

International Initiative for Impact Evaluation



WORKING PAPER 1

Some Reflections on Current Debates in Impact Evaluation

Howard White
April 2009

About 3ie

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

3ie Working Papers cover both conceptual issues related to impact evaluation and findings from specific studies or synthetic reviews.

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

© 3ie, 2009

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

SOME REFLECTIONS ON CURRENT DEBATES IN IMPACT EVALUATION

Howard White

Executive Director

International Initiative on Impact Evaluation, 3ie

Email: hwhite@3ieimpact.org

Abstract

A debate on approaches to impact evaluation has raged in development circles in recent years. This note gives some reflections on this debate through discussion of four issues. First, it is pointed out that there are two definitions of impact evaluation. Neither one 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, 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, calls for addressing contribution rather than attribution are also definitional, mistaking claims of attribution to mean sole attribution, which is does not. I then consider accusations of being 'positivist' and 'linear', which are, respectively, correct and unclear. Finally, these arguments do not mean that there is a hierarchy of methods, rather that quantitative approaches, including RCTs, are often the best available methods, having the added advantage of allowing analysis of cost effectiveness.

1. Introduction

Over the past few years there has been a somewhat heated debate about impact evaluation within the international development community. This paper offers some reflections on the debate. The main argument in the paper 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.

Since this paper provides an 'insider's' review of these debates it is useful to summarize my own experience. In 2002 I joined the Operations Evaluation Department (now called the Independent Evaluation Group, IEG) of the World Bank to undertake a series of impact evaluations.¹ Our first studies coincided with the establishment of the Development Impact Evaluation Initiative (DIME) by the Bank's research department. Initially DIME, especially within the Human Development Network responsible for health and education interventions, was a proponent of randomized control trials (RCTs). There followed some internal discussion of RCTs, but both the opposition to sole reliance on these approaches by the Bank's leading micro-econometric researcher (see, e.g. Ravallion, 2003), and the experience of the approach's limited applicability when it came to designing *ex ante* evaluations of Bank-supported projects, led to a broad acceptance of quasi-experimental approaches (most notably propensity score matching). IEG's own program comprised *ex post* studies, which were necessarily quasi-experimental. Hence IEG sided with those who saw a place for RCTs, but argued they could not be exclusively relied upon (IEG, 2006). However, the main feature of IEG's position was the importance of ensuring the policy relevance of studies, stressing the importance of mixed methods (White, 2008) and of engaging with stakeholders throughout the evaluation process. Nonetheless, IEG's position also stressed that quantitative analysis was a necessary component of impact evaluation and that potential biases had to be explicitly addressed, a position which became more consolidated as the debate progressed.

Meanwhile, IEG was engaging the Evaluation Network of the bilateral group the Development Assistance Committee (DAC) to press for 'more and better' impact evaluation. From 2002 to 2005 the Bank's presentations to the Evaluation Network were largely met with polite disinterest. The situation was different at the November 2006 meeting where there was an agreement from bilaterals to work together, leading to the formation of the Network of Networks for Impact Evaluation (NONIE) at a subsequent

¹ Six studies resulted from this work program, covering Ghana education (World Bank, 2004), Bangladesh health and nutrition (World Bank, 2005), Andhra Pradesh irrigation (World Bank, 2006), rural electrification (World Bank, 2007), a review of water supply and sanitation (World Bank, 2008), and Andhra Pradesh rural development (still on-going).

meeting in May 2007. NONIE included not only DAC but also the UN Evaluation Group (UNEG) and the Evaluation Cooperation Group (ECG) of the multilateral development banks. IEG played a lead role in the creation of NONIE and became the NONIE Secretariat. The main supporters of the creation of NONIE understood its purpose to be the promotion of 'rigorous impact evaluation'. However, it has proved to be a challenge to withstand its mission to creep into broader areas of evaluation, and to move on past methodological debates as to the meaning of impact evaluation, especially as NONIE's membership expanded to include evaluators under the umbrella of IOCE. Early in 2008, I left IEG for the new International Initiative for Impact Evaluation (3ie). 3ie is committed to enhancing development effectiveness through evidence-based policy making, and will work to both expand the evidence base and make evidence better known. I will come back to how it will do this later in this paper.

