International Initiative for Impact Evaluation

WORKING PAPER 3

Theory-Based Impact Evaluation: Principles and Practice

Howard White
June 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.

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

THEORY-BASED IMPACT EVALUATION: PRINCIPLES AND PRACTICE

Howard White
Executive Director
International Initiative on Impact Evaluation, 3ie
Email: hwhite@3ieimpact.org

Abstract

Calls for rigorous impact evaluation have been accompanied by the quest not just to find out what works but why. It is widely accepted that a theory - based approach to impact evaluation, one that maps out the causal chain from inputs to outcomes and impact and tests the underlying assumptions, will shed light on the why question. But application of a theory- based approach remains weak. This paper identifies the following six principles to successful application of the approach: (1) map out the causal chain (programme theory); (2) understand context; (3) anticipate heterogeneity; (4) rigorous evaluation of impact using a credible counterfactual; (5) rigorous factual analysis; and (6) use mixed methods.

1. Introduction

Recent years have seen increased interest in using quantitative methods to measure the impact of development programs. The work programs of organizations such as the Poverty Action Lab (J- PAL) and Innovations in Poverty Action (IPA), [1] the portfolio of studies financed under the World Bank's Development Impact Evaluation Initiative (DIME) and Spanish Impact Evaluation Fund (SIEF),[2] and the financing being made available by the International Initiative for Impact Evaluation (3ie) [3] mean that there will be hundreds of such studies five years from now, compared to just the handful mentioned in reviews undertaken in recent years (e.g. Centre for Global Development, 2006). However, the mantra of most of those supporting the move toward better impact evaluation is to understand not just what works, but why. Such insight is not given by simply reporting the average treatment effect of an intervention. He nce the statement of the Network of Networks on Impact Evaluation (NONIE): 'the application of the theory-based approach implies that a well designed impact evaluation covers both process and impact evaluation questions. Policy relevance is thus enhanced as the study can address questions of why - or why not - an intervention had the intended impact, not just whether it did' (NONIE, no date). Similarly, 3ie's guide on impact evaluation practice state that 'studies should clearly lay out how it is that the intervention (inputs) is expected to affect final outcomes, and test each link (assumption) from inputs to outcomes (sometimes referred to as the program theory). The evaluation design should incorporate analysis of the causal chain from inputs to impacts' (3ie, no date: 2).

The approach advocated here to understand why a program has, or has not, had an impact is labelled here as theory- based impact evaluation (TBIE). There is nothing new about this. Theory-based evaluation, which means examining the assumptions underlying the causal chain from inputs to outcomes and impact, is a well- established approach (see, for example, Weiss 1998, and Carvalho and White, 2004, for an application in a development setting). Elaborations of program theory have long been used by some practitioners of experimental and quasi- experimental approaches as a way of explaining their findings (Blackman and Reich, 2009: 67- 68). In her paper reviewing possible impact evaluation designs for a range of development interventions, Rogers (2009) notes that a theory- based approach would be appropriate in every case.

Although the commitment to theory -based impact evaluation is there in principle, few studies appear to meet the promise of this approach in practice. This paper is intended to help bridge that gap by laying out the steps, or principles, behind theory- based evaluation. I begin in Part 2 with an example, the Bangladesh Integrated Nutrition Project (BINP), which is then drawn on, with other examples, to illustrate the principles discussed in Part 3. Part 4 briefly compares TBIE with black box approaches and part 5 concludes.

[1] See www.povertyactionlab.org and http://poverty -action.org respectively.
[2] See www.worldbank.org/dime and www.worldbank.org/sief respectively.
[3] See www.3ieimpact.org.

2. An example – the Bangladesh Integrated Nutrition Project

This section provides a brief overview of the evaluation of the Bangladesh Integrated Nutrition Project (BINP). This case is then used to illustrate the principles behind TBIE discussed in the next section. More extended discussion of this project can be found in World Bank (2005), White and Masset (2006), and White (2005).

BINP, modelled on the acclaimed Tamil Nadu Integrated Nutrition Project (TINP) in India, was a growth monitoring project. Infants were weighed weekly at a local weighing station staffed by a village woman trained to be a c ommunity nutrition practitioner. Weight was plotted against age on a growth chart. Children who were growing insufficiently (growth faltering), or who fell too far below the reference norm (malnourished), were admitted to the program. The program consisted of both nutritional counselling and supplementary feeding. However, the project documents were clear that the main impact was expected to come through the counselling. The rationale was that ignorance, rather than poverty, was to blame for poor nutrition, an argument backed up by data showing malnutrition even in the richest quintile, and the existence of beliefs such as 'eating down', that is that a woman should eat less during pregnancy. The program also targeted pregnant women with nutritional counselling and supplementary feeding. BINP was a pilot program, later succeeded by the National Nutrition Program (NNP).

