A Contribution to Current Debates in Impact Evaluation

Howard White

Evaluation 2010 16: 153
DOI: 10.1177/1356389010361562

Invited Article

Evaluation
16(2) 153–164

DOI: 10.1177/1356389010361562
http://evi.sagepub.com
SAGE

A Contribution to Current Debates in Impact Evaluation

Howard White
International Initiative on Impact Evaluation (3ie)

Abstract

A debate on approaches to impact evaluation has raged in development circles in recent years. This paper makes a contribution to this debate through discussion of four issues. First, I point out that there are two definitions of impact evaluation. Neither is right or wrong, but they refer to completely different things. There is no point in methodological debates unless they agree a common starting point. Second, I argue that there is confusion between counterfactuals, which are implied by the definition of impact evaluation adopted in this paper, and control groups, which are not always necessary to construct a counterfactual. Third, I address contribution rather than attribution – a distinction that is also definitional, mistaking claims of attribution to mean sole attribution. I then consider accusations of being 'positivist' and 'linear', which are, respectively, correct and unclear. Finally, I suggest that these arguments do not mean that there is a hierarchy of methods, rather quantitative approaches, including RCTs, are often the most appropriate methods for evaluating the impact of a large range of interventions, having the added advantage of allowing analysis of cost effectiveness or cost-benefit analysis.

Keywords

attribution analysis; contribution analysis; counterfactuals; development; impact evaluation; selection bias

Introduction

Recent years have seen an increased focus on impact evaluation in the development community. One factor behind this trend has been the emphasis on results in the last 15 years, embodied in the Millennium Development Goals. The results agenda has seen what I, and many others, regard as a welcome shift in emphasis from inputs to outcomes. However, it has been realized that outcome monitoring does not tell us about the success, or otherwise, of government programmes or the interventions supported by international development agencies. In more recent years there has been advocacy by a number of groups for more rigorous evaluation of development programmes, most notably from the Poverty Action Lab at MIT and in the report *When Will We Ever Learn?* (CGD, 2006). The World Bank, the Inter-American Development Bank and more recently 3ie, are financing growing numbers of such studies.

Corresponding author:
Howard White, 3ie,, c/o Global Development Network, Post Box 7510, Vasant Kunj PO, New Delhi, 110070, India
Email: hwhite@3ieimpact.org

But this emerging focus has not been without controversy. Indeed there has been a somewhat heated debate about impact evaluation within the international development community, echoing similar debates in other fields of social inquiry. In this paper I seek to contribute to and hopefully clarify this debate.

The main argument I advance here is that there are a number of misunderstandings. The most important of these is that different people are using different definitions of 'impact evaluation'. Since this is a purely semantic matter, neither side is right or wrong. The definitions are just different. It makes little sense to debate on the appropriate methodology when people are in fact talking about different things. The debates become more focused and meaningful when they do address a common understanding of what we mean by impact evaluation, and I explore what I believe are some misunderstandings in this more focused debate. Finally, I address the issue of whether promotion of quantitative impact evaluation means there is a hierarchy of methods.

Controversies and confusions

Defining impact evaluation

The two sides of the impact evaluation debate are commonly talking about completely different things, but seem not to realize this.

Amongst evaluators, 'impact' typically refers to the final level of the causal chain (or log frame),[1] with impact differing from outcomes as the former refers to long-term effects. For example, the definition given by the Evaluation Network of the donor organization, the Development Assistance Committee (DAC) is 'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'. Any evaluation which refers to impact (or often outcome) indicators is thus, by definition, an impact evaluation. So, outcome monitoring can fall under the heading of impact evaluation. In addition, there are established fields of impact assessment, including participatory impact assessment, which rely largely or solely on qualitative approaches which also fall under the impact evaluation label since they are concerned with outcomes and impacts.

But this definition is not shared by many working on impact evaluation, for example in the World Bank, or, indeed, 3ie. Impact is defined as the difference in the indicator of interest (Y) with the intervention (Y_1) and without the intervention (Y_0). That is, impact = $Y_1 - Y_0$ (e.g. Ravallion, 2008). An impact evaluation is a study which tackles the issue of attribution by identifying the counterfactual value of Y (Y_0) in a rigorous manner. I leave aside for the moment what constitutes rigour, though as I shall explain, I am arguing in favour of greater use of quantitative approaches to attribution. Usually this is an analysis of outcome or impact indicators, but not necessarily so.

