a RCT

Since an RCT relies on random assignment of the treatment, this will nearly always mean that the evaluation has to be designed *ex ante*, since it is extremely unlikely that assignment of the project would have been on a random basis. Since RCTs are currently fashionable, you may encounter cases of less well-informed managers asking for an impact evaluation of a completed project, adding “and make it a RCT”. It has to be explained that this is not possible.

Using random assignment means that the evaluation affects the intervention design, at least in the selection of treated areas within the eligible population. It is very important that the implementing agency, and other key stakeholders – notably politicians – buy into the design, otherwise you may find the design compromised. In the case of a prospective impact evaluation of health insurance in India, the staff of the health ministry told us very clearly that we could assign the intervention how we liked, but that the Minister was sure to change it. So there was no sense in embarking on a RCT.

Detailed discussions with the implementing agency are required to establish the level at which the programme will be randomized (school, community, household etc.) and to identify the eligible population across which randomization will be done. If randomization is across the pipeline (see below) then the timing of this phasing in needs to be agreed. And this timing needs to allow for a baseline survey to be conducted in treatment and control areas before the intervention reaches the field.

As will be seen below, many of the common objections to RCTs are based on misconceptions, so they can be countered if raised by the implementing agency. Indeed, the random element can be a selling point. In some Latin American countries, in which lotteries are common, conditional cash transfer programmes have been allocated through a public lottery, with a well known personality, such as a soap opera star, making the draw. The transparency of this process has appealed to local political leaders who cannot be accused of corruption or favouritism in the allocation of programme resources.

Designing the RCT

A first decision is the number of treatment groups to be included. A study which compares multiple interventions is of more use to policy makers than a study of a single intervention. Which has more impact on reducing teacher absenteeism: cash incentives or improving teacher housing? Answering that question is more useful than just looking at just one of these two interventions. But a separate treatment group is needed for each intervention. So with two treatments, A and B, three groups are needed, A, B and the control C. In some cases it may be argued to be unethical to withhold treatment from a control group. This issue is discussed in more detail below. However, here I note that in medical trials, the control

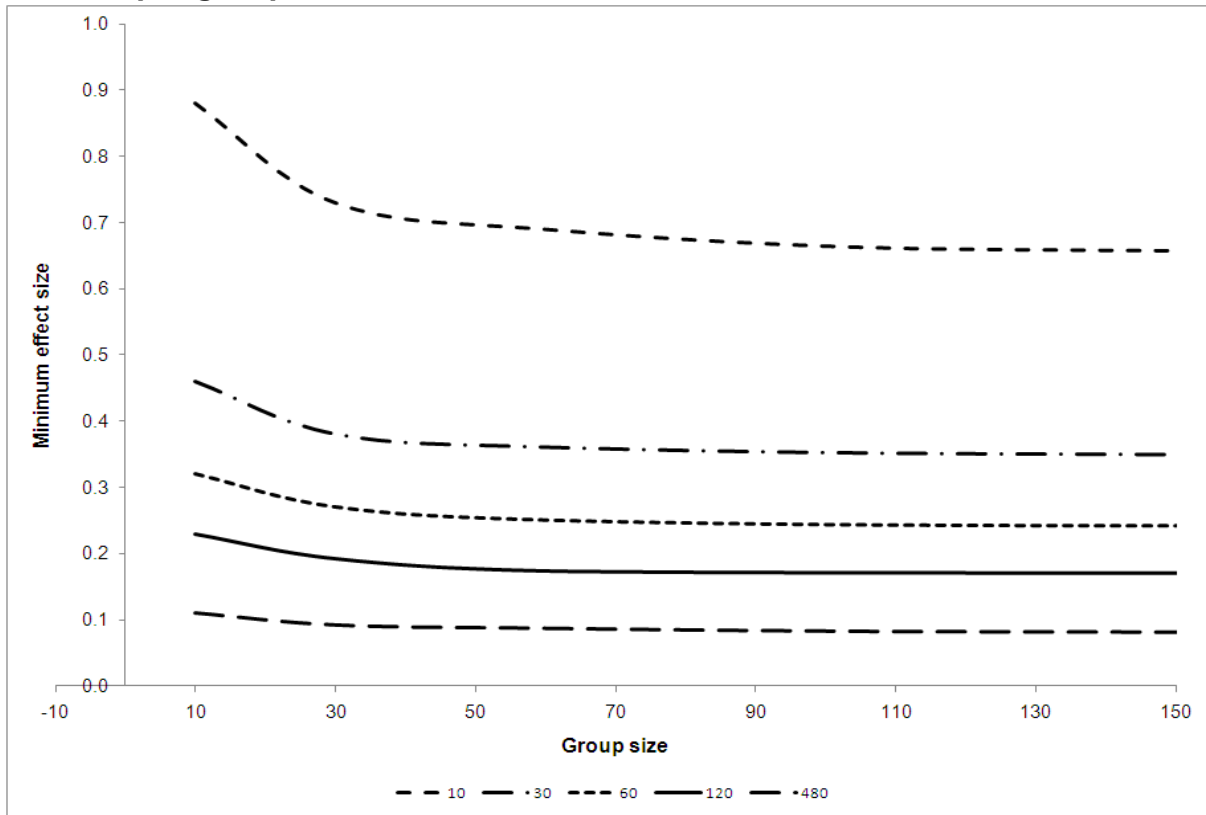
group often receives the existing treatment rather than no treatment, so the evaluation is of whether the new treatment has a larger impact than the existing treatment. If there is an established programme, it is this latter question that is of more interest to policy makers.