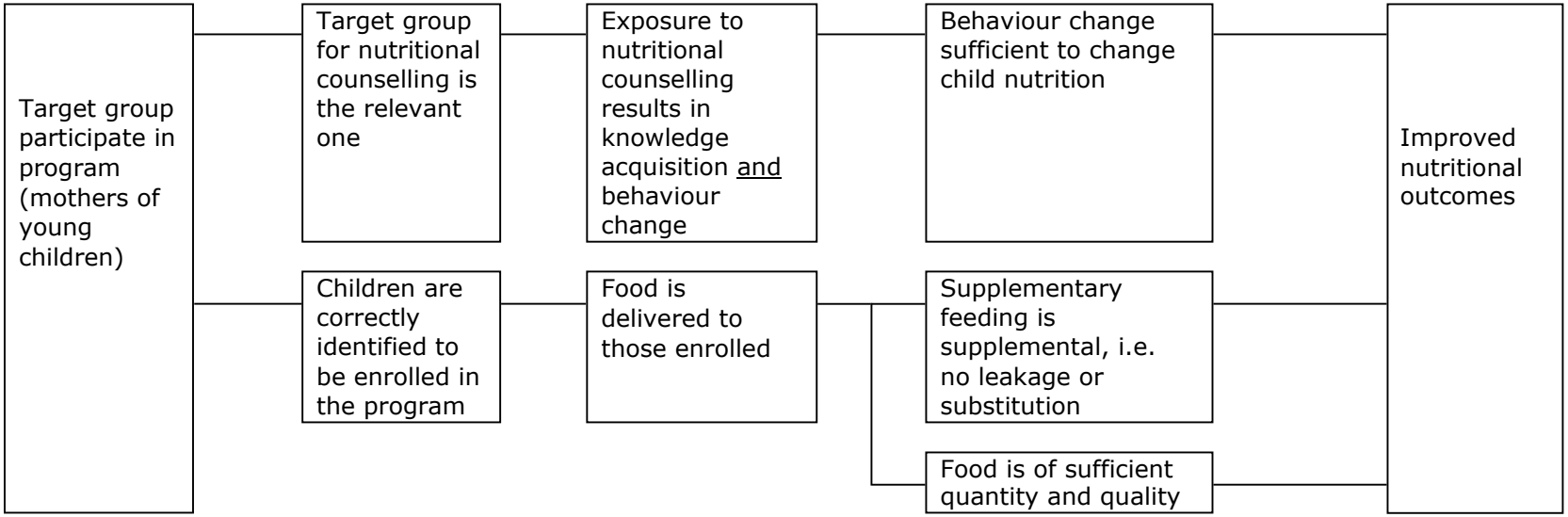
BINP was initially held to be a success. The monitoring data showed substantial falls in malnutrition, notably severe malnutrition, in the project areas. On the basis of this evidence, the Bank decided to go to scale with NNP about mid- way through BINP and prior to any evaluation. Save the Children UK issued a report critical of this decision – reporting their own data from a simple ex- post treatment versus control design which found no difference between the two areas (Save the Children, 2003).

The analysis undertaken by the Bank's Operations Evaluation Department (OED, now the Independent Evaluation Group, IEG), used propensity score matching, combining data from project areas for the treatment with data from a national nutritional survey conducted by Helen Keller International to construct a control group. This analysis found no significant impact of the program on nutritional status, although there was a positive impact on the most malnourished children.

There are rather many assumptions along the causal chain through which BINP may have been expected to have a positive impact on nutritional outcomes, some of which are shown in Figure 1.

A first issue is whether people indeed know about the program and participate – many development projects fall at the first hurdle since insufficient effort is made to explain the intervention to intended beneficiaries, or to make a realistic assessment of the relative costs and benefits for beneficiaries. But BINP did well in this respect, with around 90 percent of eligible women bringing their children, though there were some significant exceptions, as shall be seen below.

Figure 1 - Causal chain for nutrition project: nutritional counselling and supplementary feeding



Second, the people targeted have to be the right ones. The program targeted the mothers of young children. But mothers are frequently not the decision makers, and rarely the sole decision makers, with respect to the health and nutrition of their children. For a start, women do not go to market in rural Bangladesh; it is men who do the shopping. And for women in joint households – meaning they live with their mother- in-law – as a sizeable minority do, then the mother-in- law heads the women's domain. Indeed, project participation rates are significantly lower for women living with their mother- in-law in more conservative parts of the country.

Once women show up with their children to be weighed, the right children have to be admitted to the program, that is those which are growth faltering or malnourished. But the data showed substantial miss- targeting with both Type I (children not being in the program when they should be) and Type II errors (children who should not be in the program being enrolled). We tested the c ommunity nutrition p ractitioners with some sample growth charts (those used in the training), and it turned out that most could not correctly identify from the charts which children should be admitted to the program, hence the mis- targeting. This mattered a great deal for program impact, as we did find that the most malnourished children did benefit, so average impact would have been greater had the progra m concentrated on such children, but in fact resources were going to children who would not benefit.

Furthermore for the supplementary feeding to have a beneficial impact then it has to be supplementary, whereas in fact there was both leakage (the food given to someone other than the person for whom it is intended, this was particularly the case for the supplement given to pregnant women) and substitution (the food was taken in place of a meal that would otherwise have been given).

Returning to behaviour change, the behaviour change communication achieved the communication but not the desired behaviour change. That is women entered into the program did have significantly better knowledge about "good practices". But there was a substantial knowledge- practice gap: a large number of women were not putting this knowledge into practice. The reason was partly resource constraints: women in poorer households were less likely to eat more during pregnancy and those in households with land or living with an elderly male relative were less likely to take more rest during pregnancy. But it was also the mothers-in- law again. One focus group told the investigators explicitly that 'when our mothers- in- law have passed then perhaps we will do these things you are telling us, but until then we will do it the traditional way'. Finally, some behaviour changes – notably those aimed at increasing pregnancy weight gain – were unlikely to have much impact on the final outcome of low birth weight (it is the mother's pre- pregnancy weight which matters most for this).