These are completely different definitions of impact. They *may* overlap if Y is an outcome indicator. But I now believe that drawing attention to the overlap (which I have done many times), or worse still, treating the two definitions as if they are somehow the same, confuses the real issue, which is the fundamental difference between the two definitions. Since this is a purely semantic matter, neither side is right or wrong. The definitions are just different. No debate about methodology will be of any use unless we first agree which definition is being used.

Hence, many studies can be considered to be impact evaluations since they discuss outcome and impact indicators, while making no attempt to attribute changes in those indicators to the intervention. Indeed such studies often explicitly state that attribution is not possible.[2] But such studies are most decidedly not an impact evaluation to someone for whom attribution is the defining characteristic, and especially those using quantitative methods to measure impact. However many of the

objections to quantitative approaches to impact evaluation, are *not* methodological disagreements about the nature of causality, they are simply using a different definition of impact evaluation.

So much of the current debate could be avoided with some terminological clarification. The current push toward impact evaluation is for studies which can attribute changes in selected outcomes (or possibly outputs) to a specific intervention, and to do so for interventions for which quantitative methods are the best available method. It is this, second, definition which was intended in the CGD report (2006). We may wish to call these studies 'attribution analysis', rather than impact evaluation to avoid appropriating a term already in use with a different meaning, though I fear it is perhaps too late for such relabelling. But the different sides in the debate need to understand that they mean different things by 'impact evaluation'. And there is no reason at all why these quite different types of studies should adopt the same methodology.

It should also be made clear that both definitions are the basis for useful studies. However, the current focus on funding attribution studies originated from the fact that there had been an underinvestment in evaluations of this sort, a feeling articulated most clearly in the CGD report (2006). Hence there is a lack of evidence about what works and what doesn't – and at what cost. With that point in mind we can focus on a more precise question, which is must 'scientifically valid' methods of experimental and quasi-experimental methods be used to attribute observed changes to a specific intervention? Note that this wording is intended to make it very clear that there are many evaluation questions for which other methodological approaches are more appropriate. Impact evaluation is not the only game in town, and not the only one to be played in the assessment of development interventions.

Whether experimental and quasi-experimental approaches are part of the best available method depends on the nature of the intervention being evaluated: is it a small *n* or large *n* intervention? Here *n* refers to the unit of assignment of the intervention, for example, households, firms, schools, communities, districts or Ministries.

The unit of assignment is not necessarily the same as the unit of analysis. For example, we might assess the impact of a school-level intervention on student learning outcomes, or of feeder roads on household incomes. But it is the unit of assignment which drives the power calculations, and so is the relevant *n* for knowing whether statistical approaches are the best available methodology to answer the questions at hand.

For small, and often middle size, *n* then qualitative approaches are the best available methodology, such as the fuzzy set social science approaches discussed by Ragin (2000); see Khagram et al. (2009) for a discussion of their application to the field of development evaluation. Ironically, such an example of when such approaches are the best available methodology is assessing the impact of impact evaluations. How have impact studies influenced policy? Such a study would require a stakeholder mapping exercise, determining the overall drivers of change, and the place of the study, if any, in those changes – such an approach is adopted by IFPRI in its Policy-Oriented Research Impact Assessments (see http://www.ifpri.org/book-25/ourwork/researcharea/impact-assessment).

However, in some cases different quantitative approaches can be the most appropriate approach for small *n* studies, for example macroeconomic modelling to assess the impact of economic policy changes such as devaluation or changes in tariff rates (an *n* of one).

There is a need for more agreement on which of the range of available methodologies is most appropriate in which settings, but that is not the subject of this paper.

But there is a wide range of large *n* interventions for which experimental and quasi-experimental approaches can, and should be, used, and have been under-utilized to date. As a result, the evidence of the lack of evidence of what works is quite overwhelming. CGD listed several reviews which revealed the scanty nature of evidence regarding what works (2006). A more recent review by

NORAD found that most evaluation studies had little, or even no, basis for the conclusions drawn as to impact (Villanger and Jerve, 2009). Finally, a review by 3ie of the 'evaluative reports database' of the Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP) showed that not one of the 339 evaluation studies could be classified as a rigorous impact evaluation.

