International Initiative for Impact Evaluation

3ie

WORKING PAPER 13

Can we obtain the required rigour without randomisation? Oxfam GB's non-experimental Global Performance Framework

Karl Hughes
Claire Hutchings
August 2011

About 3ie

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

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

This Working Paper was written by Karl Hughes and Claire Hutchings, Oxfam GB

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

Contents

CAN WE OBTAIN THE REQUIRED RIGOUR WITHOUT RANDOMISATION? OXFAM GB'S NON-EXPERIMENTAL GLOBAL PERFORMANCE FRAMEWORK

Karl Hughes
Global Programme Effectiveness Advisor, Oxfam GB; PhD Candidate, London School of Hygiene and Tropical Medicine

Claire Hutchings
Global Monitoring and Evaluation and Learning Advisor (Campaigns), Oxfam GB

Abstract

Non-governmental organisations (NGOs) operating in the international development sector need credible, reliable feedback on whether their interventions are making a meaningful difference but they struggle with how they can practically access it. Impact evaluation is research and, like all credible research, it takes time, resources, and expertise to do well, and – despite being under increasing pressure – most NGOs are not set up to rigorously evaluate the bulk of their work. Moreover, many in the sector continue to believe that capturing and tracking data on impact/outcome indicators from only the intervention group is sufficient to understand and demonstrate impact. A number of NGOs have even turned to global outcome indicator tracking as a way of responding to the effectiveness challenge. Unfortunately, this strategy is doomed from the start, given that there are typically a myriad of factors that affect outcome level change. Oxfam GB, however, is pursuing an alternative way of operationalising global indicators. Closing and sufficiently mature projects are being randomly selected each year among six indicator categories and then evaluated, including the extent each has promoted change in relation to a particular global outcome indicator. The approach taken differs depending on the nature of the project. Community-based interventions, for instance, are being evaluated by comparing data collected from both intervention and comparison populations, coupled with the application of statistical methods to control for observable differences between them. A qualitative causal inference method known as process tracing, on the other hand, is being used to assess the effectiveness of the organisation's advocacy and popular mobilisation interventions. However, recognising that such an approach may not be feasible for all organisations, in addition to Oxfam GB's desire to pursue complementary strategies, this paper also sets out several other realistic options available to NGOs to step up their game in understanding and demonstrating their impact. These include: 1) partnering with research institutions to rigorously evaluate "strategic" interventions; 2) pursuing more evidence informed programming; 3) using what evaluation resources they do have more effectively; and 4) making modest investments in additional impact evaluation capacity.

1. NGOs and the Effectiveness Challenge

Few of us in the international NGO community would argue against the importance of accessing trustworthy feedback on whether the interventions we implement and/or support are making a meaningful difference. Such feedback can give us confidence that we are on the right track or encourage another visit to the drawing board. It can also help prove the worth of our work to donors and supporters and even motivate others to model our efforts, thereby, leveraging additional change. However, the core issue that dumbfounds many of us is how to access credible intervention effectiveness feedback – *practically*.

Since the earlier half of the 20th century, the importance of impact evaluation has been increasingly recognised from accountability, programme improvement, and knowledge generation perspectives (Valadez and Bamberger 1994; Weiss 1998; Picciotto 2003). However, assessing the effectiveness of development and, perhaps to a greater extent, policy interventions is typically not straightforward, and it has been the subject of hotly contested debate in the evaluation literature (Cracknell 2000; Wynn, Dutta et al. 2005). The roots of much of the debate lie in differing philosophical perspectives on how knowledge is acquired (epistemology), particularly in social settings. At one end of the spectrum are the "positivists." They view the randomised controlled trial (RCT) as the "gold standard" for evaluating the effectiveness of social programmes and judge the merit of other evaluation designs by how close they come to replicating it (Newman, Rawlings et al. 1994; Zhu 1999; Shadish, Cook et al. 2002). The "constructionists," on the other hand, question the objective measurement of social change and emphasise the use of qualitative and participatory methods to "socially construct" an evaluation's findings (Cronbach 1982; Guba and Lincoln 1989; Lay and Papadopoulos 2007).