In summary, project impact was undermined by weak and missing links in the causal chain. Overall, there was no project impact. The improvements shown by the project monitoring data were in fact occurring across the country, which is why Save the

Children found no difference between project and control areas. In fact , this was a trend driven by increased rice yields, higher incomes and the falling price of rice, not BINP .

Nonetheless, the analysis pointed to some clear ways in which program performance could be improved: (1) including mothers- in- law and husbands in nutritional counselling, (2) tighter targeting of the program, and (3) better targeting performance by better training of c ommunity nutrition p ractitioners, and possibly more selective recruiting of the practitioners. However, the calculations also showed it to be a very expensive intervention – one that would be difficult to take to scale on account of both management and resource constraints.

Sadly, these evaluation lessons were not taken on board. The nutrition team of the Bank were very wedded to the TINP/BINP model. It was thought to be a proven success in TINP (though no rigorous study has been conducted by today's standards of rigour), and the Bank was cla iming a success in Bangladesh also, though this claim was disputed. Having first taken part in these debates, a document later put out by the Bank's nutrition team held up BINP as a success without caveats (World Bank, 2006). Fuelled by this belief, the decision was taken to roll out NNP using the same model as BINP, despite evaluation evidence that the model which may have worked in Tamil Nadu needed some adaptation to work in Bangladesh. Three years later NNP was foundering and closed down early, the planned impact study shelved as the lack of impact was evident from weak implementation. The Bank's completion report concluded with recommendations for program reform remarkably similar to those made by OED two years earlier.

3. Principles

The six key principles of a theory - based impact evaluation are:

1. Map out the causal chain (programme theory)
2. Understand context
3. Anticipate heterogeneity
4. Rigorous evaluation of impact using a credible counterfactual
5. Rigorous factual analysis
6. Use mixed methods

Map out the causal chain (program theory)

The causal chain links inputs to outcomes and impacts. That is, the causal chain embodies the program theory (or theory of change) as to how the intervention is expected to have its intended impact. Such a theory is embedded in the traditional log frame, though the latter may not make explicit the underlying assumptions, whereas testing assumptions is central to a theory- based approach.

A common criticism of a causal chain approach that is linear, meaning either unidirectional or presenting a deterministic approach; see White (2009) for a discussion of the different meanings of the word 'linear' in evaluation discourse. But neither criticism is correct. Whilst it may be true that programme managers often do envisage a fairly simple framework linking inputs to activities to outputs to outcomes and impacts, the theory- based evaluation tests the assumptions underlying this chain of reasoning. One such assumed link is that observed outcomes are the result of project activities and outputs, and not *vice versa*. But such reverse, or bi- directional, causality is at the heart of impact evaluation debates: the selection bias caused by program placement and self-selection into the program mean precisely that outcome variables affect who participates, rather than the other way round. For example, communities with high levels of social capital are more likely to apply for funds for community development programs. These programs are meant to build social capital, but simply observing an ex- post difference in the level of social capital between treatment and control villages is more likely to reflect pre- program differences than program impact.

A more valid criticism is that the approach may be rather static, whereas interventions typically adapt and evolve. The systems described in the project document may bear little relation to how the program is being implemented, either because it has been redesigned, or because field managers have taken rather liberal interpretations of project procedures. In the former case, the program theory should reflect the new design, and the evaluation document the learning process that resulted in this design. In the latter case, any discrepancy between what is meant to be done and what is actually done is a key evaluation question: why have these differences emerged, and how do they affect program performance?

An example of project learning comes from the social funds study mentioned above. Another criticism of social fund financed investments is that they are not sustainable since no provision is made for operations and maintenance (O&M). Originally social funds utilized a central committee which approved all applications, it being assumed that line ministries, by their presence in this committee, were committing themselves to meet operational expenses when they agreed to a project. But this system did not work, so social funds began to enter into 'umbrella agreements', to cover all projects for each line ministry. This system also had failings, so some social funds sought line ministry agreement on a case- by-case basis, others required local sustainability plans, whilst others set aside resources for a maintenance fund (see World Bank, 2002).

The program theory should be dynamic in that it allows for learning from the field, which is a restatement of the need to iterate between theory and data. Model- based approaches to statistical analysis take the model as given and simply test how well the data fit the model – and practitioners have various ways of ensuring that data do so fit, as in Coase's statement that the data will confess if you torture them long enough (quote by Leamer, 1983). However, a data analysis approach allows the data to lead the theory, looking for patterns in the data. This approach sounds unst ructured, but of course no statistical exercise can be devoid of theory, since theory guides which data are collected

and analyzed in the first place. Rather theory should be ready to adapt to surprises in the data. This approach may sound akin to data min ing, but it is in fact quite different. The data miner knows what they are looking for and digs the data until they find it. The data analyst, on the other hand, is looking through the data allowing patterns, expected or unexpected, to emerge (see Mukherjee et al., 1998, Chapter 1 for a fuller discussion).

Another possibly valid criticism of this approach is that by focusing on the causal chain the study will miss unintended effects. This weakness can be avoided in two ways. First, a careful application of program theory can identify possible unintended consequences; for example thinking through environmental implications, which may have been dealt with in a rather roughshod manner by the program designers. Second, preliminary fieldwork, including participatory analysis, is an important part of evaluation design which can pick up such unintended outcomes, which can then be incorporated into the evaluation framework.