In the development literature it is often argued that interventions are complementary to one another, that is the impact of the two together is greater than the sum of the impact of the two provided individually. For example, business service training, or market access information, increase the impact of microfinance programmes; hygiene education increases the impact of improving the availability water supply and sanitation; and awareness raising amongst men will increase the impact of programmes to empower women through support for livelihoods activities. Or these different interventions may be substitutes, so the combined impact of the two is less than that of the sum. These effects can be examined with a factorial design, which has three treatment groups: A, B and a third group that receives both A and B. So, including the control, C, four groups are needed. There are limits to the extent to how many treatments, and combinations of those treatments, can be evaluated in one go since each new treatment adds to the sample size that is needed.

Once the number of treatment and control groups is decided, the next step is to perform a power calculation to determine the required sample size. Inputs into the power calculation are the number of treatment and control groups, the minimum effect size you want to observe and the corresponding level of confidence. Required sample size can be reduced by various methods of pre - matching including stratification, covariate matching and matched pair randomization. These methods help ensure the similarity of the treatment and control groups, and are often required to ensure this balance for small samples (though there is some debate on this point), but they do create a possibility of selection bias since matching is only on observables. It is usually the case that sample size should be the same for treatment and control groups (a balanced sample). If there are multiple treatments, the same control group acts for all, so the control group is the same size as that for each individual treatment.

The unit of randomization may not be the same as the unit of analysis, requiring a clustered design. For example, a treatment may be randomized at the school level, but the intervention takes place at the classroom level with outcomes measured in individual students. The standard errors must be adjusted for clustering. The main factor driving the power calculations is the number of clusters across which randomization occurs, not the total sample size. This fact is shown in Figure 2, which shows the minimum effect size that can be detected for different combinations of the number of clusters (each line in the figure corresponds to a given number of clusters, with the top line being that for 10 clusters). As shown in the figure, increasing the sample within each cluster much above 30 units does practically nothing to increase the power of the study. On the other hand, increasing the number of clusters, especially for low numbers of clusters, has a very marked impact on power.

Figure 2 Minimum effect size as a function of number of clusters sampled and sample group size



Source: derived from Bloom (2006: 21)

So if a manager says you can save time and money by going to half as many clusters but doubling the sample in each area, and that will be the same sample size, he or she is wrong. Such a step would drastically reduce the power of the sample.

Figure 2 shows the case of a simple randomization, in which information from the baseline, or other sources, has not been used to improve the match. Using information on covariates can improve precision. Suppose the desired minimum effect size is 0.4. With covariates the required sample size is close to 30 groups, where it is not much more than 10 when covariates are taken into account in assigning the treatment (Bloom, 2006: 21).¹ A better matching can also be obtained by pre-matching, i.e. taking repeated random samples after the baseline to get one in which treatment and control are balanced on key variables.