However, increasingly, evaluators are seeing the limitations in confining themselves to the boundaries of either camp. There can be, for instance, considerable barriers to undertaking RCTs – ethical, legal, political, financial, and/or practical (Rossi, Lipsey et al. 2004; Bamberger, Rugh et al. 2006). They are also no panacea; such designs, for instance, are inappropriate for evaluating "small n" interventions (White 2010) or those that are either underdeveloped or evolving (Veerman and van Yperen 2007). On the other hand, many question the ability of constructionist approaches to address attribution issues and untangle the effects of an intervention from change(s) that might have taken place anyway. While the "paradigm wars" may not be entirely over, many evaluators have attempted to move things forward by acknowledging the merit of both quantitative and qualitative approaches and advocating for "mix-methods" (Greene, Benjamin et al. 2001; Chen 2005; Voils, Sandelowski et al. 2008; White 2008).

Unfortunately, however, employing mixed-methods is not, by itself, a magic bullet; we still must get our hands dirty if we are to access the credible effectiveness feedback we are after. Complementing the "traditional" NGO approach of administering questionnaires to our "beneficiaries" with focus group discussions and participatory exercises or vice-versa is not going to cut it. Taking a mixed-methods approach does not liberate us from the need for rigour and, perhaps most importantly, irrespective of the approach we take, we must come to terms with the fact that there are no shortcuts. Evaluation is research, and, like all credible research, it takes time, resources, and expertise to do well. This is true for both quantitative and qualitative approaches.

2. Flirting with global outcome indicators

2.1 How to Demonstrate Effectiveness Ineffectively and at Great Cost

With an increasingly sceptical public, the need for governments to justify and/or cut aid budgets, and pressure from donors to demonstrate "results" and "value for money," we have little choice but to step up our game. We are being challenged by the relative success of organisations such as the Abdul Latif Jameel Poverty Action Lab (J-PAL) and the International Initiative for Impact Evaluation (3ie) in promoting more rigorous impact evaluation in general and RCTs in particular. It is no longer acceptable for us to maintain the status quo or say that what we do is intrinsically not evaluable. The challenge remains how can international NGOs – who often work in numerous countries and diverse contexts and pursue "bottom-up" programming – reliably understand and "sum up" their effectiveness, particularly when their core business is not research.

What can be dubbed as "global outcome indicator tracking" is one popular option being pursued by several international NGOs who are attempting to improve their ability to demonstrate their effectiveness. This essentially involves defining core impact/outcome indicators to measure specific improvements in people's lives and other change variables, and then having all relevant projects/programmes systematically capture data on these indicators. Here is an example of such an indicator:

- *% of people whose net income generated within target value chains has increased, by sex*

The idea, then, is for income-related data to be periodically collected in relation to all the organisation's value-chain enhancement interventions from representative samples of supported producers. All of these producers – taken together – in effect would be a global cohort, and an assessment is then made on the extent to which their income from participation in targeted sectors has changed over time. The status of the indicator could, of course, be disaggregated by country, region, sex, value chain, etc.

So what would this information tell us? Let's say, for example, that baseline data were captured on this indicator in the near future, and – over the course of the next few years – the global economy expands considerably and disproportionably in favour of lower income countries. Would a progressive improvement in the status of the indicator over time indicate changes brought about through the organisation's support or something that would have simply happened anyway? From the vantage point of the counterfactual account of causality – now in widespread use in the social sciences (Morgan and Winship 2007) in general and impact assessment in particular (Khandker, Koolwal et al. 2010) – the answer is no, particularly without the backing of complementary evaluative evidence. While it is relatively straightforward to link outputs, e.g. boreholes constructed, civil servants trained, etc., to workings of an intervention and, by extension, an agency, this becomes more difficult as we move up the causal chain (White 2005). Outcome/impact level changes are affected by numerous factors, e.g. sectoral trends, external events, and maturation. As such, simply observing positive change in an outcome indicator, even following the successful implementation of an intervention, is insufficient to evidence that this intervention, in particular, was responsible for causing the change. Moreover, one can imagine the level of effort, complexity, and costs associated with collecting, quality controlling, and aggregating global outcome data. While it may be possible to generate some interesting statistical trends, it would certainly not enable an organisation to credibly demonstrate *its* effectiveness.