In brief, the quantitative methods used should be able to deal with the problem of 'selection bias'. The bias arises from either programme placement (project communities, households, firms, etc. are chosen by some systematic criteria) or self-selection (communities or people opt in or out of programmes). Now, if the characteristics determining programme participation are correlated with the outcomes of interest, comparison of outcomes between the participants (the treatment group) and a comparison group will result in a biased estimate of the changes brought about by the programme. This is not an unlikely situation. If these selection characteristics can be measured, then they can be controlled for statistically (quasi-experimental methods). But if they cannot be measured (so-called unobservables) and change over time, then randomization (randomized control trials, RCTs), and only randomization, can produce numerical estimates of impact which are free from bias – though, as critics correctly point out, even randomization may not do the trick (Deaton, 2009; Scriven, 2008).

The so-called methodological debate, more specifically the critique by evaluators of the use of quantitative methods, including RCTS, is missing the point. The problem is not that there are growing numbers of studies applying experimental and quasi-experimental approaches to assess the impact of development interventions. There was a clear 'gap'. Previous evaluation studies typically made statements that attribution was not possible. But it is possible, and these studies are filling that gap. However, a problem remains. The problem is if all these studies do is estimate an average treatment effect: measurement is not evaluation.[3] Rigorous impact evaluation is a building block, a necessary component, of a good-quality impact evaluation, but it is not all there is to it.

The current debate revolves around selection bias and how it should be handled. While I agree that this source of bias should be addressed where present, I see this as a technical issue in evaluation design which has to be dealt with appropriately. But the two main points are: 1) to embed the analysis in a mixed-methods approach which unpacks the causal chain, and 2) to use the impact estimate as the basis for analysis of either cost effectiveness or the rate of return. I have dealt with the first of these issues elsewhere (see White 2008 and 2009). Here I expand briefly on the latter.

Forty years ago the value of having good-quality measures of intervention benefits would not have been questioned. Projects were typically subject to cost-benefit analysis, which required quantification of the benefit stream; that is, the difference in outcomes (such as agricultural value added) resulting from the project. Cost-benefit analysis was explicitly based on with versus without analysis; that is, establishment of a counterfactual of what outcomes would have been in the absence of the project. This counterfactual was commonly generated through use of a comparison group – what recent developments do is give more robust estimates of these benefits.

Cost-benefit analysis fell into general disuse in the 1980s and early 1990s as a result of two related trends (see White, 2005, for more discussion). First was the increase in social sector interventions which were, mostly mistakenly, seen as not being amenable to quantitative analysis. The inappropriateness of these methods seemed even clearer once projects began with objectives such as 'empowerment' and 'building social capital'. The second trend was the rise of participatory approaches to development, including participatory evaluation. Such an approach rejects evaluation against objectives set by outsiders – who may well not have the required knowledge of local circumstances – in favour of local narratives and judgements of 'success'. But the rise of quantitative impact evaluation should enable a resurgence of cost effectiveness and cost-benefit analysis.

In summary, there are good, policy-relevant reasons for making quantitative estimates which attribute changes in outcomes to the intervention. To do this, evaluation designs must take account of selection bias where it arises, which will usually require resorting to a comparison group constructed using either an experimental or quasi-experimental approach. These estimates should be part of a broader evaluation design, and the benefits compared to the costs of the intervention and so guide the allocation of resources.

In the remainder of this article I am addressing impact evaluations whose explicit purpose is an analysis of attribution. I have noted already that this is just one definition of impact evaluation and that the other definition is not wrong, but different, and that that other definition opens the door to alternative methodologies. But I do subscribe to the view that there is a need for a greater number of quality studies addressing attribution using quantitative methods.

Comparison groups and counterfactuals

A further area of confusion relates to counterfactuals and control groups. Once we move to a discussion of attribution I believe we are necessarily dealing with counterfactuals, though they may be implicit.

It is over 35 years since Lewis's influential paper established the counterfactual approach as the principal legitimate approach to the analysis of causation. Prior to that, regularity analysis (which could be manifested as statistical correlations) ruled the day. But Lewis argued that regularity analysis would fail to detect causation, or incorrectly do so, for a number of relatively straightforward reasons. The counterfactual approach remains a powerful one in scientific circles – see, for example, the recent collection of Collins et al. (2004), despite critiques.

