International Initiative for Impact Evaluation
Working Paper 9

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

Howard White
January 2013

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 responds to demands for better evidence to 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. The paper was first published in February 2011. This is the second edition.

Photograph: Christelle Chapoy
3ie Working Paper Series Production Team: Mukul Soni, Rajesh Sharma and Radhika Menon

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, and to Yasmin Khalafallah for assistance in preparation of the final version. The usual disclaimer applies. Email: hwhite@3ieimpact.org.

1. Introduction

The focus on results has been 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 amongsst economists and political scientists, of a range of approaches which do directly tackle this question of what difference an intervention has made, that is, its impact. 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 and ethical concerns in designing and implementing an RCT are then discussed before moving on to some of the criticisms which are commonly made of this approach, and, finally, challenges for practitioners of RCTs.

2. What is impact evaluation?

Within the development community many think of 'impact' as meaning long-run effects. This usage is contained in the DAC definition of impact,[1] 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 program made: did it improve lives, save lives even? Impact evaluation is a 'with versus without' analysis: what happened with the program (a factual record) compared to what would have happened in the absence of the program (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, program 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 programs supported by donor funds either directly (project aid) or indirectly (program aid) should clearly play an important role in addressing the issue of that agency's contribution to development.

[1] The DAC definition of impact is: 'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended' (OECD-DAC, 2002).

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 (households, schools, clinics, firms, villages, districts or whatever) then statistical analysis is the most appropriate means of constructing a counterfactual. Where n is not large then the most appropriate methods are deductive approaches based around causal chain analysis, such as process tracing; see White and Phillips (2012) for a discussion of small n approaches.

For large n impact evaluation designs, 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 that of how to identify a suitable comparison group.

3. The problem of selection bias

The problem of selection bias arises because program participants are not a random sample of the population as a whole. Rather those in the program are selected through both program placement and self-selection. Program 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 programs. They do so voluntarily, and those choosing to participate may have different characteristics to those who do not do so.

Problems occur if the factors affecting whether a group or individual participate in a program 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 program 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.

As an example of selection bias from program 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.

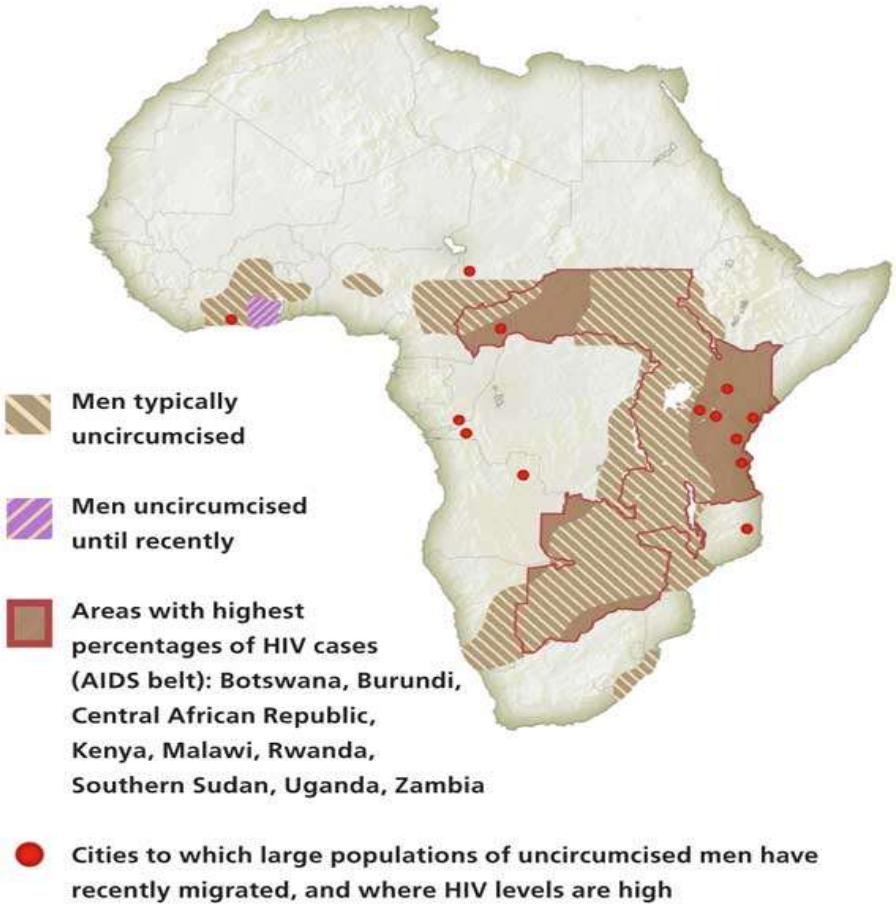
As an example of selection bias from program placement, consider a community-driven development intervention, such as a social fund. Communities make a proposal to the district administration for funds, to be managed by a community level committee, to undertake a project such as build or renovate the school or clinic, or build 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 to 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 program more likely (see World Bank, 2002 and 2005, and Vajja and White, 2008 for further discussion; similar findings emerge from more recent studies with respect to social cohesion, see King et al., 2010, Casey et al., 2011, and Humpreys et al., 2012). 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. Most deliveries are at home. So 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 survey 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 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 changes 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 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).

4. What to do about selection bias

The problem of selection bias is that the group subject to the intervention is systematically different to 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. Those receiving the programme are often referred to 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 of the two groups 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 and 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 program 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 program 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. In interpreting the impact estimate it need be remembered that this estimate only applies to those at the threshold. So the estimate cannot be used for a general calculation of cost effectiveness.

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.

5. Issues in implementing a RCT

Preparing for a RCT

Since a 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 (there are rare cases, such as school voucher programmes, in which assignment is often random). And some programmes allow for natural experiments, which applies the same analytic methods, but does not require random assignment.

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. Or they may also ask for an experimental design, but say, 'not a RCT as they are expensive'. Experimental designs are RCTs, there are no alternative experimental designs.

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 program 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 programs 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 program resources.