2.2 Setting Out on a Road Once Travelled

Like many other NGOs, Oxfam GB seriously considered adopting a global outcome indicator tracking approach. This began in 2006 with the "Indicator Feasibility Study," which set out to develop and test a fairly large number of global outcome indicators – 34 in total, spread over the organisation's five strategic aims. Ten projects were identified to field test the indicators over a period of two years. "These indicators were meant to measure the outcomes and impact of the majority of our programmes in a wide range of contexts, and could be synthesized and further analyzed to obtain a more aggregate picture of Oxfam's impact as an organisation"(Shroff and Stevenson 2008). However, before any assessment could be made on the value of these data to the organisation, the study failed for practical reasons. During its first year, only half of the projects collected data; three dropped out due to "human resource issues," one experienced funding constraints, and the staff members of the final project did not find any of the indicators relevant (Ibid).

The Indicator Feasibility Study was subsequently abandoned, given a senior management steer to direct evaluation efforts to other organisational priorities. In the intervening years, the pressure for Oxfam GB to demonstrate its effectiveness as an organisation only intensified, coinciding with changes in the UK's political landscape in general and escalating debates on aid effectiveness in particular. In response, the global outcome indicator tracking approach resurfaced.

3. Working out a workable compromise: Oxfam GB's global performance framework

3.1 But It Ain't Just about Indicators!

Coming up with Oxfam GB's new global indicators was the main agenda item for the semi-annual meeting of the Internal Division's Programme Leadership Team (PLT) in the summer of 2010. This time, however, senior leaders from headquarters and the regions, rather than technical specialists, were asked to get involved in identifying a much smaller number of outcome indicators, to encourage their buy-in and ownership and ensure that the initiative would be taken as a serious corporate priority.

From the perspective of Oxfam GB's Programme Performance and Accountability Team (PPAT), the key challenge was to ensure that indicators were understood as being only a small part of the effectiveness puzzle; that evidencing effectiveness is as much, if not more, about the evaluation methods that underlie indicators, rather than the indicators themselves. To this end, we wrote a background paper for PLT, reviewing the pros and cons of different possible approaches to evidencing effectiveness, including global outcome indicator tracking.

However, while many understood the wider challenge, simply identifying an initial list of indicator areas proved significantly challenging. Our team was subsequently tasked with shaping the event's outputs into more polished, conventional indicators in collaboration with the organisation's programme policy advisors. These indicators are presented in Annex 1.

3.2 Project Effectiveness Auditing: An Alternative Way of Operationalising Global Indicators

Once the indicators were agreed, we were tasked with proposing a measurement approach that would allow the organisation to credibly "sum up" its effectiveness on these indicators. Demonstrating the impact of our organisations would be challenging but technically possible if all of our various interventions were rigorously evaluated as part of standard practice. In particular, we could commission systematic reviews, using appropriate means of bringing together the findings and lessons from evaluations by thematic area. For large *n* interventions, in particular, we could use similar quantitative measures for selected outcomes and then statistically aggregate these findings through meta-analysis (Greenhalgh 1997; Higgins, Green et al. 2005). Unfortunately, most of our organisations do not have the evaluation basics in place needed for systematic reviews, leave alone meta-analysis. Some of our organisations may be ahead of the game, having carried out rigorous evaluations on specific interventions. However, these interventions are likely unrepresentative. As such, synthesising the findings of these evaluations would result in biased representations of overall organisational effectiveness.

Reflecting on these issues resulted in our team's proposal to senior management: We continue on our quest to strengthen monitoring, evaluation, and learning (MEL) for all our projects and programmes. However, with approximately 400 projects closing in any given year, it would be difficult to ensure that all are evaluated to acceptable standards, at least in the short to medium term. Narrowing in on a smaller sample, then, was found to be the only feasible alternative. We were then confronted with two issues: First, if we were to *purposively* select the sample of projects, we could be accused of "cherry picking." Randomly selecting the sample was, therefore, agreed as the only credible way forward. Once this was agreed, a second issue arose around the timing of the selection. If we were to randomly select the projects as they were starting-up, they would automatically move into the organisation's "spot light," thereby, likely resulting in their receipt of special attention. This could, again, result in a biased understanding and representation of our effectiveness. Consequently, it was decided that the sampling frames would be comprised of closing and sufficiently mature projects. This approach, however, has its own particular shortfall – missed opportunities to collect relevant baseline data to support the evaluation processes. While the majority of Oxfam GB's projects are expected to collect baseline data, it is assumed that there will be, in many cases, considerable data gaps, e.g. lack of data from control/comparison populations. This important issue is revisited below.