Although it may not always be necessary or useful to make the counterfactual explicit, in attributing changes to outcomes to a development intervention it most likely is useful to have an explicit counterfactual. An explicit counterfactual does not necessarily mean that one needs a comparison group, though often it will.

As already discussed, we are defining impact as $Y_1 - Y_0$. What happens with the intervention is observed, this is Y_1. What we don't know is what would have happened without the intervention (Y_0). There are various ways of getting an estimate of Y_0. A common, though usually unreliable, one is the value of Y before the intervention; that is, the before versus after approach. This approach is unreliable since other things affect Y, so not all of the change in Y can be attributed to the intervention, which is what the before versus after approach does. However, there are some cases in which there are no other plausible explanations of the change in Y so before versus after *will* suffice. A good example is the impact of a water supply project on the time household members spend collecting water. The average time falls after the project. The only plausible explanation is the improved proximity of water. In this case there is no need for a comparison group from outside the project area – the most meaningful comparison is with the treatment group before the intervention (a comparison group would give a less accurate estimate in this case). But having no comparison group is not the same as having no counterfactual. There is a very simple counterfactual: what would Y have been in the absence of the intervention? The counterfactual is that it would have remained unchanged, that is, the same as before the intervention.

Supporting a before versus after approach, Scriven argues that it is clear that 'a program that gives food or shelter to those who need it immediately has beneficial consequences' (2008: 20). But this actually is *not* true for more complex outcomes, and even if it is, 1) there is still a counterfactual, and 2) quantification allows analysis of the efficiency of the intervention in achieving welfare outcomes.

We might also think that before versus after is adequate if there is no observed change in Y after the intervention. Y has remained unchanged, so clearly the intervention had no impact. Once again there is a counterfactual here, though there is no comparison group. The counterfactual is the same as in the previous example, which is that if there were no intervention then we'd expect Y to keep its pre-intervention value. But we might well be on shaky ground here. Perhaps there is a downward trend in Y, for example, declining yields in a fragile ecological zone. So observing the same value before and after the intervention is in fact a positive impact from the intervention as without it yields would have been lower. Unless we have both trend data to suggest Y is not changing over time, and good arguments as to why other factors will not have affected it during the intervention, then a stronger counterfactual is needed, which is what a comparison group can provide.

It is objected that attribution statements, causality statements, in many areas of science are made without a counterfactual. For example, there is no counterfactual in the analysis of the moon causing tides or to explain night and day by the revolutions of the Earth around its axis. Actually there *is* an implicit counterfactual – for example, no moon then no, or more likely lesser, tides. There is no comparison group. Hence it clearly cannot be claimed that a comparison group is always required to demonstrate causation. There is a counterfactual, but not one which needs be made explicit in this case.

What we are interested in is the underlying causal chain. My approach to impact evaluation is to start with the outcomes and impacts and to identify the range of factors which influence these outcomes. I then ask whether the project outputs will have any effect on these causal factors. That is, I build up the programme theory. The job of a quality impact evaluation is to trace the causal chain through from inputs to outcomes and impacts, and different approaches most applicable to analysing different parts of it (see White, 2009, for an elaboration of the approach). A note of caution can be added since there are cases in which evidence precedes theory, most famously the origins of evidence-based medicine in the West. Working in a Vienna hospital in the 1840s, Semmelweis noted that maternal mortality from hospital deliveries was 30 percent, much higher than that in home births. Although the germ theory of disease would not be established for another three decades, Semmelweis got doctors to wash their hands before performing deliveries, reducing the mortality rate to just 2 percent. Despite this achievement, Semmelweis was widely attacked by the Viennese medical establishment as his recommendations lacked a theoretical foundation (for a recent description of this episode in the history of quantitative evidence making, see Ayres, 2008).

Usually a theory will help. But there can be cases when, as for Semmelweiss, the evidence is just so compelling. Such seems to be the case with 'Scared Straight' schemes in which teenage boys are given a taste of prison in order to discourage them from a life of crime. These programmes have been subject to RCTs in several US states and the UK, finding at best no impact, the average treatment effect from the meta-analysis being negative (Petrosino et al., 2003). So it is sufficient to note that such schemes simply don't work (though qualitative data have provided reasons, which is that many of the young men are actually rather attracted to prison life, plus that they make some useful contacts there).