I have maintained a 'middle position' in the debate between those promoting quantitative approaches and those calling for a larger range of evaluation approaches to be employed. Whilst this middle ground has remained discouragingly empty, it is a vantage point which allows one to see the strengths and weaknesses of both sides. It also allows one to see that the sides largely speak past one another with very limited engagement. Hence some quite substantial confusion remains. The main purpose of this paper is to clarify some of these confusions. I begin, however, with a brief apparent digression on education in the United States.

2. American Evaluation Association versus Institute of Education Sciences

There are many parallels between the current debates in international development circles and that in the US evaluation community surrounding the activities of the Institute of Education Sciences (IES). The Institute was created by the 2002 Education Sciences Reform Act, partly with the purpose of conducting 'scientifically valid' evaluation of education programs. Scientific validity was defined by the Act as a study which 'employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible'.² In October the following year IES issued a 'proposed priority' for the use of its funds which stated that 'evaluation methods using an experimental design are best for determining project effectiveness... If random assignment is not feasible, the project may use a quasi-experimental design with carefully matched comparison conditions'.³

This proposal was discussed at the annual conference of the American Evaluation Association (AEA) and then on the Association's listserv, EVALTALK. The discussion culminated in a statement from AEA to IES which argued that 'RCTs are not always best for determining causality and can be misleading,' pointing out that 'in medicine, causality

² <http://www.ed.gov/policy/rschstat/leg/PL107-279.pdf>

³ <http://www.eval.org/doi.fedreg.htm>

has been conclusively shown in some instances without RCTs, for example, in linking smoking to lung cancer and infested rats to bubonic plague. The secretary's proposal would elevate experimental over quasi-experimental, observational, single-subject, and other designs which are sometimes more feasible and equally valid'.⁴ But there was a backlash within AEA, with some high-profile members feeling they had not been adequately consulted. A counter 'not the AEA statement' was issued, supporting the IES's stance, asserting that 'Randomized controlled trials have been essential to understanding what works, what does not work, and what is harmful among interventions in many other areas of public policy including health and medicine, mental health, criminal justice, employment, and welfare'.⁵ One of the authors of the counterstatement went so far as to say that 'the AEA now has the same relationship to the Field of Evaluation as the Flat Earth Society has to the Field of Geology'.⁶

This debate is sometimes referred to during current debates in international development, being presented as how the evaluation community faced off an attempt to make RCTs the sole means of evaluating education programs. This view misrepresents the episode in various ways. First, AEA was not united on the issue, with prominent members supporting the IES. Second, the attempts to shift IES were unsuccessful; the 'What works clearing house' of education studies maintained by IES only includes those studies which adopt 'a randomized trial, a regression discontinuity design, or a quasi-experiment with equating of pre-test differences'.⁷ Third, even those opposing IES did not have a fundamentalist opposition to RCTs. Scriven, who was one of the signatories of the AEA statement, agreed that 'we have not used RCTs when we should have many, many times. There have been many occasions when we could have pulled off RCTs, when we could have staffed them with competent people, and this is still the case in the present, and that was the best design around... the theoretical advantages [of RCTs] in validity aspects of it are undeniable'.⁸

But, finally, and of most relevance to current debates, was that the statement can be seen as an overreaction for two reasons. First the IES was not advocating sole use of RCTs, though it did say they should be used when possible. Second, the statement referred to the activities of one division of IES, which had been specifically set up for the purpose of 'scientific evaluation'. It was not making a ruling regarding all evaluation of education programs across the whole US. As we shall see that disquiet about RCTs has been equated with a general opposition to quantitative methods amongst some evaluators in the development debate.