In sum, our proposal to senior managers on how to operationalise the global indicators, as set out in Oxfam GB's Global Performance Framework, was to randomly select projects from a sampling frame comprised of closing or sufficiently mature projects and use relatively rigorous methods to evaluate the extent to which they have generated change. We have labelled this approach *Effectiveness Auditing*. Our plans, in particular, are as follows: Each financial year, sampling frames of projects earmarked to close or are otherwise sufficiently mature will be developed for each of the six thematic areas presented in Annex 1. Many of the projects will fall under more than one thematic area. When the approach is brought to scale, at least seven projects will be randomly selected under each thematic area. They then will be evaluated during the course of the year, specifically to assess the extent they have generated change in relation to the global outcome indicator in question. Standardised data collection instruments will be used

where appropriate, so that data can be aggregated from different projects. In addition to the organisational learning benefits elaborated on below, it is expected that this approach will, when combined with global output monitoring data, enable communication statements such as:

> "*2011 saw the successful delivery of 51 OGB supported livelihood interventions. These provided vital income generation support to 242,454 smallholder producers (over 156,000 of whom were women) from 36 low-income countries. A random sample of these interventions were independently evaluated and found to have improved consumption and expenditure by an average of 30%."*

4. Horses for Courses: The Large and Small *n* Divide

4.1 Choosing the Right Causal Inference Tool for the Job

While there is resistance in some circles of the NGO community to the call for more RCTs in the international development sector, it is important to understand why the "randomistas" (Ravallion 2009) hold this particular design in such high regard. While more thorough and technical descriptions can be found elsewhere (Shadish, Cook et al. 2002; Duflo, Glennerster et al. 2008), the basics are as follows: We want to know the extent a particular intervention has affected a particular outcome, e.g. household income. If it were possible to know what the status of this outcome would have been in the absence of the intervention, we could compare it (known as the counterfactual outcome) with the observed outcome. The difference between the two would be the intervention's effect.

We can, of course, never really know for certain what would have happened to a particular individual, household, community, etc. had we never intervened. However, the situation is different if the numbers of units we are targeting is large. Specifically, if we were to randomly assign a significant number of units to both intervention and control groups, the statistical distribution of their characteristics – particularly those that affect outcome – will be very much the same. As such, we can use the *observed* outcome of the control group to estimate the *counterfactual* outcome of the intervention group. In the language of the impact evaluation literature, both groups, have the same potential outcomes (Morgan and Winship 2007). Other impact assessment designs that do not randomise intervention exposure are viewed as inferior, given that there is usually no way of being unequivocally certain that the potential outcomes of the intervention and comparison populations are the same.

We can also now clearly see why the RCT design is inappropriate when the number of units being targeted is small, e.g. policy decision-makers in country X. Large numbers of units need to be randomly assigned to intervention and control groups, so that both groups are statistically equivalent. In fact, the more heterogeneous the population, the greater the number required. If we were only targeting a few units, randomly assigning their exposure to a given intervention would be futile from a causal inference point of view; the two groups would, more than likely, simply be too dissimilar to be comparable. Fortunately, the counterfactual or potential outcomes framework is not the only approach to credible causal inference (Brady 2004; Hedström 2008). There are other

approaches that are more appropriate for small *n* interventions, one of which is presented below. The next two subsections, in particular, present the approaches Oxfam GB is pursuing under its effectiveness auditing approach to evaluate both large *n* and small *n* interventions, respectively.

4.2 Mimicking Experiments Non-experimentally

As discussed above, while RCTs may very well be the "gold standard" evaluation design for large *n* interventions, it is unrealistic for most of our organisations to take them on board *as part of regular practice*. Even if we can overcome the ethical and political hurdles associated with randomisation, such designs are expensive and often very challenging to successfully implement. However, over the last several decades, significant developments have taken place in drawing causal inferences from non-experimental or observational data (Imbens and Wooldridge 2009).