The question of unintended effects is also linked to the issue of 'whose theory?' A good theory- based design will take into account competing theories as to how a program works. Program managers will have one view, but field staff, beneficiaries or other commentators may have quite different perspectives. For example, social fund projects (development funds disbursed at community level) were argued by program managers to have positive effects on institutional development at local and national level from learning by example (seeing what the social fund did) and learning by doing where these agencies were involved in social fund implementation. However, critics argued that social funds bypassed existing government procedures, thus undermining them directly (by taking staff) and less directly by disrupting optimal resource allocation by line ministries. The evaluation thus considered both the official programme theory and the competing anti- theory (see World Bank, 2002, for the full study, Carvalho et al., 2002 for a summary, and Carvalho and White, 2004 for a presentation of the theory- based approach used).

The usual starting point for putting together the programme theory will be the project documents. If there is a logical framework (log frame) then this framework will embody the programme theory. However, it is unusual for a project document to make explicit all the underlying assumptions, though some of these may appear as 'risks'. A next step is to run the proposed programme theory by programme managers. Even if they had not thought it out explicitly before they will have views on any such document that is produced. This exercise is a good opportunity to engage programme managers allowing them to influence evaluation design in beneficial ways. [4] A second step is to read existing evaluation studies and the academic literature, if any, on the intervention being evaluated or similar programs, which will identify weak links in the causal chain. For

[4] The most usual response of programme managers is that it is not an appropriate time to evaluate the programme because it has just been redesigned, they have just done their own study, there has been a change in government, Minister or project manager etc. These objections should usually be politely ignored, as of course should be any attempt to influence findings. But it makes sense to pick up on what programme managers believe are important evaluation questions.

example, mis- targeting is an oft-cited problem, especially of microfinance programs (e.g. Mosley and Hulme, 1996). A more nuanced point is that micro-finance for women may in fact be utilized by male household members, which affects the impact on final outcomes such as child health and nutrition. The next perspectives to incorporate are those of fieldworkers and beneficiaries. A useful question for any evaluator to ask themselves is 'how will an average villager experience this project? How will they get to know if it? Why would they get involved?' It is useful to try this, though development anthropology has taught us that local perspectives of projects may be very different from what is expected, because of differing perceptions, needs or a simple failure to communicate on the part of project staff.

Understand context

Understanding context is crucial to understanding program impact, and so designing the evaluation. Context means the social, political and economic setting in which the programme takes place, all of which can influence how the causal chain plays out. The impact of an identical program can differ in different contexts: as with the apparently successful TINP model not working so out so well in Bangladesh. However, 'identical programs' are something of an ideal, rarely achieved in messy field conditions – which is in itself an important part of context. Furthermore, as outlined below, an understanding of context will help anticipate heterogeneity, it will also help generalization.

Understanding context means a thorough reading of project documents prior to embarking on evaluation design, but also exposure to a broader literature (anthropology and political economy), as discussed under the use of mixed methods below.

Understanding context also helps generalization. Studies of World Bank support to basic education in Ghana and of maternal and child health in Bangladesh were overall success stories. In the Ghanaian case large scale school rehabilitation and textbook provision had made significant contributions to improved enrolments and learning outcomes (World Bank, 2004). There were two important contextual aspects behind this result. First, was that, following years of crisis, the school system was in a very sorry state indeed, with inadequate infrastructure and virtually no school supplies. School renovation and textbook supply had an impact in this context which it may not have done had schools already been relatively well functioning. Second, there was strong political support for the program, which helped ease implementation (the program was part of a wider educational reform). Government commitment was also a key ingredient in the success of the aid- financed planning which resulted in an accelerated demographic transition in Bangladesh with dramatic falls in mortality and fertility (World Bank, 2005). The country went from next to no facilities immediately after independence, being written off as a basket case in the subsequent famine, to having a nation-wide decentralized health and family planning system, down to doorstep delivery of contraceptive services, in a ten year period. Similarly ambitious programs may falter if government does not have the will to see them through.

Anticipate heterogeneity

Understanding context helps anticipate possible impact heterogeneity. Impact (that is the treatment effect) can vary according to intervention design, beneficiary characteristic or the socio- economic setting. Examining the underlying theory can help expose possible heterogeneity and allow the evaluation design to anticipate it. Anticipating likely heterogeneity matters for two reasons. First, the power calculations for sample size need reflect the levels of disaggregation that will be used in the analysis: the greater the degree of disaggregation the larger the required sample (for both treatment and control). Second, simple probability suggests that if we test for impact in twenty different, arbitrarily defined, sub- groups then we will find a significant impact in one of those at the five percent level. Good practice, required for medical RCTs, requires that the sub-groups to be tested are defined before data collection. The theory- based approach assists in the pre- identification of such groups and provides a plausible explanation for such differential impact. There is, however, a caveat arising from the need to iterate between model and data.

Consider child feeding programs, malnourished children are more likely to respond with weight gains than are already well nourished children, though extremely poorly nourished children may have diarrhoea which prevents effective feeding and weight gain. Better targeted programs will thus have a higher average impact, and that impact will be greatest in the lean season – as was indeed found to be the case with BINP. Younger children are likely to benefit more; children who have suffered stunted growth in infancy will not experience marked height gains from feed ing in later years. Similarly cognitive gains from better nutrition appear to be captured under three years of age. Hence, impact varies by beneficiary age and pre - existing nutritional status, the latter having a seasonal element. Impact can also vary according to socio- economic status; for example, substitution (using 'supplementary feeding' to replace an existing meal) is more likely in poor households.