The above information should all be recorded in a study protocol. In the medical field, most journals require that the protocol has been published before the study commenced as a pre-requisite for publication of the findings. The same approach has not been adopted in the development field to date, despite its advantages, namely peer review of the proposed study design and reducing the scope for data mining or selective reporting of findings. The

¹ This example assumes that the covariates predict 60 percent of the unexplained variation in the outcome variable.

protocol should also describe the eligible population and how they were identified (administrative data, listing and so on), and the procedure used to randomly assign the treatment. This procedure should rely on random numbering from a random number table or generated by a computer. 'Pseudo-randomization' such as using alphabetical lists can produce systematic biases.

The design needs to allow for possible impact heterogeneity from a number of sources. The assumption is that the treatment is homogenous, which seems reasonable for medical trials in which people take a pill, but is less so for development interventions in which capacity to implement may vary greatly, or may vary according to contextual factors such as accessibility. It is harder to get staff to go to remote areas, or to stay there if they are initially enticed. Or impact may vary according to beneficiary characteristics: younger children respond more to feeding programmes and with greater impact on cognitive development, the better off are more likely to benefit from microfinance as they have the resources (land, labour, vehicles etc.) required to utilize the loan productively and so on. Or impact can vary according to context: a school feeding programme can increase student alertness and so learning outcomes in a well functioning school, but will be of no use if teachers are absent. Such heterogeneity can be captured by sub-group analysis. The power calculations need to allow for the intended sub-group analysis which will be done.

The standard in medicine is that all intended sub-group analysis must be recorded in the protocol beforehand. The reason is to prevent data mining. Chance will throw up some significant relationships, if you try enough sub-groups ('this drug works if administered on a Thursday to people with a d in their name'). I am a bit ambivalent about transferring this practice to the analysis of development interventions. I come from an exploratory data analysis tradition, in which the analyst's job is to seek explanations consistent with the patterns in the data rather than impose a model or theory without reference to those data. Hence it is possible that sub-groups may only emerge as the evaluation proceeds, from engaging with either quantitative or qualitative data. So I would argue that additional sub-group analysis can be added if it is well supported by other data or arguments as to why it is a meaningful sub-group to be analyzed.

The design should also identify and capture spillover effects. In the best known case in development literature, deworming selected children will have beneficial effects on children in neighbouring households (Miguel and Kremer, 2004). More complicated is if there are possible spillovers into the intended control groups, say by word of mouth for information campaigns, labour market effects for public works programmes and so on. Such spillovers can be reduced by using a list of eligible clusters which are not contiguous. But doing so means they are further apart, so the quality of the match may well be poorer for all sorts of reasons, especially in smaller samples. There is thus a trade-off between being close and far, and one that has to be determined on a case by case basis depending on the likelihood of such spillovers and the heterogeneity of the eligible population.

Conducting the RCT

The RCT begins with a baseline. It might be thought that since randomization ensures similarity of treatment and control then one can simply compare the difference in outcomes

at endline. But statistics tells us randomization will not always result in well matched samples, so we do need check for the quality of the match. And even if it's fine, it's not perfect. So a double difference estimate will always be preferred. Besides which there are other sorts of data we may require for other aspects of the evaluation for which the baseline will prove useful.

The randomization protocol should state how refusals as well as cross-overs (those in the control getting the treatment) are to be treated. Once the intervention starts, a record should be made of refusals and cross-overs.

A great threat to the integrity of the RCT design is the danger of contamination, that is, that the control group receives an intervention which affects the outcomes of interest.² In Nicaragua the control group were given a programme by the local governor precisely because they were not receiving the treatment, and in Andhra Pradesh the donor went ahead and scaled up an HIV/AIDS programme before the pilot was finished, thus contaminating all the controls (Samuels and McPherson, 2010). It is unlikely that contamination can be prevented. Data must be collected to know whether contamination has occurred or not. If contamination is universal across treatment and control, the RCT is measuring impact in the presence of that intervention. If contamination is restricted to the control, and is universal, then the RCT is comparing the two interventions. If contamination is partial then the contaminated clusters can be dropped, or subgroup analysis conducted if sample size allows.

Objections to RCTs

The use of RCTs to evaluate socio-economic development interventions, both in the developed and developing world, has been controversial. This section reviews the objections.

A first objection is simply the idea of 'experimenting on people' as suggested by the name experimental design. But all new policies are essentially experiments. We try a new policy and then decide to continue it or not hopefully based on evidence of how well it works. So, unless we are committed to never trying out new policies or programmes, then this particular argument against 'experimental designs' does not have much merit. The stronger argument concerns the ethics of having an untreated control group.