⁴ Journal of MultiDisciplinary Evaluation, Number 3, October 2005, http://www.wmich.edu/evalctr/jmde/content/JMDE%20Num%203_files/Webpages%20JMDE%2003/JMDE_003_Part_I.htm#_Toc116196689

⁵ *Ibid.*

⁶ *Ibid.*

⁷ <http://ies.ed.gov/ncee/wwc/twp.asp>

⁸ *Ibid.*

3. Controversies and confusions

Defining impact evaluation

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

The tradition in evaluation has been that 'impact' refers to the final level of the causal chain (or log frame),⁹ with impact differing from outcomes as the former refers to long-term effects. For example, 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'. Any evaluation which refers to impact (or often outcome) indicators is thus, by definition, an impact evaluation. Hence, for example, 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. 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, $\text{impact} = Y_1 - Y_0$ (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 rigor. 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, whilst making no attempt to attribute changes in those indicators to the intervention. Indeed such studies often explicitly state that attribution is not possible.¹⁰ But this is most decidedly not an impact evaluation to someone for whom

⁹ 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.

¹⁰ Whilst 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."

attribution is the defining characteristic. Many of the objections that you don't necessarily need quantitative methods to do impact evaluations, 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. It is this, second, definition which was intended in the report of the Centre for Global Development, *When Will We Ever Learn?* 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 relabeling. 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 need 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 under-investment in evaluations of this sort, a feeling articulated most clearly in *When Will We Ever Learn?* 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 for attributing observed changes to a specific intervention?* My answer is that, where these methods can be used then they should be. Where they cannot, then other approaches can be used, but they will not usually give the numerical precision which opens up realms of policy relevance.

The current debate revolves around selection bias and how it should be handled. Whilst I agree that this source of bias should be addressed where present, I see this as a technical issue in evaluation design, whereas the main point is to realize the value of having quantitative measures of impact. In the 1960s and seventies, this rationale would not be 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.

These approaches fell into general disuse in the 1980s and early nineties 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.

Whatever the strengths of newer forms of projects and participatory approaches, the consequence of these trends was little reliable evidence has been produced on whether development spending is actually doing any good or not. Let us be clear this is not an outsider's subjective perspective: many studies explicitly state that they cannot make a clear statement as to impact (I have been involved in such evaluations myself in the past). *When Will We Ever Learn?*, listed several reviews which revealed the scanty nature of evidence regarding what works (CGD, 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 (Jerve and Villanger, 2008). 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. It is difficult to avoid the conclusion that there is no evidence to show that most development interventions actually have a significant effect on their intended outcomes, let alone if they do so in a cost effective manner.

Hence there is renewed interest in numerical estimates of program impact. But whilst development evaluators had shifted their attention from quantitative assessments, there were methodological advances in econometrics regarding the problem of 'selection bias'. The bias arises from either program placement (project communities, households, firms etc. are chosen by some systematic criteria) or self-selection (communities or people opt in or out of programs). Now, if the characteristics determining program 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 program. This is not an unlikely situation. If these selection characteristics can be measured, then they can be controlled for statistically. But if they cannot be measured (so-called unobservables) and change over time, then randomization, 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).

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.

In the remainder of this paper 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.

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 are necessarily dealing with counterfactuals, though they may be implicit. It may not always be necessary or useful to make the counterfactual explicit, though 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 discussed above, for economists impact = $Y_1 - Y_0$. 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; i.e. 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, i.e. the same as before the intervention.

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 , e.g. 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 that is 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 - e.g. 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. But there is a counterfactual, but not one which need 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 program 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 analyzing different parts of it. 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%. 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).

However, usually a theory will help. But there can be cases when, as for Semmelweis, 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 programs have been subject to RCTs in several US states and the UK, finding at best no impact. So it is sufficient to note that such schemes simply don't work (though qualitative data have provided reason, which is that many of the young men are actually rather attracted to prison life, plus making some useful contacts). 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, whilst 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 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? IEG's 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. 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 program 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 mentioned above. 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. 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 per cent over the period of the intervention. The impact question is how much of that 12 per cent can be attributed to the project? Even if the outcome is dichotomous, 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%. A good study would dig a bit deeper, and say that since Y changed by P% over the period of the intervention, say, a quarter 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 analysis.

This use of the word 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. 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 – it happens or it doesn't. 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 e.g. 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.