For these reasons the trend in feeding programs has shifted away from school- based feeding to targeting those under the age of three, such as the program in Bangladesh discussed in detail above. But school feeding can still be expected to yield learning gains. Calorie deficiency makes children tired and listless, so a feeding program can make them more attentive in class; with the caveat that most people are sleepy after a good meal, so timing matters. But the setting matters for learning gains to be realized from more attentive children. A crucial assumption for all interventions is that the right constraint is being tackled. An attentive child is no use if the teacher is absent, and will likely learn less if there are no learning materials. So the impact of feeding programs can be expected to be greater in well- functioning schools than in poorly equipped ones in which teacher absenteeism is rife. A similar point has been made with respect to conditional cash transfers, which increase demand for schooling, but may not improve learning outcomes, or even enrolments, if there are supply - side constraints (Ravallion, 2009).

Another aspect of heterogeneity is the possible complementarity between interventions; for example microfinance has a large impact if accompanied by business support services. Or possibly the two are substitutes, where the impact of the two c ombined is less than the sum of the two separately. Designs that explore such complementarities are clearly of great policy relevance.

Impact can also vary across time, despite the usual (often implicit) assumption of linear impact trajectory (Woolcock, 2009). A linear impact trajectory is different from the previously-discussed criticisms of unidirectional causality or of being a static approach to a continuously changing program implementation. Even when the program design remains unchanged and the causal direction has been established, the impacts of the intervention may change over time and findings will be very sensitive to the point in time in which impact is measured. For example, for projects that try to increase the participation and empowerment of marginalized groups the literature suggests that the most likely shape of such projects' impact over time is a J curve; that is, things get worse before they get better. T his is an area that has not been sufficiently explored using TBIE, but one to which it lends itself particularly well. In the case of the BINP program discussed earlier, the program may have induced initial conflicts between the women and their husbands and mothers- in- law because of the increased awareness of the women, something which may explain no nutritional impacts, but possibly a longer- term evaluation may detect a positive effect given the broader social changes increasing the status of women in rural Bangladesh.

Identifying heterogeneity is linked to generalisability. An RCT in Ke nya, South Africa and Uganda tested the impact of male circumcision on the transmission of HIV/AIDS, finding that circumcised men were significant less likely to contract the disease [see, for example, Wawer et al., 2008, on Uganda]. One aspect of heterogeneity was age. There should be one month abstinence following circumcision so the wound can heal; sex in that period is higher risk, not lower risk. Carrying out the procedure on, say, 12- year old boys does not carry this risk of one month's high- risk exposure. But older males they are often unable to abstain for a whole month, thus reducing the beneficial impact of the treatment. Nonetheless, the studies found a reduction in the risk of transmission between 30 and over 50 percent as a result of circumcision. This level of impact can only be generalized to populations with similar patterns of sexual behaviour. In a community in which men practiced abstinence, single-partner relationships or universal condom use then there would be no impact from the interve ntion.

Impact

Rigorous evaluation of impact using an appropriate counterfactual is of course a key component of TBIE. The appropriate counterfactual is most usually defined with reference to a control group, which has to be identified in a way in which avoids selection bias, meaning the use of either experimental or quasi-experimental approaches. Having panel data helps strengthen the design, so baselines – designed in such a way to allow re - identification of sample households – are to be encouraged. Where they are not available

they might be recreated using existing data sets or recall, though caution need be exercised with the latter (see Bamberger, 2009). In addition to selection- bias, important issues to consider in the design are the possibility of spill over effects (the control is affected by the intervention) and contagion or contamination (the control is affected by other interventions)

Rigorous factual analysis

The counterfactual analysis of impact needs to be supplemented by rigorous factual analysis of various kinds. Many of the links in the causal chain are based on factual analysis. In the case of BINP this included poor targeting and the reasons for it, identification of leakage, and the fact that improved knowledge was not turned into practice.

Targeting analysis is the most common form of factual analysis which should be a part of most, if not all, impact studies: who benefits from the program? To the extent that there is a defined target group, then what is the extent of the targeting errors; such errors can be quantified and their source identified, as was done in Bangladesh. Targeting analysis should be carried out at different levels. In the case of social funds it was found that the use of poverty maps meant that social funds in many countries were focused on the poorest districts, but that within those districts it was the better off communities which were more likely to access project resources (World Bank, 2002). Conversely, in the case of rural electrification, better off communities were more likely to connect, but poorer households in connected communities remain unconnected for many years on account of their inability to afford the connection charge (World Bank, 2008).

Targeting analysis has to be done using a representative data set. By collecting samples which allow for selection bias, impact evaluation datasets are often not representative of the population as a whole and so cannot be used to answer a question such as 'what percentage of the poorest 20 percent benefit from the project?' unless sampling weights are available to make the sample representative.