Designing the RCT (treatment arms, power calculations and all)

All impact evaluations should adopt a theory-based design employing mixed methods (White, 2009, 2010a and 2011). But I focus here on design issues specific to randomization.

The key questions in designing an RCT are:

1. What treatment is being tested?
2. How many treatment arms will there be?
3. What will be the unit of assignment?
4. How large a sample do I need? (which depends on the design the RCT)
5. How will I randomize?

What treatment is being tested? The treatment being evaluated needs to be clearly defined. It may be a very straightforward 'single component' intervention such as water chlorination or building a road. But most interventions have multiple components. Water chlorination comes with some institutional structure to ensure technical and financial sustainability, any training necessary to support these structures, and information to intended users on how and why to chlorinate.

It is not a problem for a RCT if the intervention being delivered is a multi-component one. But, for a single treatment arm study, it does need to be ensured that the same treatment is being applied in all project areas in the same way.

But it is also possible to break the components down to make a multi-treatment arm study, which may prove more useful in informing programme design.

How many treatment arms will there be? 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? Or as another example, one treatment arm will get, say, supplementary feeding to tackle child malnutrition. The second treatment arm could have nutritional counselling (for an example of such a design evaluating World Vision programmes in Haiti see Ruel et al., 2008). These are multiple treatment arm studies, in which a separate treatment group is needed for each intervention. But the same control group acts for all so the control group is the same size as that for each individual treatment.

Should we have a control group? Just having two treatment arms will let us compare which of the two treatments is most effective. If we also have a no treatment control group arm we can also measure the absolute impact and cost effectiveness of the two treatments. So with two treatments, A and B, three groups are need, A, B and the control C.

There may also be cases for multiple untreated arms, for example if there are expected spillover effects. In the best known case in the 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 programs and so on. In this case one control arm are those not directly receiving the treatment but who might experience spill over effects, which may also be considered an 'indirect treatment' arm – for example non-beneficiaries in beneficiary communities, provided there is random assignment within the community. The 'pure control' should be a group which will be free from spillovers.

Obtaining a pure control is best done 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.

In some cases it may be argued to be unethical to withhold treatment from a control group. But the control group need not be a 'no treatment' control group. Indeed, in clinical trials the norm is that the control group receives the existing standard of care. Heart patients are not left unmedicated which would clearly be unethical. Hence, clinical trials as usually multiple treatment arm studies in which one arm gets the existing treatment. With such a design the evaluation question being answered is whether the new programme works better than the existing programme. This question is usually the one of interest to policy makers. They are less interested to know whether the new programme is better than doing nothing at all.

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 programs; hygiene education increases the impact of improving the availability water supply and sanitation; and awareness raising amongst men will increase the impact of programs 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 sum. These effects can be examined with a factorial design (a special case of multiple treatment arm 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.

What will be the unit of assignment? The unit of assignment is the level of which random assignment takes place. It may not be the same as the unit of analysis. In a simple RCT, the two are same. Random assignment takes place at individual level, and the individual is the unit of analysis. Some interventions, notably vouchers, allow randomization at the individual level.

But mostly individual level randomization is not possible either for logistical reasons or because of the problems of giving to one person and not their neighbour. Hence most development impact evaluations are cluster-RCTS. The unit of assignment is a cluster, each

cluster containing more than one unit which will receive the treatment. 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. Common clusters are villages (e.g. Gram Panchayats in India), sub-districts, blocks in urban areas (e.g. barangays in the Philippines), community groups (e.g. co-operative associations) or schools.

Clustering is usually more feasible and reduces the logistical costs of data collection. But the standard errors need to be adjusted for clustering: they will be larger than they would be had the same number of observations been collected through simple random assignment. Hence the sample for a cluster design needs to be larger than that for a simple RCT. This point is elaborated below in the discussion of power calculations.

How large a sample do I need? Power calculations are performed to determine the sample required to detect an impact if there is one. The main determinants of power are:

- Sample size: the larger the sample, the greater the power, though power is reduced if that sample is split into clusters

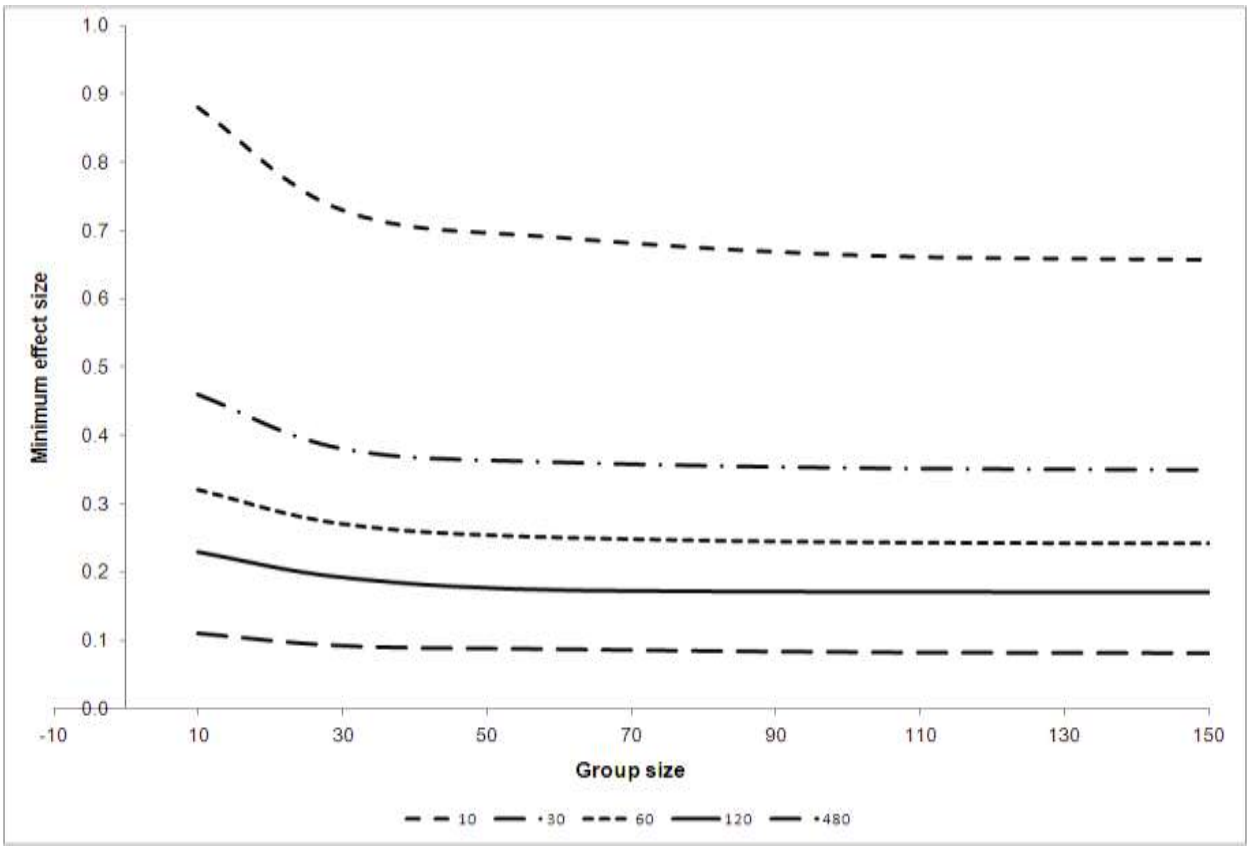
- Confidence required: the norm is 95 percent, which is perhaps rather high, having it so high increases the chance that we will conclude there is no impact when in fact there is one;