That is a case when the intervention has never worked. A different treatment for juveniles – boot camps – has rather more mixed results. The Campbell review of 32 boot camp evaluations, while reporting no overall impact, finds five in which there is a significant positive impact (and eight where it has a significant negative impact) (Wilson et al., 2008). Hence further examination of why it worked in some cases and not in others may seem helpful – though the effect size remains small compared to other interventions.

Consider a business services project. Some causal elements seem straightforward, such as the delivery of counselling services. One can test whether entrepreneurs have acquired the knowledge

they are meant to have acquired through counselling and training. But perhaps this is something they knew already (as in the case of an extension project in Kenya evaluated by the then OED (now IEG); World Bank, 1999) – before versus after analysis will tell us this if we have baseline data, but if not a comparison group of peers will be useful. If it is new knowledge, do they put it into practice? The World Bank analysis of a nutrition project in Bangladesh found this was often not the case. More in-depth qualitative interviews are most likely the best means of identifying the reasons for this knowledge–practice gap: in Bangladesh it was the influence of mothers-in-law (White and Masset, 2007). This is a straightforward factual analysis. If entrepreneurs do adopt the advice, does it affect profitability? A before versus after analysis won't work as market conditions may have also changed, affecting profitability. A simple comparison group will almost certainly suffer from selection bias, if only the self-selection bias of the more motivated applying for the programme and sticking with it. Hence for the analysis of the outcome indicator a comparison group derived by experimental or quasi-experimental means is necessary, and this will usually be the case. Consider the water supply project already mentioned. Before versus after indeed suffices for analysis of changes in time use, but if we wanted to know the impact on child diarrhoea then a comparison group would be necessary.

Contribution versus attribution

Attribution is usually seen as a substantial problem. Again, it is first necessary to sort out definitions. By attribution, I mean attributing observed changes to the intervention being studied. For many development agencies, attribution means attributing observed changes to their aid dollars. This latter meaning of attribution raises additional issues such as fungibility and attribution in the presence of multiple donors, which are beyond the scope of this article. But for now, I stress that my definition requires that we link observed changes to the intervention, regardless of the source of funding.

Many suggest that it is difficult, if not impossible, to attribute a change in outcomes to a specific intervention since there are so many different factors involved, so we had best look instead for a contribution. This argument confuses attribution with sole attribution. It is not being argued that the intervention was the sole cause of observed an observed change. Many outcomes of interest are not dichotomous, so, for example, infant mortality in a particular region may have fallen by 12 percent over the period of the intervention. The impact question is how much of that 12 percent can be attributed to the project? Even if the outcome is dichotomous, (i.e. an outcome happens or it does not) then the impact question is how the intervention affected the probability of the event occurring, which is indeed what the impact on the mortality rate tells us for any given infant.

Hence a common finding in an impact evaluation would be that intervention X caused outcome Y to change by p percent. A good study would dig a bit deeper, and say that since Y changed by P percent over the period of the intervention, say, a quarter ($p/P = 0.25$) of the overall change can be attributed to the intervention. That is, the analysis of attribution allows identification of how much the intervention contributed to the overall change in the outcome of interest. In this sense, attribution analysis also addresses contribution.

This use of 'contribution analysis' is not the same as Mayne's 'contribution analysis'. As Mayne argues, there may be cases in which collecting data on trends in outcomes and plausible explanatory factors of observed trends may suffice to show that the intervention contributed to the outcome. However, Mayne is clear that this method is not an approach to impact evaluation. It is an evaluation tool which will serve some purposes, but quantifying impact, in the sense used here, is not one of them. He states explicitly that 'an [impact] evaluation study probably remains the best way to address [the attribution] problem, if one has the time, money and expertise' (2001).

This same argument also addresses some of the more philosophical objections regarding causation. Suppose both A and B are, individually, sufficient to cause C, but just one of them is necessary (the problem of over-determination). Hence if A happens C happens and there is a causal relation. But if A happens and then B happens, then C would happen even if A had not happened. This problem occurs when the outcome is dichotomous. And it most certainly is an issue in some areas of development evaluation, such as how policy dialogue affects policy outcomes. But the outcome of interest usually varies over a range for example, enrolment or mortality rates. In this case both A and B can contribute to changes in C.