A second point with respect to targeting is that it is a bivariate exercise, requiring plotting or tabulating participation against the characteristics of interest (these characteristics may be individual, household or community ones). A quasi- experimental approach requires a multivariate analysis of program participation, but it is generally a mistake to use these results for the targeting analysis which should rely on the descriptive statistics. Whether the program reaches the bottom 20 percent is a statement based on a bivariate tabulation, not the statistical significance of a quintile term in a multivariate regression. What the regression can do is highlight the factors which do drive participation and so help explain the bivariate targeting outcomes. For example, a multivariate analysis of a project in India might show significant lower participation from

tribal populations, which are amongst the poorest in some program areas, thus explaining poor targeting performance.[5]

An example of an under- used form of factual analysis is testing whether people who have been exposed to training have learned, and put into practice, new approaches as intended by the training. The BINP study showed that mothers acquired knowledge but many did not put it into practice. And community nutrition practitioners could conduct weighing sessions, but, crucially, had not learned to correctly interpret the growth charts. Such an analysis is not often done, but clearly there is great scope for it. Do trained teachers know about improved teaching methods and put them into practice? A World Bank study in Ghana suggested that many do not.[6]

As in the Bangladesh case, factual analysis can often highlight a crucial break in the causal chain and so explain low impact. Another OED study found the training and visit agricultural extension service in Kenya to have no impact on yields. In principle the project funded new agricultural research in research stations, the lessons of which were passed onto extension workers and then to farmers. In practice the lessons from the research were not passed onto extension workers, who were giving messages to farmers telling them to adopt practices most had already adopted long ago (World Bank, 2000).

However, sometimes what appears to be a requirement for a factual analysis, may in fact require a counterfactual. A school capitation grant is intended to increase both enrolments and learning outcomes. But how does it do this? Such an explanation must surely rest on the uses of the money. This might sound like a straightforward factual analysis – tracing the use of funds: checking how much does indeed reach schools and how it is spent, both of which are indeed useful parts of the study. However if schools already had some resources at their disposal then there is the possibility of fungibility. A before versus after analysis of spending patterns might yield a valid counterfactual in this case, though a treatment versus con trol analysis of school improvements and materials acquisition is likely to be a stronger design.

Mixed methods

'Mixed methods' is the combination of qualitative and quantitative approaches in a single evaluation. All quantitative studies have some measure of qualitative analysis – at least reading the project documents – so it is a question of degree.

---

[5] More specifically a wealth term is significant when a tribal dummy is excluded, but becomes insignificant once the latter variable is included. Hence it is tribal status rather than poverty per se which is driving participation. This approach may not always be possible on account of the high degree of co-linearity amongst the possible explanatory variables, such as those mentioned and education and location.

[6] Classroom observation would be the best way to measure practice, but was excluded on the grounds of expense. One might think that simply asking teachers about methods would be biased since they would report using improved methods even if they don't, but in practices teachers proved either surprisingly candid, or their lack of knowledge was such that they did not know which were the 'right answers'.

The call for mixed methods generally comes from proponents of qualitative approaches. But in the development field qualitative approaches have dominated evaluation until very recently, so a major step toward mixed methods is in fact the increased use of rigorous quantitative methods in qualitative studies. However, here I pay attention to increasing the use of qualitative data in quantitative studies, an issue I have dealt with at more length in White (2008). I will make three general points.

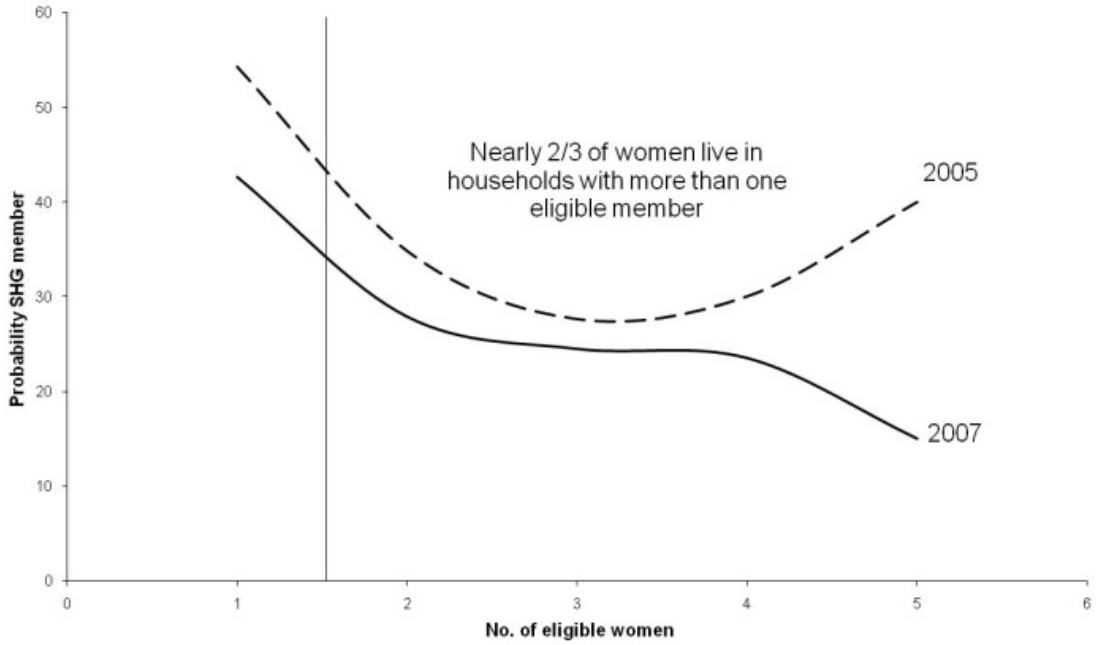
First, use of qualitative data means a wide range of activities, not just arranging for some focus groups (in my view one of the weaker forms of qualitative data, unless done really well). It includes, for example, reading of anthropological and political literature of the intervention context to inform evaluation design. In the Bangladesh case, identification of the 'mother- in- law' effect came from reading anthropolo gical literature (notably White, 1992). This insight led us to unpack the household roster section of the questionnaire to identify those women living with their mother in law (e.g. daughter-in- law of household head, where spouse of household head also pre sent, spouse of household head, where mother of head also present, and sister- in- law of household head, where mother of head also present), and so carry out quantitative analysis informed by a qualitative insight.