There are several different approaches that can be undertaken. What they all share in common is their attempt to "mimic" randomisation. Approaches such as multivariable regression and propensity score matching (PSM), for instance, do this by controlling for observed differences among intervention and comparison groups. Heckman's control function approach attempts to tackle selection bias head-on by directly controlling for the unobserved determinants of outcome that are correlated with participation. The instrumental variable and regression discontinuity designs, on the other hand, exploit the presence of quasi-random factors that affect programme participation. Finally, the difference-in-differences approach uses outcome trends experienced by the comparison group to estimate what the outcome trend of the intervention would have been in the absence of the intervention (Morgan and Winship 2007; Khandker, Koolwal et al. 2010). While there are inherent limitations associated with each of these approaches, there is evidence that non-experimental approaches can reasonably replicate treatment effect estimates generated by experiments, particularly under certain conditions, e.g. controlling for key factors that determine both selection and outcome and using geographically proximate comparators (Cook, Shadish et al. 2008; Handa and Maluccio 2010).

Assessing the impact of large *n* projects under Oxfam GB's project effectiveness auditing approach primarily involves the use of PSM and comparison populations. However, taking advantage of emerging opportunities to complement this with one or more of the other approaches is not ruled out. The general approach is as follows: Upon understanding the nature of the project and its target population/group, efforts are made to identify appropriate comparators. In most situations, the comparators are people/households residing in adjacent communities or sub-districts not reached by the project's interventions. Given that we are interested in estimating the counterfactual, it is important for the comparison populations/groups to be as similar as possible to the intervention populations/groups. Following Handa and Maluccio (2010), this can be best achieved if the comparators reside in the same setting as the project's target group. This is because they are more likely to possess similar characteristics and have been subjected to the same external influences during the project's life span. However, this proximity criterion needs to be balanced with consideration for spill-over effects generated through the neighbouring population's inadvertent exposure to the project's interventions. Understanding who the project's participants are and how they were selected for and/or selected themselves into the intervention group is also critical, so

that relevant differences can be controlled for during statistical analysis. Moreover, efforts are being made to reconstruct baseline data through respondent recall (Bamberger 2009), particularly for information we assume can be reliably remembered, e.g. ownership of particular household assets.

The above approaches were piloted in Somaliland and Tanzania in the context of a disaster risk reduction (DRR) project and value-chain development project, respectively, in the early part of 2011. A summary of the latter project, including the data collection and analysis processes employed and the main effect estimates identified, is presented in Annex 2.

4.3 Searching for Signatures and Smoking Guns

As mentioned above, the potential outcomes framework is not the only approach to causal inference that has traction among scholars; credibly evidencing the mechanisms through which an intervention brought about its effect(s) is also viewed by many as a valid method of causal inference (Brady 2004; Hedström 2008). The best case scenario is when counterfactual and mechanism-based approaches are used together, i.e. where there is both a rigorous estimation of what would have happened in the absence of the intervention, coupled with strong evidence of what mechanisms were at work to bring about the change (Reynolds 1998). Unfortunately, however, as mentioned above, the former approach is not suitable for small *n* interventions. Such interventions, then, must rely primarily on the latter, and this is the impact assessment approach Oxfam GB is pursuing for its policy influencing and citizen engagement interventions.

Process tracing is a qualitative research method that attempts to identify the causal processes – the causal chain and causal mechanism – between a potential cause or causes, e.g. an intervention, and an effect or outcome, e.g. changes in local government practice (George and Bennett 2005). Reilly (2010) elaborates:

> *Process tracing is a robust technique to test theories of causality-in-action by examining the intervening steps….It has been used within the fields of political science, comparative politics, organizational studies, and international relations, in addition to examining cognitive processes underlying decision-making, creativity, and problem solving….It is used to "unwrap" the causal links that connect independent variables and outcomes, by identifying the intervening causal processes, i.e., the causal chain and causal mechanisms linking them.*

In short, process tracing involves evidencing the specific ways a particular cause produced (or contributed to producing) a particular effect.