Positivist and linear

Quantitative impact evaluation is accused of being positivist and, apparently, worse still, 'linear'. Regarding the former, the whole business of policy advice appears to fall firmly in the positivist realm. A policy maker wants to hear, "if we do policy X we believe it will have impact Y", not "if we do policy X, we really have no idea what will happen (as it's not possible to say since previous experiences of policy X have been in different times and places so there is no way of knowing if what happened there then and there will happen here and now)". I will return to this issue of generalisability (external validity). But first let's deal with 'linear'.

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 model specifications. It might be argued that X only has an impact in the presence of factor Z, an 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 use of linear critics appears to refer to one-way causation. That is, that 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 program 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, which was discussed above. 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 amongst likely offenders, which does not mean that none will re-offend. Statistical analysis is of stochastic relationships, meaning precisely that there are also some unknown elements ('the error term') which affect outcomes.

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

4. Is there a hierarchy of methods?

The possibility of demonstrating causality through other means is not open to question. The points where critics, in my view, 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 method must be used.

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 tests were conducted on dogs. But although RCTs were not used, quantitative methods were used to compare health outcomes in smokers and non-smokers, whilst controlling for age, sex, occupation, environmental conditions and so on had to be used to link smoking to lung cancer. This is a decidedly quantitative approach, and wholly consistent with IES' statement that such 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 paper, 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. Indeed a theory-based approach will usually combine methods: quantitative and qualitative and evaluation approaches. Hence, I believe the argument between proponents of realistic evaluation and RCTs is over-stated. 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. There is no reason why an RCT cannot be embedded in a theory-based approach, though I realise that at present most are not.

5. Summary

There are two definitions of impact. One refers to outcomes and long term effects, and any analysis discussing these is an impact evaluation definition. The second definition is about attribution. Neither is right or wrong, they are just different; both sorts of impact evaluation can yield information of relevance to policy makers. This paper uses the second definition, arguing that there is a lack of studies of this type. It is also argued that quantitative estimates of impact, where possible, have added policy relevance.

Conducting attribution analysis implies that there is a counterfactual, though it may be left implicit. This does not necessarily mean that there has to be a comparison group, although this is frequently the best way of constructing the counterfactual. Arguments that contribution is more relevant than attribution are misplaced, since attribution does not mean sole attribution, i.e. it does mean contribution.

Finally, none of this means there is a hierarchy of methods. Rather one should adopt the best available method, which will very often, though by no means always, be quantitative.

References

Supercrunchers, Ayres (2007), *Why thinking by numbers is the new way to be smart*, Bantam Books.

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

Ravallion, Martin (2004), 'Comments' in Keith Pitman, Osvaldo Feinstein and Gregory Ingram Evaluating Development Effectiveness, *New Brunswick: Transaction Books*.

Ravallion, Martin (2008), Evaluating Anti-Poverty Programs, T.P. Schultz and John Struass *Handbook of Development Economics Volume 4*.

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

White, Howard (2006), Impact Evaluation: the experience of the World Bank's Independent Evaluation Group, Washington D.C.: World Bank.
<http://ideas.repec.org/p/pramprapa/1111.html>

White, Howard (2008), Of Probits and Participation: The Use of Mixed. Methods in Quantitative Impact Evaluation, *IDS Bulletin VOL 39; NUMB 1*, pages 98-109.

World Bank (1999), Agricultural extension: the Kenya experience: an impact evaluation # 198 World Bank.
<http://lnweb90.worldbank.org/oed/oeddoclib.nsf/InterLandingPagesByUNID/B728D887FC2B754D852568BD005A8C19>

World Bank (2004), Buildings and Learning Outcomes: an impact evaluation of World Bank support to basic education in Ghana [Washington D.C.: IEG, World Bank].

World Bank (2005). Maintaining Momentum to 2015? An impact evaluation of interventions to improve maternal and child health and nutrition in Bangladesh [Washington D.C.: IEG, World Bank].

World Bank (2006), An impact evaluation of the Andhra Pradesh Second and Third Irrigation Projects: poverty reduction with limited economic benefits [Washington D.C.: IEG, World Bank].

World Bank (2007), Welfare impact of rural electrification - The welfare impact of rural electrification: a reassessment of the costs and benefits [Washington D.C.: IEG, World Bank].

World Bank (2008), What works in WSS? A review of impact evidence, [Washington D.C.: IEG, World Bank].