Quantitative impact evaluation can also make more complex causal inferences regarding context. Context is one aspect of impact heterogeneity. That is, impact varies by intervention, characteristics of the treated unit and context. A study which presents a single impact estimate (the average treatment effect) is likely to be of less use to policy-makers than one examining in which context interventions are more effective, which target groups benefit most and what environmental settings are useful or detrimental to achieving impact. Hence it could be shown that an educational intervention, such as flip charts, works but only if teachers have a certain level of education themselves, or only if the school is already well equipped with reading materials, or the pupils' parents are themselves educated.

'Linear' is not always bad

Being 'linear' is apparently a self-evident bad thing to many in the evaluation field, but in fact it is not that clear what is meant by it. For those of a mathematical or modelling frame of mind, the term implies a linear relationship, meaning that a unit change in X causes a fixed increment in Y, regardless of the value of X. Economists would usually expect there to be diminishing returns to scale, so that as X increases it starts to have less and less impact on Y: maybe after a certain level the effect is even negative. Alternatively, there could be a 'threshold effect' with a certain amount of X needed before there is any effect. All of these possibilities are readily captured by logarithmic, quadratic or spline function (interactive terms) model specifications. It might be argued that X only has an impact in the presence of factor Z, a hypothesis which is readily tested by the use of an interactive variable (a new variable which is the product of X and Z).

An alternative criticism of linear appears to refer to one-way causation. That is, quantitative impact evaluation assumes that X affects Y without allowing for the fact that Y may also affect X. Such a critique is patently false, as the whole fuss about selection bias is precisely about bi-directional causality: schools with school improvements perform better (X to Y), but it was the better performing schools that were most likely to get the intervention in the first place (Y to X), because the programme picks them, or because their management has its act together to make a successful application. It is to remove the possibility of this second link that random allocation of the treatment is proposed.

The accusation 'linear' also appears to mean that models of the causal chain imply an automatic (or deterministic) relationship. This point is sometimes conflated with the point that there are multiple determinants, so sole attribution is not possible. But this criticism is incorrect. Statistical modelling reflects trends. To take a classic controversial case, someone from a disadvantaged background is more likely to resort to crime, but not all (or even the majority) of people from such backgrounds do so. Similarly, an intervention can reduce the crime rate among likely offenders, which does not mean that none will reoffend. Statistical analysis is of stochastic relationships, meaning precisely that there are also some unknown elements ('the error term') which affect outcomes.

Another interpretation of linear is that it misses the multiple causal pathways which occur in real interventions. Actually this problem does not matter for a purely outcome-oriented evaluation with a black box approach, which can tell you whether there is an impact or not using treatment versus control designs no matter how complex the intervention. But for theory-based approaches which attempt to unpack the causal pathways it is indeed a challenge, but not an insurmountable one. Models are developed precisely to capture more complex causal mechanisms, including multiple channels, confounding variables and bi-directional causality.

Bi-variate causality is at least a two-equation model, but can be more. In the example just given school improvements affect learning outcomes (equation 1), but school management capacity affects both receiving school improvements (equation 2) and learning outcomes (equation 3).

A further, and related, use of linear in fact refers to single-equation modelling of a multi-equation system, summed up as 'it's all terribly complex', or more formally as 'these approaches may be all very well for very simple interventions, but they can't be used for the more complex interventions of the sort commonly implemented today'. This statement needs to be understood to respond to it.

Rogers (2008) distinguishes between complex and complicated, the latter meaning an intervention with many components. In the latter case one or more components may be amenable to rigorous impact evaluation. And designs are available to examine whether different components complement (reinforce) one another, or possibly substitute (counteract) each other. It has to be recognized that not all components of an intervention will be amenable to the same approach. Budget-support and sector-wide approaches are said not to be suitable for large *n* studies. But such a view misses the point that such aid supports government spending, and the government programmes so supported can be the subject of such studies. At the same time, the donor wishes to affect the level and composition of government spending, for which there are a range of alternative analytical approaches, and to influence policy for policy dialogue, which has also been subject to a sizeable literature, both quantitative and qualitative.[4] A holistic approach has to be taken to the evaluation design, selecting the appropriate methodology for the different evaluation questions which arise to assess the intervention as a whole.