An important component of process tracing is to consider alternative, competing explanations for the observed outcome in question, until the explanation(s) most supported by the data remains (Patton 2008a). If these alternative explanations have already been identified, "process verification" is directly undertaken. This involves considering, specifying, and documenting what kinds of evidence, if found, would either validate or exclude each of these alternative explanations. However, in many cases, some or all of the possible and plausible explanations for the observed outcome will not have been identified in advance. "Process induction" is, consequently, undertaken. This involves undertaking exploratory, inductive research to identify plausible alternative

explanations, which are then developed into more thorough explanations or hypotheses that can be tested via "process verification," as explained above.

Process tracing therefore works through affirming explanations that are consistent with the facts and rejecting those that are not. This is much like a detective who pursues possible suspects and clues, "…constructing possible chronologies and causal paths both backward from the crime scene and forward from the last known whereabouts of the suspects" (Bennett 2008). However, there is the possibility that the available evidence is not sufficient to verify or eliminate all investigated explanations. Hence, it is possible for the findings of such studies to be inconclusive.

The above approach was piloted under Oxfam GB's Pan African Fair Play campaign and Raising Her Voice project in Indonesia. A summary of the former is presented in Annex 3.

5. Harnessing Potential for Organisational Learning

Ensuring credibility and rigour in the 42 effectiveness audits that will take place each year is going to be challenging in and of itself. However, the other challenge will be to ensure they support us to increasingly improve our effectiveness. The heart of this challenge is getting any resulting learning fed back into decision-making. Incorporating insights from Michael Quinn Patton's Utility Focused Evaluation (UFE) wherever possible will be critically important in this respect (Patton 2008b). This involves things such as identifying and/or cultivating in-country effectiveness champions and making sure staff and partners (including country directors and programme managers) are involved at key points in the process and intimately understand what is taking place. It will further involve opening up space for some of their own evaluative questions to be answered. Making sure that technical evaluation reports are accompanied with more accessible versions (in the local vernacular where relevant) with actionable recommendations is also critical. Attempting to integrate principles from theory based evaluation (Weiss 1998) is particularly relevant for the latter. This is intended to place both ourselves and implementing staff in a better position to understand how the intervention brought about evidenced change or why it failed to do so. Part of this entails unpacking the intervention's theory of change and ensuring that data are captured on key intermediary outcome variables. However, obtaining basic intervention exposure data is also valuable to avoid Type III error, i.e. falsely concluding that a poorly implemented intervention – that would have been effective if properly implemented – is intrinsically ineffective (Dobson and Cook 1980).

If organisational commitment can be maintained, the results of the effectiveness audits should provide us with reasonably credible feedback on the effectiveness of our interventions and support us to increase our overall effectiveness as an organisation. Moreover, looking into the future, we hope to identify particular interventions that appear to be generating significant impact. We then want to drill down on them further through additional research – preferably in partnership with our research team – to better confirm and understand the nature of the impact, as well as how it was brought about and any contextual supporting factors that may have played a role. We also want to do something similar in cases of unexpected and/or unacceptable failure, i.e. for

interventions where we expected impact but where none was detected. Again, the objective, at the end of the day, is to better understand what works and what doesn't and feed this learning into strategic decisions about which interventions to scale up and which to scale back, thereby, enabling the organisation to more effectively fulfil its mission.

6. Are There Other Options?

Carrying out "effectiveness audits," as described above, may not be realistic for many international non-governmental organisations (INGOs) that are also seeking to better understand and demonstrate their effectiveness, and Oxfam GB is certainly not advocating that this approach is a panacea for the INGO sector or even itself. Indeed, there are other options that similarly avoid going down a "development lab" route (Roetman 2011) that are realistic for INGOs to pursue. There are also a number of practical alternatives available for those that are not yet in a position to rigorously assess impact but are, nonetheless, seeking to improve their effectiveness. Several suggestions are presented below that could either be pursued separately or alongside the effectiveness auditing strategy.