Is it right to withhold the treatment from a part of the eligible population? There are several justifications for doing so:

1. We actually don't know if the programme works or not, that is why we are evaluating it. For example, there may be unanticipated adverse side effects. Withholding an ineffective or even harmful programme is not unethical.
2. It is very rarely the case that a programme is extended to the whole eligible population on day one. For budgetary reasons, the implementing agency, especially

² Spillover effects which affect the control group are a special case of contamination which were discussed above.

NGOs, may only intend to ever treat a proportion of the eligible population. Or for logistical reasons, the intervention may be being rolled out over time, so there will be a untreated population for at least some months, possibly two or three years. Hence the order of treatment can be randomized, that is 'randomization across the pipeline'. The best known RCT in the developing world – the evaluation of the Mexican conditional cash transfer programme, Progresa - adopted this approach. In the initial phase, the programme was a pilot programme for 506 communities, just half of which received the programme at first, the other half acting as a control group for two years. So in most cases there is anyway an untreated section in the eligible population, at least temporarily, and the evaluation is just exploiting that fact for the purposes of assessing the impact of the programme.

3. The really unethical thing is not the withholding of the programme, perhaps temporarily, from some group. The really unethical thing is the spending billions of dollars each year on programmes that don't work. And without rigorous impact studies, including RCTs, we won't know if they work or not. The sacrifice of having a control group is a small one compared to the benefits of building an evidence base about effective development programmes.

I find the above arguments quite compelling. But they are not complete. It is not that we are just leaving the control group untreated, we are going into these areas and collecting data, but giving them nothing at the end if it is not a pipeline randomization. It is all very well, and easy, to say that the sacrifice is worth it, but it is not our sacrifice. I believe the ethical issues involved here have received insufficient attention amongst RCT practitioners. The fact is that outsiders entering a community, especially foreigners, raise expectations. Those expectations must be managed. The usual line of 'this research will not benefit you directly but will benefit people like you' may be insufficient to ensure cooperation. Remuneration for taking part should not be ruled out. There are, however, two problems. One is that providing remuneration will create an incentive for local people to influence sample selection. The second is that the remuneration may have an impact on the outcomes of interest, thus biasing the impact estimates. Both of these problems can be addressed by making the contribution at the community level - \$200 for the village development fund, exercise books and pencils for the school and so on – and doing so at endline only. Study budgets should make a provision for such ex-post incentives. Having said that, in our study in Ghana the enumerators typically gave the respondent the pencil they had been using at the end of the interview, which generally made them happy and is on a scale unlikely to bias the findings.

A second objection to RCTs is that they are expensive. They are expensive because they involve primary data collection. But they are no more expensive than any other study requiring data collection on a similar scale. Indeed quasi-experimental designs (PSM and RDD) require throwing out parts of the data, so can prove more expensive.

A third objection is that RCTs are not really feasible for development programmes. As explained above, quantitative impact evaluations are only feasible for large n interventions. However, a RCT is not feasible for all large n interventions either for technical reasons (the study is being done *ex post*, it is a national programme and so on) or for political ones (it is

not possible to get stakeholder buy in to randomization assignment of the programme). Some years ago it was suggested that perhaps 5 percent of aid money could be spent on programmes which are amenable to RCTs. Whilst that is not a lot of the aid programme, it is still quite a substantial number of RCTs. And given that practically none had been conducted up to that date, it was an argument for doing more. But in the intervening years we have seen RCT designs being used to evaluate programmes in a wide range of sectors. Whilst the largest share of studies are still in health and education, there have also been RCTs of interventions in climate change, governance, women's empowerment, micro credit and access to finance, and so on. It remains the case that there are of course some things that cannot sensibly be evaluated with a RCT, but it also remains the case that there are many, many opportunities for further learning about what works from such studies.

The fourth criticism is that 'what works?' isn't the right question, or it is at best only part of the question. At 3ie, this question is made into three: what works, and why, and for how much? So-called 'black box' impact evaluations which don't seek to unpack the causal chain, to understand why a programme does or does not work in a particular setting are of far less benefit to policy makers than those that do. As I have elaborated elsewhere (White, 2009 and 2011), answering the why question means drawing on a broader range of data and approaches – but the rigorous analysis of impact is a crucial part of the design. We are seeing increasing attention to the underlying theory of change of the intervention by impact evaluation researchers. It is also far more useful to know at what cost the improvement in outcomes has been achieved. Cost effectiveness analysis, or cost benefit analysis when there are multiple outcomes, does not at present feature in RCT design as frequently as it should.