Complex interventions may be of various types (Rogers): those with recursive causality (i.e. feedback or reinforcing loops), disproportionate or 'tipping point' relationships and emergent outcomes. The first and second of these issues has already been dealt with. The first problem is precisely that addressed by experimental and quasi-experimental approaches and the second a form of non-linearity readily accommodated in statistical analysis.

This leaves us with the problem of emergent outcomes. We may not know what the outcomes of an intervention will be in advance: what the outcomes will be can only be discovered as the intervention proceeds. I deal with this question in part by turning to the charge that impact evaluation is undemocratic.

Is democratic impact evaluation possible?

Jennifer Greene has stated that impact evaluation is 'inherently undemocratic'.[5] What I take this to mean is that the way in which such studies are conducted and utilized is not very participatory. The approach to most impact evaluation studies is what Robert Chambers calls 'extractive' – that is, the researchers go and collect the data, scurry back to their ivory tower to analyse it, publish their article and move on. In contrast 'empowering' research takes place within the community, allowing them to define its parameters, take part in the analysis, so by deepening their understanding of their situation they are able to overcome obstacles, with research findings reinforcing this effort.

A literal interpretation of this approach, running Stata classes in the African bush, is clearly nonsense. But the impracticality of that proposition should not make us turn away from useful principles derived from a more empowering approach to impact evaluation. First, does the intervention address issues which matter to the local population? This is the standard evaluation question of relevance, but it can also guide impact evaluation design in selecting the indicators to be analysed. Second, did the intervention tackle perceived obstacles to improving wellbeing, which may well change as new understandings and experiences 'emerge' when a new programme is implemented. Local analysis should inform the analysis of the casual chain. And, finally, findings and interpretation should be fed back to the local community in an interactive manner. These mechanisms are being put in place as impact evaluations move more strongly to a mixed methods approach.

Is there a hierarchy of methods?

Evaluations should be issues-led not methods-led. And having determined the evaluation questions, the best available method should then be used to answer them. These methods may, or may not be, quantitative.

The points where critics need to give some ground is to accept that: 1) there is a problem of selection bias which needs addressing in any assessment of attribution; 2) RCTs are very often the most appropriate means of dealing with this problem; and 3) if not RCTs, then some other quantitative (quasi-experimental) method will be the best available method for large *n* interventions.

Indeed this argument does appear to be implicitly accepted since critics often conflate RCTs with all quantitative approaches. For example, it is pointed out that proof of the adverse impact of smoking was not proved by experiments, since it would not have been ethical to do so. This claim is not entirely true since large-scale smoking experiments were conducted on dogs. But although RCTs were not used, quantitative methods were used to compare health outcomes in smokers and non-smokers, while controlling for age, sex, occupation, environmental conditions and so on, in order to link smoking to lung cancer. This is a decidedly quantitative approach, and wholly consistent with the view that other quantitative methods should be used when RCTs are not possible. Hence the smoking example, and many others, do support the use of quantitative methods. As argued in this article, a further reason for quantification is the possibility of analysing cost effectiveness, or better still conducting a cost-benefit analysis.

However, I don't believe these statements mean that there is a hierarchy of methods. It is rather that we want to use the best available method. There are many settings when quantitative methods will be the best available method. But there are also many cases when they are not. And as argued, a theory-based approach will usually combine methods: quantitative and qualitative and evaluation approaches. Hence, I believe the argument between proponents of theory-based evaluation and RCTs is overstated. A theory-based approach provides a framework for an evaluation. It still needs an analytical approach to determine if outcomes have changed as a result of the intervention, so experimental or quasi-experimental approaches are embedded in a mixed-methods approach. It is to be hoped that traditional development evaluators, and more quantitatively oriented researchers, can work together with such approaches to produce studies to influence policy and improve lives.

Notes

1. Many evaluators now object to the log frame saying it is 'old fashioned', or, worse still, linear. However the log frame is alive and well in development agencies.
2. While compiling a preliminary impact evaluation database for NONIE a major donor told me that 'all their evaluations were impact evaluations'. It was indeed true that the ToR included impact analysis, most

commonly appearing the final report along the final lines: 'It is not possible to attribute these changes to the activities supported by the project.'