- *Partnering with research institutions to rigorously evaluate "strategic" interventions*
  Partnering with universities and their equivalent can, of course, be a sensible approach. Universities, in the UK at least, are under pressure to demonstrate their "real world" impact and are, consequently, increasingly interested in pursuing such partnerships. Moreover, there will likely be increasing funding available for this, particularly if we can work with such institutions to develop technically strong impact evaluation proposals. However, we should be aware that undertaking such research is no "walk in the park," as illustrated by the HIV/AIDS Alliance's experiences with an RCT in Andhra Pradesh, India (Samuels and McPherson 2010). Even with the involvement of external researchers, such designs inevitably take up staff time and may necessitate significantly modifying the intervention's implementation strategy. Contamination of the control group is also a serious risk, necessitating a high degree of co-ordination with and support from key stakeholders. Given the level of effort and resources required to ensure success, we may want to reserve this precious strategy for those innovations that are under-researched, have significant scale-up potential, and are compatible with our respective missions and organisational objectives. We should avoid getting ourselves involved in such impact evaluations simply because funding is available.

- *Making modest investments in additional technical capacity and unleashing it*
  With the ever increasing pressure to demonstrate results, we must become more effective as a sector in both successfully adapting to and shaping this new reality. However, we cannot be successful at either unless we make more serious efforts to scale-up our technical capacity. This amounts to more than simply employing more Monitoring and Evaluation (M&E) officers and the like; we need to take a serious look at the structures of our organisations and create space for the recruitment and retention of high-calibre impact evaluation and research specialists. Currently, impact evaluation enthusiasts emphasise undertaking more RCTs, particularly because of their high internal validity. Technical evaluation specialists could support us to effectively implement hybrid RCT designs for large *n* interventions, even without

necessarily having to partner with research institutions, e.g. pipeline or randomised encouragement designs (Duflo, Glennerster et al. 2008). This would only require relatively minor tweaks to our modi operandi. Such specialists would also be essential in the credible pursuit of conventional quasi-experimental designs where randomisation is not feasible, as well as other alternative approaches such as dose response analysis (Plautz and Meekers 2007) and pattern matching (Trochim 1989).

Moreover, much of our focus and, arguably, comparative advantage, lies in undertaking small *n* interventions, e.g. those related to organisational capacity building and advocacy. As argued above, rigorously evaluating such interventions requires a different approach to causal inference, a role most effectively filled by appropriately trained qualitative researchers. Having such staff within our organisations to shape and manage such evaluations would, therefore, be clearly advantageous.

- *Pursuing more evidence informed programming*
  More often than not, we spend insufficient time designing and developing our programmes and/or supporting our partners to do the same. We may consult with communities and other stakeholders and even invest significant time in fleshing out the intervention's logic and/or developing its logframe with programme staff and partners. However, how often do we look at the existing literature on the effectiveness of interventions that are similar to those we plan to implement or even advocate for in our policy work? We should be mindful, however, that the available evaluative evidence is hardly ever in an accessible format or even accessible to our organisations for that matter, and we should throw this challenge back to the impact assessment and academic communities.

  Nevertheless, as communicated at the *Mind the Gap* conference (Cuernavaca, Mexico, June, 2011), significant effort has gone into bridging the evaluation gap in recent years, with the number of rigorous impact evaluations in international development sector now in the hundreds. Systematic reviews are also becoming more popular. Such reviews seek to identify, review, and synthesise all high quality studies on a particular research question, e.g. the effectiveness of a particular intervention. Not all of the standard community-level NGO-type interventions have been systematically reviewed to date, nor do those that have been reviewed necessarily provide conclusive answers on what works and for whom and under what conditions. However, such reviews do exist on several of our sector's core interventions, e.g. water and environmental sanitation (Waddington and Snilstveit 2009), HIV behaviour change (Noar, Palmgreen et al. 2009), and micro-finance (Stewart, van Rooyen et al. 2010). And we should be using their findings and those of other rigorous evaluations, where relevant, to inform our programming. Simply doing so would lighten the evaluative burden for such interventions, given that the effectiveness question has already been answered in large measure, thereby, allowing us to concentrate our main efforts on ensuring effective implementation.

- *Making better use of existing evaluation resources*
  We could do more to ensure that the M&E resources we do have deliver better "value for money." At the moment, much of our evaluative efforts are spent on evaluation designs that focus on establishing whether intended outcomes have materialised, rather than on assessing the contribution of our interventions to such changes.