- Minimum effect size: the smallest effect policy makers would expect to see from the intervention, a smaller sample is required if the minimum effect size is large (you need less observations to capture a large effect); and

- Intra-cluster correlation coefficient (ICC): a measure of how similar the units are within each cluster. It is because units within a cluster are similar to each other that a larger sample is needed for cluster RCTs than simple RCTs. Hence the larger the intra-cluster correlation coefficient the larger the sample you need. Ideally data for the ICC come from the population of interest, which may require a pilot survey. However, it is common to get the coefficient from existing surveys of similar populations. Or just to assume ICC= 0.2, which is really not to be recommended.

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. The power of a study is greater the smaller the effect size it can detect. Each line in Figure 2 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 same sample size, he or she is wrong. Such a step would drastically reduce the power of the sample.

Once the number of treatment groups is decided, and if to have a 'no treatment' control, the next step is to perform a power calculation to determine the required sample size. As indicated above, 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. Consider the case of matched pair randomization. A cluster-RCT is being conducted of an education project in India using 60 schools, 30 treatment schools and 30 controls. Suppose just two of the schools are in areas in which the population is predominately tribal. It is very likely education dynamics are different in these two communities to the other villages. With straightforward randomization it is quite likely that these two villages could be? both in either the treatment or control. Matched pair randomization forms 30 pairs of villages based on observable characteristics, and then randomly assigns one of the pair to the treatment

group. So the two tribal villages could be one pair, ensuring that one is in the treatment group and the other in the control. More formally speaking, the matching ensures that the sample is balanced, that is that treatment and control groups have the same average characteristics.

So it appears that these methods help ensure the similarity of the treatment and control groups, and are often required to ensure this balance for small samples. There is some debate on this point, but in defence of pre-matching see Imai et al. (2009).

It is usually the case that sample size should be the same for treatment and control groups (a balanced sample). For multiple arm studies, each arm will usually be the same size.

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. As just explained, 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, compared to not much more than 10 when covariates are taken into account in assigning the treatment (Bloom, 2006: 21).[2]

How to randomize? The first step is to define the eligible population, which should of course be done by the implementing agency. Randomization requires a list of eligible units at the level of assignment of sufficient size to obtain the sample size required for the study.

There are at least three ways in which the eligible population can be used for random assignment:

- Simple randomization allocates units to treatment and control arms
- Pipeline randomization: all units will receive the project but over time, so it is the time of entry to the programme which is randomly assigned. The best known example of this approach is Mexico's conditional cash transfer, Progressa. In the initial phase the program was a pilot program for 506 communities, just half of which were received the program at first the other half acting as a control group for two years. The communities were randomly allocated into the two groups to receive the programme in years one and three. Those receiving the programme in year three served as a control group for two years.
- Raised threshold randomization: expands the eligible population and randomly allocates within that group. This technique is less common and so explained at greater length below.
- Encouragement designs, which are used for universally available but not adopted programmes and policies. The treatment group is provided with an encouragement to take up the intervention but this encouragement should not affect the intervention (see Gertler et al, 2010: 69-79).

A raised threshold design expands the eligible group in some way with no, or minor, modifications to the programme targeting mechanism. For example, the evaluation of

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

vocational training centres in Colombia asked each training centre to identify not 25 students from the 100 or so applicants from the next course, but 30 (Attanasio et al., 2011). The study team then randomly picked 25 of those 30 to be in the programme, with the five not selected being in the control group. There is virtually no change in programme design here: the centres still admit 25 students per course, and on average over 80% of those (21 out of 25) are students they would have picked in the absence of the evaluation. As far as the students are concerned, 25 get accepted and the rest are rejected. And the vast majority of those rejected would have been rejected in the absence of the programme. So the programme is hardly affected at all, but a valid control group has been created through the raised threshold.

A variation on the above design would be to ask the centres to identify 20 they want to enrol, and the next 10, picking five at random from the last 10. This design variation ensures that the centres for sure get the better students, and that, on average 22.5 of the 25 enrolled would have been enrolled in the absence of the evaluation. But the impact estimate is different than the earlier design, as it is measuring the impact on 'marginal students' not on the best students.

Raised threshold design can be applied, as in the above example, by raising the eligibility threshold such as a credit scoring, entry grade or poverty line. For example, if a programme is targeted at the 60 poorest schools in a province, make a list of the 100 poorest, and pick 60 at random to receive the treatment. Or, pick the worst 30 schools to definitely be in the programme, and then 30 of the remainder to also be in it.

The same idea may also be applied geographically. So if a programme is to be conducted in 30 villages, pick 60 villages, and randomly allocate the programme to half of those villages.