The range of techniques goes from 'develo pment tourism' (spending a day or so in the field), through the toolkit of PRA, to embedding an anthropologist in the project area, the latter being an under- exploited approach which could be used on longer-term studies. My second point is that, although development tourism is much derided, it is an essential part of TBIE. There really is no substitute for spending time in the field yourself, and it is difficult to know how data can be sensibly analyzed without this field exposure (it shows when it is the case). Spending just a few days being exposed to project implementation in a range of settings – and very preferably not just those chosen by the project staff – will help both design and implement the study. It is also useful to visit non- project areas.

I could give many examples of insights from discussions with project staff, beneficiaries and other stakeholders in the field. I will give just two. The first is from an evaluation of a rural livelihoods project which included loans through women's self - help groups. One man complained that his unmarried daughter of 22 could not get a loan since his wife had already received one. This comment led to the insight that the villagers regard the loan to the household not to the individual, a fact which drove the much lower participation rates of women in households with more than one woman eligible to join a SHG (Figure 2). The project's aim is that all eligible women should participate, but this is not a realistic goal so long as benefits are at the household, not individual, level. The second example is of how a well- chosen quote can powerfully make a general point. In the fieldwork for the Zambian social fund evaluation it was striking how everyone – from managers, to program staff, to villagers – would say 'the community' chose the project, although it was evident that a more selective process was actually at work (see White and Vajja, 2008, for a longer discussion). However, the fact that 'the community' was in fact a rather narrow construct, meaning the projec t committee, was well captured by a

regional program officer who, answering his mobile phone, told us 'I have to go now, I have a community in my office'.

Figure 2 - Program participation rates in Self Help Groups in Andhra Pradesh by number of eligible women in the household



Source: IEG survey data

Since data sets are available which allow impact evaluation without new data collection (something which is to be encouraged, since we are too ready to collect new data whilst existing data sets are under- exploited), there is a danger of researchers conducting impact studies with no exposure at all to the intervention. Such studies are very likely to lack policy relevance owing to weak understanding of how the intervention actually works.

Finally, the budget should allow for some action research type activities, where puzzles in the data are followed up with additional field work. The focus groups on the reasons for the knowledge practice gap in Bangladesh are an example of such work. Another example comes from the just -mentioned study of finance to Self Help Groups in Andhra Pradesh in India. We had panel data, and the survey included a standard LSMS - type module on household enterprises. Analysis of these data showed low returns to most enterprises, including a significant minority of loss making activities. But the data in the module were really too blunt an instrument to understand how these enterprises were functioning. Hence, we commissioned what I would call some 'quantitative ethnography' to re- visit all the households that had been surveyed and declared an enterprise. The re - visits used a semi- structured questionnaire to identify the daily cash flow of the enterprises and labour inputs from household members (and employees, though these

were very rare). The results indeed confirmed the low level of income from these activities (Rs. 20- 30 a day was not unusual, compared to a daily wage of Rs. 50- 70), and the risky nature of many (livestock death, especially of goats, and insufficient market size).

4. TBIE versus black box approaches

Theory- based impact evaluation may be contrasted with a 'black box' approach. The latter often simply reports an impact – being interested in the statistical significance of the coefficient for the average treatment effect, but makes no attempt to answer the why question. This paper has sought to show how to tackle the why question and the benefits of doing so. However, some caveats are in order.

Criticisms of reporting an average treatment effect should not be overstated. Heterogeneity matters, as does understanding the context in which a particular impact has occurred. But it will rarely be the case that the average treatment effect (usually both the treatment of the treated and the intention to treat) is not of interest. Indeed it is very likely to be the main parameter of interest. It would be misleading to report significance, or not, a particular sub- group if the average treatment effect had the opposite sign. Moreover the average treatment effect is the basis for cost effectiveness calculations.

Second, TBIE unpacks the causal chain in various ways. It tries to disentangle the various stages of the causal chain, but also which bits of an intervention work and which bits don't. This might be done through regression analysis. For example, the BINP study presents regressions on the determinants of the knowledge- practice gap. But such regression-based approaches, which rely on sample selection models and parametric specification of the relationship being examined, have many critics, who favour either experimental or quasi- experimental approaches such as propensity score matching and regression discontinuity design. These rigorous approaches can accommodate analysis of which bits of the program work, but the intervention has to be set up to allow intervention design to vary across groups – e.g. some entrepreneurs get loans, some get business support services and some get both. In practice a TBIE will combine such rigorous impact estimates as can be made with other approaches to unpacking the causal chain.

Finally, what is inside the black box may be so messy that it is sometimes best left unopened? The World Bank study of rural electrification examined the impact of electrification on fertility. Access to electricity does significantly reduce fertility (World Bank, 2008). The study was able to demonstrate one possible channel which seemed to be at play (access to television increasing contraceptive knowledge), and one which was not ('alternatives to sex' reducing sexual activit y). But there are many other possible channels, such as income effects, other educational benefits and so on. In such cases,

where all channels cannot be separated out, then a reduced form impact estimate can be the best way to go.