A fifth objection from programme staff is that they don't want to assign the programme at random, as they want to target it. This objection is a mis-understanding. Randomization is done across the eligible population, not the population as a whole. Going back to an earlier example, there have been randomized control trials of the impact of male circumcision on HIV/AIDS transmission in Kenya, South Africa and Uganda. But the researchers did not pluck names out of the phone book and go around with a pair of scissors. The trial was advertised, and those registering with the project assigned to a "treat now" group and a "treat in two years" group.³

RCTs of development interventions are criticized as they don't attain the triple blinding ideal of medical trials: blinding of the treated as to if they are in treatment or control, blinding of the person delivering the treatment as to whether it is the treatment or a placebo, and blinding of the researcher analyzing the data (Scriven, 2008). The first two kinds of blinding are clearly not possible, but the third is. To my knowledge it is not practised in the analysis of development interventions, but it should be. The other two issues deserve more attention than they have received to date. To add to the research agenda to the possible biases from non-blinded trials should be added investigation of placebo and Hawthorne effects. As an example of the latter, the fact that data collection can raise awareness of the

³ In South Africa the intervention was found to be so effective that the trial was ended prematurely and the control group treated early.

issues being addressed by the intervention amongst the control group and so cause a change in behaviour in that group.

Finally, RCTs are said to have limited external validity as they are often small scale trials run on a resource-intensive basis often with foreign researchers or their students running the intervention. The impact will be quite different once the programme goes to scale using local implementation agencies and, probably, a lower level of resources. This argument also has some validity. Attempts should be made to ensure that the experimental pilot is as close to how the programme will be in the scaled up version as possible.

Summary

RCTs have become popular in the development community as part of the response to the results agenda. These studies are well placed to address the question of which programmes work or not. And, properly designed, they can be embedded in a broader evaluation design which also addresses questions of why an intervention works in a specific context or not, and at what cost. Usually the appropriate design will not be a simple randomization drawn from an eligible list. So there are several decisions to be made in designing a RCT, and these decisions should be recorded in a study protocol which is placed in the public domain before the study begins. Many of the criticisms of RCTs can be responded to. But some require more reflection and research and have implications for the conduct of RCTs.

References

- Bloom, Howard (2006) 'The Core Analytics of Randomized Experiments for Social Research', *MDRC Working Papers on Research Methodology*,
- Duflo, Esther, Rachel Glennerster and Michael Kremer (2006) 'Using Randomization in Development Economics Research: A Toolkit', Department of Economics, Massachusetts Institute of Technology and Abdul Latif Jameel Poverty Action Lab,
<http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf> (accessed 23/10/10)
- Miguel, Ted and Michael Kremer (2004) 'Worms: identifying impacts on education and health In the presence of treatment externalities' *Econometrica*, **72**(1): 159–217.
- Vajja, Anju and Howard White (2008) 'Can the World Bank Build Social Capital? The Experience of Social Funds in Malawi and Zambia', *Journal of Development Studies*, **44**(8): 1145-1168.
- Samuels, Fiona Samuels and Sam McPherson (2010) 'Meeting the challenge of proving impact in Andhra Pradesh, India', *Journal of Development Effectiveness*, **2**(4).
- Scriven, Michael (2008) 'A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research' *Journal of MultiDisciplinary Evaluation* **5**(9): 15-24.
- White, Howard (2009) 'Theory-based impact evaluation: principles and practice', *Journal of Development Effectiveness* **1**(3)
- White, Howard (2010) 'A Contribution to Current Debates in Impact Evaluation' *Evaluation*, **16**: 153-164
- White, Howard (2011) 'Achieving high quality impact evaluation design through mixed methods: the case of infrastructure', *Journal of Development Effectiveness* **3**(1) (forthcoming)
- World Bank (2002) 'Social Funds: an evaluation' Washington D.C.: OED, World Bank.
- World Bank (2005) 'The Effectiveness of World Bank Support for Community-Based and Community-Driven Development', Washington D.C.: OED, World Bank.