The stage of how to randomize is to actually allocate units of assignment to the different arms. Assignment is most simply done by assigning each unit a number and using a random number generator, such as in Excel or Stata. Sometimes a lottery is held, which may be done in public to increase transparency. 'Pseudo-randomization' is not a valid basis for a RCT, for example using alphabetical lists can produce systematic biases.

Reporting study design

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 is now being recommended in the development field (e.g. Rasmussen et al., 2011). J-PAL have a 'hypothesis registry' for RCTs, and 3ie is putting in place a registry for socio-economic development impact evaluations using both experimental and quasi-experimental designs.

Registries allow for peer review of the proposed study design (though the J-PAL registry does not do this), and reduces the scope for data mining or selective reporting of findings. The 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.

The design needs 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 programs 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 program can increase student and 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 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 meaningful sub-group to be analyzed.

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 are to be treated, and well as cross-overs, those in control getting the treatment. 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.[3] In Nicaragua the control group were given a program 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 program before the pilot was finished, thus contaminating all the

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

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. For example, in China, eye glasses were distributed to secondary school students who needed them, helping to improve their test scores. However, endline data showed increased use of eyeglasses in six comparison townships (Glewwe et al., 2012). Discussions with project staff showed that the doctors doing the eye tests had glasses left over from the treatment townships, so gave them away in the comparison communities. The study had used a matched pair randomization design and so was able to drop the pairs with contaminated controls with no risk of bias. This example shows the importance of collecting appropriate data along the causal chain in both treatment and comparison areas.

6. 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, programmes and projects 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 programs, 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 program 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 program is not unethical.

2. It is very rarely case that a program is extended to the whole eligible population on day one. For budgetary reasons the implementing agency, especially 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 the 'randomization across the pipeline' described in the previous section. 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 program.

3. The really unethical thing is not the withholding the program, perhaps temporarily, from some group. The really unethical thing is the spending billions of dollars each year on programs 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 programs.

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 is, but it is not our sacrifice. I believe the ethical issues involved here have received insufficient attention amongst RCT practitioners, as indeed have those of ensuring that the treatment group receive a genuine intervention which is not merely of academic interest. 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 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 programs. 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 program and so on) or for political ones (it is not possible to get stakeholder buy in to randomization assignment of the program). Some years ago it was suggested that perhaps 5 percent of aid money could be spent on programs which are amenable to RCTs. Whilst that is not a lot of the aid program, 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 programs 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.

Moreover, RCTs need not be that disruptive to the design of programmes as is often believed. Various designs make very small differences to programme implementation. Pipeline randomization reaches all the planned eligible population, just affecting the order in which they are reached. Raised threshold designs barely affect programme design or beneficiaries at all. Encouragement designs, which are used for universally available but not adopted programmes and policies, do not affect the programme at all (see Gertler et al, 2010: 69-79).

The supposed disturbance caused by the impact evaluation can also be less than imagined as programmes often have larger coverage than is required for evaluation purposes. Impact evaluation can be made more palatable by emphasizing that only a small part of the programme is affected by the planned evaluation. For example, consider a programme will be rolled out to 500 communities over five years at the rate of 100 a year. Suppose that power calculations show that only 60 communities are needed for the impact evaluation design, 30 treatment and 30 control. Hence the programme team can be told that for just over 10 percent of the communities you want to randomize the order in which communities get the programme, picking 30 at random for the first year and 30 at random for the last year. For the vast majority of communities, including all those getting the programme in years two, three and four, there is no change to the programme.

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 program 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 program staff is that they don't want to assign the program 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 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 the phone book and go round 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.

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 practiced 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 of data collection can raise awareness of the 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 program 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 like to how the program will be in the scaled up version as possible. This issue should not be taken likely, as is the first of the challenges identified in the final section.

7. Challenges for RCTs

The use of RCTs in development evaluation has made considerable strides in recent years. But there remain challenges which need to be addressed if these studies are to maintain their credibility and acceptability. Many of these issues are lessons learned from managing the first years of 3ie's impact evaluation programme.

The first challenge is to make impact evaluations more like evaluations and less like research studies. Too many impact evaluations focus on a research question of interest to the authors whilst ignoring evaluation questions of interest to policy makers. This issue affects choice of intervention to be evaluated, evaluation design and the way in which the findings are reported. The focus should be on evaluating real-life programmes, not on interventions designed by researchers. Indeed, as Sherman et al. (2002:6) point out, even the focus on evaluating programmes is a limitation as the vast majority of government resources are spent on 'practices' not programmes. Practices are the on-going routine activities we take for granted. But an evidence-based approach means nothing should be taken for granted.

Contrary to what many of the new generation of economists conducting impact evaluations seem to believe, the development community does have considerable experience in implementing programmes which needs to be built upon not ignored. The design of the evaluation should address evaluation questions of interest across the causal chain, using both counterfactual and factual analysis (White, forthcoming). And the report should address those questions, not only an academic research question. Researchers are of course free to produce separate, academically-oriented papers in which much of the evaluation material may not be included. But an evaluation of the programme is also required.

The most common example of this problem is that of participation rates. Many interventions, notably micro insurance programmes, suffer from low participation rates, that is intended beneficiaries are not terribly interested in the intervention. To researchers this fact is often treated as a technical problem. They use the estimate of the intention to treat effect (the average impact on all people targeted by the programme, whether they participated or not) as a means to get at the treatment of the treated effect (the average

impact on those who actually participated). The latter has been referred to the 'true impact'. I don't care much for semantics, but this 'true impact' if this is what you wish to call it, is pretty irrelevant if fewer than 10 percent of intended beneficiaries want to take part in the programme.

The best way to present results, which is done in too few studies, is to use cost effectiveness analysis. This approach would thus include the costs of trying to reach the whole targeted population, any additional costs for those who actually participate, divided by total benefits, which of course only come from those who did participate. A policy maker is unlikely being told simply that 'this intervention can boost girls' enrolment by 20 percent'. They will ask 'what will it cost to do that?', and 'is it cheaper than other ways of reaching the same objective'. So any policy relevant impact evaluation will report cost effectiveness; or cost benefit analysis where there are multiple outcomes.