5. Conclusions

This paper endorses calls to produce a greater volume of rigorous quantitative studies of what works in development. However, the policy relevance of such studies will be far greater if they also shed light on why interventions to, or do, not work. It is widely agreed that theory- based impact evaluation (TBIE) can yield the necessary insights. However, many new studies fail to meet the promise of the theory- based approach, making speculations as to the reasons for impact, or differences in impact, rather than having a solid empirical analysis to explain them.

I have presented an example of a TBIE in practice, and how the approach leads directly to policy conclusions to enhance program impact. Doing this required application of a set of principles which are elaborated above. The program theory need be elaborated in a flexible way, ready to adapt to changing circumstances in the field, and to take on board competing theories and unintended consequences. Rigour needs to be combined in both factual as well as counterf actual analysis, which will mean using a mix of methods. The program theory has to be set in the social, political and cultural context of the intervention, which will be one means of highlighting expected heterogeneity of impact.

Acknowledgment

The aut hor thanks Marie Gaarder for comments on an earlier version of this paper. The usual disclaimer applies. The views expressed here are those of the author and cannot be taken as those of 3ie, or any of its members or supporters.

References

3ie guide for grantees (no date) '3ie impact evaluation practice: a guide for grantees', http://www.3ieimpact.org/page.php?pg=overview (accessed June 1, 2009).

Bamberger, Michael (2009) 'Strengthening the evaluation of program effectiveness through reconstructing baseline data' *Journal of Development Effectiveness* **1**(1): 37- 59.

Blackman, Leonard and Stephanie Reich (2009) 'Randomized control trials: a gold standard with feet of clay?' in Stewart Donaldson, Christina Christie and Melvin Mark (eds.) *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* [Thousand Oaks, California: Sage].

Carvalho, Soniya, Gil Perkins and Howard White (2004) 'Social funds: participation, social capital and sustainability' *Journal of International Development* **14** 611- 625, 2002.

Carvalho, Soniya and Howard White (2004) 'Theory- based evaluation: the case of social funds' *American Journal of Evaluation* **25**(2) 141-60, 2004.

Centre for Global Development (2006) *When Will We Ever Learn?* [Washington D.C.: Centre for Global Development].

International Initiative for Impact Evaluation, 3ie (no date) '3ie Impact Evaluation Practice: a guide for grantees' http://www.3ieimpact.org/doc/3ie%20impact%20evaluation%20practice.pdf (accessed June 1, 2009).

Leamer, E. (1983) 'Let's take the con out of econometrics', *American Economic Review*, **23**(1), 31 –43.

Mosley, Paul and David Hulme (1996) *Finance Against Poverty* [London: Routledge].

Mukherjee, Chandan, Marc Wuyts and Howard White (1994) *Econometrics and Data Analysis for Developing Countries* London: Routledge.

NONIE (no date) 'NONIE statement on impact evaluation' http://www.worldbank.org/ieg/nonie/members.html (accessed June 1, 2009).

Ravallion, Martin (2009) 'Evaluating three stylized interventions', *Journal of Development Effectiveness* **1**(3).

Rogers, Patricia (2009) 'Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation' *Journal of Development Effectiveness* **1**(3).

Save the Children (2003) *Thin on the Ground. Questioning the evidence behind World Bank-funded community nutrition projects in Bangladesh, Ethiopia and Uganda* . [London: Save the Children UK].

Weiss, Carol (1998) *Evaluation: methods for studying programs and policies*. Prentice Hall: New York.

Wawer M, Kigozi G, Serwadda D, et al. Trial of Male Circumcision in HIV+ Men, Rakai, Uganda: Effects in HIV+ Men and in Women Partners. 15th Conference on Retroviruses and Opportunistic Infections; 2008; Boston, MA; 2008.

White, Howard (2005) 'Comment on Contributions Regarding the Impact of the Bangladesh Integrated Nutrition Project' *Health Policy and Planning* **20**(6), 408- 411.

White, Howard (2008) 'Of Probits and Participation: the use of mixed methods in quantitative impact evaluation" *IDS Bulletin*, 2008.

White, Howard (2009) 'Some reflections on current debates in impact evaluation' *3ie Working Paper No. 1* [New Delhi: International Initiative for Impact Evaluation].

White, Howard and Edoardo Masset (2006) 'The Bangladesh Integrated Nutrition Program: findings from an impact evaluation *Journal of International Development* **19**: 627- 652, 2006.

White, Howard and Anju Vajja (2008) 'Can the World Bank Build Social Capital?: Community Participation in Social Funds in Malawi and Zambia' *Journal of Development Studies* **44**(8): 1145- 1168.

White, Sarah (1992) *Arguing with the crocodile: gender and class in Bangladesh*, London: Zed.

Woolcock, Michael (2009) 'Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy' *Journal of Development Effectiveness* **1**(1): 1- 14.

World Bank (2000) *Agric ultural extension: the Kenya experience* [Washington D.C.: OED, World Bank].

World Bank (2002) *Social Funds: assessing effectiveness* [Washington D.C.: OED, 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.: OED, World Bank].

World Bank (2006) *Repositioning Nutrition as Central to Development: a strategy for long term large-scale action* [Washington D.C.: World Bank].

World Bank (2008) *The welfare impact of r ural electrification: a re-assessment of the costs and benefits* [Washington D.C.: IEG, World Bank].