3. I owe this snappy summary of the critique – 'measurement is not evaluation' – to Jocelyne Delarue of the Agence Française de Development (AFD). In their own programme, AFD has run into the problems mentioned here arising from relying solely on academics whose main interest is novel and clever techniques to derive an average treatment effect.
4. The seminal quantitative study looking at the impact of donor-supported policy changes while allowing for the endogeneity of policy change is Goldstein and Montiel (1984). The literature in general has used mixed methods, relying on more qualitative approaches to assess policy dialogue (e.g. Mosley et al., 1991; White and Dijkstra, 2004).
5. Presentation to conference 'Perspectives on impact evaluation', 31 March–2 April 2009, Cairo.

References

Ayres, I. (2008) *Supercrunchers: Why Thinking by Numbers is the New Way to be Smart.* New York: Bantam Books.

CGD (2006) *When Will We Ever Learn?* Washington, DC: Center for Global Development.

Collins, J., N. Hall and L. A. Paul (2004) *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Deaton, A. (2009) *Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development*. NBER Working Paper, 14690. Cambridge, MA: NBER.

Goldstein, M. and P. Montiel (1986) 'Evaluating Fund Stabilization Programs with Multicountry Data: Some Methodological Pitfalls', *IMF Staff Papers* 33: 304–44.

Khagram, S., C. Thomas, C. Lucero and S. Mathes (2009) 'Evidence for Development Effectiveness', *Journal of Development Effectiveness* 3(1): 247–70.

Lewis, D. (1972) 'Causation', *Journal of Philosophy* 70: 556–67.

Mayne, J. (2001) 'Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly', *Canadian Journal of Program Evaluation* 16(1): 1–24.

Mosley, P., J. Toye and J. Harrigan (1991) *Aid and Power*. London: Routledge.

Petrosino A., C. Turpin-Petrosino and J. Buehler (2003) '"Scared Straight" and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency' (updated C2 review), *Campbell Collaboration Reviews of Intervention and Policy Evaluations* (C2-RIPE). Philadelphia, PA: Campbell Collaboration, Nov.

Ragin, C. (2000) *Fuzzy Set Social Science*. Chicago: University of Chicago Press.

Ravallion, M. (2008) *Evaluating Anti-Poverty Programs: T. P. Schultz and John Strauss Handbook of Development Economics*, vol. 4.

Rogers, P. (2008) 'Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions', *Evaluation* 14(1): 29–48.

Scriven, M. (2008) 'A Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research', *Journal of MultiDisciplinary Evaluation* 5(9): 11–24.

Villanger, E. and A. Morten Jerve (2009) 'Assessing Aid Impact: A Review of Norwegian Evaluation Practice', *Journal of Development Effectiveness* 1(2): 171–94.

White, H. (2005) 'Challenges in Evaluating Development Effectiveness', *IDS Discussion Paper* 42, also published in Pitman et al., http://129.3.20.41/eps/dev/papers/0504/0504014.pdf

White, H. (2008) 'Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation', *IDS Bulletin* 39(1): 98–109.

White, H. (2009) 'Theory Based Impact Evaluation: Principles and Practice', *Journal of Development Effectiveness* 1(3): 271–284.

White, H. and G. Dijkstra (2004) *Beyond Conditionality: Program Aid and Development.* London: Routledge.

White, H. and E. Masset (2007) 'The Bangladesh Integrated Nutrition Program: Findings from an Impact Evaluation', *Journal of International Development* 19: 627–52.

Wilson, D. B., D. L. MacKenzie and F. N. Mitchell (2008) 'Effects of Correctional Boot Camps on Offending', *Campbell Systematic Reviews* 2003:1 (updated Feb. 2008). Oslo: Campbell Collaboration.

World Bank (1999) *Agricultural Extension: The Kenya Experience*. An Impact Evaluation, 198. Washington, DC: World Bank. http://lnweb90.worldbank.org/oed/oeddoclib.nsf/InterLandingPagesByUNID/B728D887FC2B754D852568BD005A8C19

World Bank (2005) *Maintaining Momentum to 2015? An Impact Evaluation of Interventions to Improve Maternal and Child Health and Nutrition in Bangladesh.* Washington, DC: IEG, World Bank.

Howard White is the Executive Director of the International Initiative for Impact Evaluation (3ie), based in New Delhi, India. Please address correspondence to: 3ie,, c/o Global Development Network, Post Box 7510, Vasant Kunj PO, New Delhi, 110070, India. [email: hwhite@3ieimpact.org]