Another advantage of using cost-effectiveness or CBA is that it shifts attention from the statistical significance of a coefficient to its importance, that is the magnitude of the effect. As Ziliak and McCloskey (2008) argue at length, the t-statistic has gained unwarranted attention as the sole measure of the extent of a relationship between two variables, but tiny effects can be significant. Most policymakers, or indeed researchers, cannot meaningfully interpret the policy importance of a regression coefficient. Hence the need to convert the coefficient to some more meaningful metric. Cost effectiveness does this.

But just because an intervention is cost effective may not mean it should be taken to scale or replicated elsewhere. What we are doing with scaling up in evidence-based development is social engineering, a term which has a bad name. We need to learn from history and be more cautious about the success we will gain when going to scale.

It should not be scaled up unless we can be confident that implementation at scale will be the same as that of the smaller pilot which was subject to impact evaluation. Researcher driven interventions are particularly prone to this problem, both because they rely on young researchers to support implementation rather than the local NGO staff who will be responsible for implementation at scale, and because there may be little ownership by the local NGO. Buying an NGO's services to implement a programme is not the same as ownership.

The intervention should perhaps not be replicated elsewhere because of external validity concerns. It is sometimes argued that RCTs have weak external validity. Their external validity is not inherently weaker than that of any evaluation of a single intervention, and their strong internal validity at least makes a basis for making evidence-based inferences. A good theory-based design will help understand the context in which a programme has or has not worked, and so help with these inferences as to where the programme might sensibly be replicated.

But the stronger response is that we should not rely on the findings of a single programme alone. In natural sciences and medicines, new study findings are replicated by other scientists as a means of testing their validity. 3ie funds a replication programme, which tests the accuracy and robustness of study findings using the original study data. Equally

important, and stressed by Karlan and Appel (2011), is the need to carry out many studies of similar programmes in different settings.

These findings should be summarized in a systematic review (see Waddington et al., 2012, for a guide to systematic reviews in international development). The strength of a systematic review is that it summarizes all available rigorous evidence. The recent controversy over deworming illustrates the case for reviews. The deworming study of Miguel and Kremer (2004) has been used to support an expansion of deworming on the basis of its impact on school attendance. But a systematic review did not support this finding (Taylor-Robinson et al., 2012). The controversy continues as the Taylor-Robinson review excluded interventions which also treated schistosomiasis. But combined treatment is recommended where schistosomiasis is a problem, and a new review will look at the impact of combined treatments.

The deworming debate points to the importance of sub-group analysis to examine heterogeneity in systematic reviews. As Cartwright (2007) emphasizes, policy makers do not want to know 'does this policy work on average?' they want to know, will it work for these people in this place. This is not a new challenge Greenhalgh (2003) makes the same point in her book *How to Read a Paper*, that doctors want to know 'will this treatment work for my patient?' Sub-group analysis and meta-regressions help answer these questions.

The final challenge is the ethical issue. There has been an enormous increase in data collection in developing countries in the last decade. Surveys are time consuming for respondents. So we have to really believe that what we are doing is worthwhile not just for us, but for the poor people whose time we are taking in conducting our studies. This consideration seems not to weigh heavily with many researchers, but clearly it should.

3ie's vision is to improve lives with impact evaluation. And that can and should be done. So let us not damage the reputation of evidence-based development with using the time of the poor to implement and evaluate ill-conceived interventions. Rather let us engage with priority questions of most importance to policy makers and poor people in developing countries, and so use evidence to improve policies, programmes and projects, spend development resources more effectively, and so truly to improve lives.

References

Attanasio, O., Kugler, A.D. and Meghi, C,,2011. Subsidizing Vocational Training for Disadvantaged Youth in Developing Countries: Evidence from a Randomized Trial. *American Economic Journal: Applied Economics,* 3 (3), 188-220.

Bloom, H., 2006. The Core Analytics of Randomized Experiments for Social Research. *MDRC Working Papers on Research Methodology*. New York: MDRC. http://www.mdrc.org/sites/default/files/full_533.pdf (accessed 30/12/12).

Cartwright, N., 2007. Are RCTs the Gold Standard? *BioSocieties,* 2, 11–20.

Casey, K., Glennerster, R. and Miguel, E., 2011. The GoBifo Project Evaluation Report: Assessing the Impacts of Community Driven Development in Sierra Leone. Report for 3ie, http://www.3ieimpact.org/evidence-hub/publications/impact-evaluations/gobifo-project-evaluation-report-assessing-impacts

Duflo, E., Glennerster, R. and Kremer, M., 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)

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B. and Vermeersch, C. M. J., 2010. *Impact Evaluation in Practice*. Washington D.C.: World Bank. Online at http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf (accessed 31/12/12).

Glewwe, P., Park, A. and Zhao,M., 2009. Visualizing development: Eyeglasses and academic performance in rural primary schools in China (Working Paper No. 12-2). Retrieved from Center for International Food and Agricultural Policy, University of Minnesota website http://purl.umn.edu/120032 (accessed 30/12/12).

Greenhalgh, T., 2003. *How to Read a Paper*. 2nd edition.

Karlan, D. and JAppel, J., 2011. *More Than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy*. Boston: Dutton Books.

Imai, I., King,G. and Nall, C., 2009. The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science,* **24**(1), 29-53.

King, E., Samii, C. and Snilstveit, B., 2010. Interventions to promote social cohesion in sub-Saharan Africa, *Journal of Development Effectiveness,* 2 (3), 336-370.

Humphreys, M., De la Sierra, R.S., Van der Windt, P., 2012. *Social and Economic Impacts of Tuungane Final Report on the Effects of a Community Driven Reconstruction Program in Eastern Democratic Republic of Congo*. New York: Colombia University. http://cu-csds.org/wp-content/uploads/2012/06/20120622-FINAL-REPORT.pdf (accessed 29/12/12).

Miguel, T. and Kremer, M., 2004. Worms: identifying impacts on education and health In the presence of treatment externalities. *Econometrica*, 72 (1), 159–217.

Organisation for Economic Co-operation and Development – Development Assistance Committee (OECD-DAC), 2002, 'Glossary of Key Terms in Evaluation and Results-based Management', Working Party on Aid Evaluation, OECD-DAC, Paris

Rasmussenab, O.D., , Malchow-Møllera, N. and Andersena, T.B., 2011. Walking the talk: the need for a trial registry for development interventions. *Journal of Development Effectiveness,* 3 (4), 502-519.

Ruel, M.T., Menon, P., Habicht, J.P., Loechl, C., Bergeron, G., Pelto, G., Arimond, M., Maluccio, J., Michaud, L. and Hankebo, B., 2008. Age-based preventive targeting of food assistance and behaviour change and communication for reduction of childhood undernutrition in Haiti: a cluster randomised trial., *Lancet,* 371 (9612),588-95

Samuels, F. and McPherson, S., 2010. Meeting the challenge of proving impact in Andhra Pradesh, India. *Journal of Development Effectiveness*, 2 (4), 468-485.

Scriven, M., 2008. A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research. *Journal of MultiDisciplinary Evaluation,* 5 (9), 15-24.

Sherman, L.W., Farrington, D., Welsh, B.C. and MacKenzie, D.L., 2002. Preventing Crime. In: Sherman, L.W., Farrington, D., Welsh, B.C. and MacKenzie, D.L., *Evidence-Based Crime Prevention.* London and New York: Routledge.

Taylor-Robinson, D.C., Maayan, N., Soares-Weiser, K., Donegan, S. and Garner, P., 2012. Deworming drugs for treating soil-transmitted intestinal worms in children: effects on nutrition and school performance. Cochrane Library. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000371.pub5/abstract;jsessionid=F36D7D4D76C97DCD449726D690534777.d04t03 Accessed 1/1/13.

Vajja, A. and White, H., 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.

Waddington, H., White, H., Snilstveit, B., Hombrados, J.G., Vojtkova,M., Davies, P., Bhavsar, A., Eyers, J., Koehlmoos, T.P., Petticrew, M., Valentine, J.C. and Tugwell, P., 2012. How to do a good systematic review of effects in international development: a tool kit. *Journal of Development Effectiveness,* 4 (3), 359-387.

White, H., 2008. Of probits and participation: the use of mixed methods in quantitative impact evaluation. *IDS Bulletin*, 39 (1), 98-189.

White, H.2009. Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness,* 1 (3), 271-284.

White, H., 2010a.A Contribution to Current Debates in Impact Evaluation. *Evaluation,* 16, 153-164

White, H., 2010b. Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness*, 1 (3), 271-284.

White, H., 2011. Achieving high quality impact evaluation design through mixed methods: the case of infrastructure, *Journal of Development Effectiveness,* 3 (1), 131-144.

White, H., forthcoming.The use of mixed methods in randomized control trials. In: Donna Mertens (ed.) *New Directions in Evaluation.*

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.

Ziliak, S. and McCloskey, D., 2008. *The Cult of Statistical Significance: now the standard error costs us jobs, justice and lives*. Ann Arbor: University of Michigan Press.

ANNEX

Impact Evaluation Glossary

Attribution
The extent to which the observed change in *outcome* is the result of the *intervention*, having allowed for all other factors which may also affect the *outcome(s)* of interest.

Attrition
Either the drop out of *participants* from the *treatment group* during the *intervention,* or failure to collect data from a unit in subsequent rounds of a *panel data survey*. Either form of attrition can result in *biased* impact estimates.

Average treatment effect
The average value of the *impact* on the *beneficiary* group (or *treatment group*). See also *intention to treat* and *treatment of the treated.*

Baseline survey and baseline data
A survey to collect data prior to the start of the *intervention*. Baseline data are necessary to conduct *double difference* analysis, and should be collected from both *treatment* and *comparison* groups.

Before versus after
See *single difference*.

Beneficiary or beneficiaries
Beneficiaries are the individuals, firms, facilities, villages or similar that are exposed to an intervention with beneficial intentions.

Bias
The extent to which the estimate of *impact* differs from the true value as result of problems in the evaluation or sample design (i.e. not due to *sampling error*).

Blinding
A process of concealing which subjects are in the *treatment group* and which are in the *comparison group*, which is single-blinding. In a double blinded approach neither the subjects nor those conducting the trial know who is in which group, and in a triple blinded trial, those analyzing the data do not know which group is which. Blinding is generally not practical for socio-economic development interventions, thus introducing possible *bias*.
Cluster sample A multi-stage *sample design*, in which a sample is first drawn of geographical areas (e.g. sub-districts or villages), and then a sample of households, firms, facilities or whatever, drawn from within the selected districts. The design results in larger standard errors than would occur in simple random sample, but is often used for reasons of cost.

Comparison Group
A group of individuals whose characteristics are similar to those of the treatment groups (or participants) but who do not receive the intervention. Under trial conditions in which the evaluator can ensure that no *confounding factors* affect the comparison group it is called a *control group*.

Confidence level
The level of certainty that the true value of *impact* (or any other statistical estimate) will be included within a specified range.

Confounding factors
Factors (variables) other than the programme which affect the *outcome* of interest.

Contamination
When members of the comparison group are affected by either the *intervention* (see *spillover effects*) or another intervention which also affects the *outcome* of interest. Contamination is a common problem as there are multiple development interventions in most communities.

Control Group
A special case of the *comparison group*, in which the evaluator can control the environment and so limit *confounding factors*.

Cost-benefit analysis (CBA)
A comparison of all the costs and benefits of the *intervention*, in which these costs and benefits are all assigned a monetary value. The advantage of CBA over analysis of *cost effectiveness*, is that in can cope with multiple outcomes, and allow comparison in the return to spending in different sectors (and so aid the efficient allocation of development resources).

Cost-effectiveness
An analysis of the cost of achieving a one unit change in the *outcome*. The advantage compared to *cost-benefit analysis*, is that the, often controversial, valuation of the outcome is avoided. Can be used to compare the relative efficiency of programs to achieve the outcome of interest.

Counterfactual
The state of the world in the absence of the *intervention*. For most impact evaluations the counterfactual is the value of the *outcome* for the *treatment group* in the absence of the *intervention*. However, studies should also pay attention to unintended outcomes, including effects on non-beneficiaries.

Dependent variable
A variable believed to be predicted by or caused by one or more other variables (*independent variables*). The term is commonly used in *regression* analysis.

Dichotomous variable
A variable with only two possible values, for example, "sex" (male=0, female = 1). The *dependent variable* in the *probit* participation equation estimated for *propensity score matching* is a dichotomous variable for which participate=1, didn't participate=0.

Difference-in-difference
See *double difference*.

Double difference
The difference in the change in the outcome observed in the treatment group compared to the change observed in the comparison group; or, equivalently, the change in the difference

in the outcome between treatment and comparison. Double differencing removes selection bias resulting from time-invariant unobservables. Also called Difference-in-difference. Compare to single difference and triple difference.

Dummy Variables
A dichotomous variable commonly used in *regression* analysis. *Impact evaluation* often uses a dummy variable for program participation (participate=1, didn't participate=0) as an *independent variable* in a regression in which the *dependent variable* is the *outcome* variable.

Effect Size
The size of the relationship between two variables (particularly between program variables and outcomes). See also *minimum effect size*.

Eligible population
Those who meet the criteria to be *beneficiaries* of the *intervention*. The population may be individuals, facilities (e.g. schools or clinics), firms or whatever.

Encouragement design
A form of *randomized control trial* in which the treatment group is given an intervention (e.g. a financial incentive or information) to encourage them to participate in the intervention being evaluated. The population in both treatment and control have access to the intervention being evaluated, so the design is suitable for national-level policies and programmes.

***Ex ante* evaluation design**
An impact evaluation design prepared before the intervention takes place. Ex ante designs are stronger than ex post evaluation designs because of the possibility of considering random assignment, and the collection of baseline data from both treatment and comparison groups. Also called prospective evaluation.

***Ex post* evaluation design**
An *impact evaluation* design prepared once the *intervention* has started, and possibly been completed. Unless there was *random assignment* then a *quasi-experimental design* has to be used.

Experimental Design
See *Randomized Control Trial*.

External Validity
The extent to which the results of the *impact evaluation* apply to another time or place.

Facility survey
A *survey* of a *sample* of facilities (usually for health or education, but could apply to police stations, training facilities and so on) that aims to assess the level and quality of all elements required to provide services. The *unit of observation* is the facility, though data may also be collected on staff in a separate facility staff survey (e.g. a teacher survey). If a facility survey is conducted alongside a household survey it is important that the *survey instruments* include information so as households can be linked to the facilities they use for the purposes of data analysis.

Factorial design
A *randomized control trial* with multiple treatment arms, in which one arm receives treatment A, a second arm treatment B, and a third both treatments (A+B). There may also be a fourth no treatment *control group*.

Hypothesis
A specific statement regarding the relationship between two variables. In an *impact evaluation* the hypothesis typically relates to the expected *impact* of the *intervention* on the *outcome.*

Impact
How an intervention alters the state of the world. *Impact evaluations* typically focus on the effect of the *intervention* on the *outcome* for the *beneficiary population*.

Impact evaluation
A study of the *attribution* of changes in the *outcome* to the *intervention*. Impact evaluations have either an *experimental* or *quasi-experimental* design.

Impact heterogeneity
The variation in *impact* as a result of differences in context, beneficiary characteristic or implementation of the *intervention*.

Independent Variable
A variable believed to cause changes in the dependent variable, usually applied in *regression* analysis.

Intention to treat estimate
The average treatment effect calculated across the whole *treatment group*, regardless of whether they actually participated in the intervention or not. Compare to *treatment of the treated.*

Internal Validity
The validity of the evaluation design, i.e. whether it adequately handles issues such as *sample selection* (to minimize selection bias), *spillovers*, *contagion,* and *impact heterogeneity*.

Intervention
The project, program or policy which is the subject of the *impact evaluation*.

Large n impact evaluation
Studies applying statistical means to construct a *counterfactual*, which requires a sufficiently large sample size (n) to ensure statistical *power*.

Logic model
Describes how a program should work, presenting the causal chain from inputs, though activities and outputs, to outcomes. While logic models present a theory about the expected program outcome, they do not demonstrate whether the program caused the observed outcome. A theory-based approach examines the assumptions underlying the links in the logic model.

Matching
A method utilized to create *comparison groups*, in which groups or individuals are matched to those in the *treatment group* based on characteristics felt to be relevant to the *outcome(s)* of the *intervention*.

Meta-analysis
The systematic analysis of a set of existing evaluations of similar programs in order to draw general conclusions, develop support for hypotheses, and/or produce an estimate of overall program effects.

Minimum effect size
The smallest effect size the researcher deems necessary to detect in the *impact evaluation*. Used to perform the *power calculation* necessary to determine required *sample size*.

Mixed methods
The use of both quantitative and qualitative methods in an impact evaluation design. Sometimes called Q-squared or Q2.

N
Number of cases. Uppercase "N" refers to the number of cases in the population. Lower case "n" refers to the number of cases in the sample.

Outcome(s)
A variable, or variables, which measure the *impact* of the *intervention*.

Panel data and panel survey
Data collected through consecutive surveys in which observations are collected on the same sample of respondents in each round. Panel data may suffer from *attrition*, which can result in *bias*.

Participant
An individual, facility, firm, village or whatever receiving the *intervention*. Also known *treatment group.*

Pipeline approach
An *impact evaluation* design in which the *comparison group* are those who have not yet received the intervention, but who are scheduled to do so. The assumption is that there will be no *selection bias*, since both *treatment* and *comparison* groups are to receive the interventions. However, the quality of the *matching* should be checked, since later *participants* may differ from those treated earlier.

Power
The ability of a study to detect an *impact*. Conducting a *power calculation* is a crucial step in impact evaluation design,

Power calculation
A calculation of the sample required for the *impact evaluation*, which depends on the *minimum effect size* and required level of *confidence*.

Primary Data
Data collected by the researcher specifically for the research project.

Propensity Score Matching (PSM)
A *quasi-experimental design* for estimating the *impact* of an *intervention*. The outcomes for the *treatment group* are compared to those for a *comparison group*, where the latter is constructed through matching based on propensity scores. The propensity score is the probability of participating in the intervention, as given by a *probit regression* on observed characteristics. These characteristics must not be affected by the intervention. PSM hence allows matching on multiple characteristics, by summarizing these characteristics in a single figure (the propensity score).

Quasi-Experimental Design
Impact evaluation designs used to determine impact in the absence of a *control group* from an *experimental design*. Many quasi-experimental methods, e.g. *propensity score matching* and *regression discontinuity design*, create a *comparison group* using statistical procedures. The intention is to ensure that the characteristics of the *treatment* and *comparison groups* are identical in all respects, other than the intervention, as would be the case from an *experimental design*. Other, *regression*-based approaches, have an implicit *counterfactual,* controling for *selection bias* and other *confounding factors* through statistical procedures.

Random assignment
An *intervention* design in which members of the *eligible population* are assigned at random to either the *treatment group* or the control group (i.e. *random assignment*). That is, whether someone is in the treatment or control group is solely a matter of chance, and not a function of any of their characteristics (either observed or unobserved).

Randomized Controlled Trial (RCT).
An *impact evaluation* design in which *random assignment* has been used to allocate the *intervention* amongst members of the *eligible population*. Since there should be no correlation between *participant* characteristics and the *outcome*, and differences in *outcome* between the treatment and control can be fully attributed to the intervention, i.e. there is no *selection bias*. However, RCTs may be subject to several types of *bias* and so need follow strict *protocols*. Also called *Experimental sesign*.

Regression Analysis
A statistical method which determines the association between the *dependent variable* and one or more *independent variables*.

Regression discontinuity design (RDD)
An *impact evaluation* design in which the *treatment* and *comparison* groups are identified as being those just either side of some threshold value of a variable. This variable may be a score or observed characteristic (e.g. age or land holding) used by program staff in determining the *eligible population*, or it may be a variable found to distinguish *participants* from non-participants through data analysis. RDD is an example of a *quasi-experimental design*.

Replication
Independent verification of study findings. Internal replication attempts to reproduce study findings using the same dataset, whilst external replication evaluates the same intervention in a different setting or at a different time. Internal replication may be pure replication, which uses the same data and model specification, or may test robustness to different model specifications, estimation methods and software.

Sample
A subset of the *population* being studied. The sample is drawn randomly from the *sampling frame*. In a simple random sample all elements in the frame have an equal probability of being selected, but usually more complex sampling designs are used, requiring the use of *sample weights* in analysis.

Sampling Frame
The complete list of the *population* of interest in the study. This is not necessarily the complete population of the country or area being studied, but is restricted to the eligible population, e.g. families with children under five, or female –headed households. For a *facility survey*, the sampling frame would be all facilities in the area of study. If a recent sampling frame is not available then one needs to be constructed through a field-based listing.

Secondary Data
Data that has been collected for another purpose, but may be reanalyzed in a subsequent study.

Selection Bias
Potential biases introduced into a study by the selection of different types of people into treatment and comparison groups. As a result, the outcome differences may potentially be explained as a result of pre-existing differences between the groups, rather than the treatment itself.

Sampling error
The error which occurs as estimates are used making data from a sample rather than the whole population.

Sample weights
A technique used to ensure that statistics generated from the *sample* are representative of the underlying *population* from which the sample is drawn. Sample weights should normally be used, though there is debate as to what to do when using *propensity score matching*, this is an alternative weighting system.

Single difference
Either, the comparison in the outcome for the treatment group after the *intervention* to its *baseline* value (also called *before versus after*), or an *ex post* comparison in the outcome between the *treatment* and *control groups.* Compare to *double difference.*

Small n impact evaluation
The set of best available methods when n is too small to apply statistical approaches to constructing a *counterfactual.*

Spillover effects
When the *intervention* has an *impact* (either positive or negative) on units not in the treatment group. Ignoring spillover effects results in a *biased* impact estimate. If there are spillover effects then the group of *beneficiaries* is larger than the group of *participants*. When the spillover affects members of the *comparison group*, this is a special case of *contagion*.

Survey
The collection of information using (1) a pre-defined *sampling* strategy, and (2) a *survey instrument*. A survey may collect data from individuals, households or other units such as firms or schools (see *facility survey*).

Survey instrument
A pre-designed form (questionnaire) used to collect data during a *survey*. A survey will typically have more than one survey instrument, e.g. a household survey and a *facility survey*.

Systematic Review
A synthesis of the research evidence on a particular topic, such as the effectiveness of water supply and sanitation, obtained through an exhaustive literature search for all relevant studies using scientific strategies to minimize error associated with appraising the design and results of studies. A systematic review is more thorough than a literature review. It may use the statistical techniques of a *meta-analysis*, but need not necessarily do so.

Theory-based impact evaluation
A study design which combines a *counterfactual* analysis of impact with an analysis of the causal chain, which mostly draws on factual analysis.

Theory of change
Laying out the underlying causal chain linking inputs, activities, outputs and outcomes, and identifying the assumptions required to hold if the intervention is to be successful. A theory of change is the starting point for *theory-based impact evaluation.*

Treatment group
The group of people, firms, facilities or whatever who receive the intervention. Also called *participants*.

Treatment of the treated
The treatment of the treated estimate is the *impact* (*average treatment effect)* only on those who actually received the *intervention*. Compare to *intention to treat*.

Triple difference
The comparative or differential impact on two groups, calculated as the difference in the double difference impact estimate for each group compared to a no treatment *comparison group*. A significant triple difference estimates demonstrates the presence of *impact heterogeneity*.

Unit of analysis
The class of elemental units that constitute the population and the units selected for measurement; also, the class of elemental units to which the measurements are generalized.

Unobservables
Characteristics which cannot be observed or measured. The presence of unobservables can cause selection *bias* in *quasi-experimental designs*, if these unobservables are correlated with both participation in the programme and the *outcome(s)* of interest.