Indeed, the capture of quality baseline and endline data on "objectively verifiable indicators" on only the intervention group has been widely touted in our sector as good practice. We have come to believe that targeted change in outcome indicator *Y* from ?₂ to ?₂ or ?₂ means that our programme was effective, despite the fact that a myriad of other factors may have been responsible for the change. Karlan and Appel (2011), in particular, "…consider it unethical to measure impact so badly…." Part of the problem lies in the successful institutionalisation of results based management (RBM) in general and the infamous logframe in particular, which focuses enquiry on the tracking of outcome indicators, rather than critically testing the assumptions embedded within the intervention's theory of change.

How then could we make better use of the resources we currently spend on evaluation? One would be to direct resources towards strengthening basic implementation monitoring. As illustrated by Williams (2007) in the context of a reading improvement programme in Malawi, many interventions are ineffective simply due to implementation deficiencies or lack of intervention uptake, rather than intrinsic design shortfalls. Being more strategic about what monitoring data are collected and how they are used to inform management decision-making and support programme strengthening processes in "real time," following the approach of developmental evaluation (Patton 2011), would surely make many of our organisations more effective – *even where we are unable to rigorously answer the effectiveness question*. This is particularly relevant for complex interventions, where strategies and interim outcomes are emergent and effectiveness relies heavily on the programme's ability to respond flexibly to unforeseen opportunities.

Another pragmatic step would be to invest more in the recruitment and management of evaluation consultants. Spending time and effort in developing clear, realistic terms of reference (ToRs) is critical. Often, the questions posed to evaluators, e.g. the extent the project reduced poverty or empowered women, are not only too many and too vague but also typically impossible to credibly answer, particularly given prevailing budget, time, and/or data constraints. Furthermore, a lack of technical oversight and management often leads to a poor application of methodological rigour. Nevertheless, the evaluators' final findings are usually accepted, by both ourselves and our donors, with minimal scrutiny (Nelson 2008). These issues call for the technical involvement of capable evaluation staff from our organisations in overseeing external evaluations, and we should not be naive as to how much time and effort this entails.

A final radical suggestion to make better use of our evaluation resources is to do away with collecting baseline data on large *n* interventions altogether, particularly when the key barrier to collecting baseline data on a control/comparison population is financial. Rather, it would be more credible from an impact evaluation perspective to use these same resources earmarked for the project's baseline survey to collect data on a comparison population ex-post. And if implementing staff and/or partners can be convinced to implement the intervention in randomly chosen locations among a pool of potential locations, then all the better. In fact, two of the main reasons why baseline data are collected in RCTs is to both see if randomisation was successful, i.e. whether the observable characteristics of the intervention and control groups are balanced, and to improve precision and estimation during statistical analysis. If one can assume that the randomisation process was successful, the collection of baseline

data is actually not entirely necessary; the intervention and control groups can simply be directly compared ex-post. Now if the random selection of communities is not possible, we can still collect data on both intervention and comparison groups and subsequently statistically adjust for any differences between them. While this approach is not the gold standard, it is certainly more rigorous than our usual before and after comparisons.

7. Concluding Thoughts

There is no doubt that the hearts of many NGOs are in the right place; most of us truly want to change the world, even at the cost of eventually working ourselves out of jobs. However, to quote the title of Karlan and Appeal's (2011) recent book, we need more than good intentions. If we really want to make a significant contribution to reducing global poverty, oppression, inequality, injustice, environmental degradation, etc., we must not only continue to work hard but also smarter. This means moving away from interventions that don't make much of a difference to those that do. Part of this entails being more strategic about what we get ourselves involved with in the first place, requiring more investment in intervention design and even turning down donor funding when conditions are unfavourable. However, we are only going to be effective if the interventions we implement are themselves intrinsically effective. This brings us back to the challenge highlighted in the first section of this paper: Accessing credible intervention effectiveness feedback is no easy task, and most of our organisations are not set up as "development labs."

This paper has suggested several possible strategies that we, as a sector, can pursue to practically, yet meaningfully, step up our game on the impact evaluation front. Partnering with research institutions to rigorously evaluate selective interventions is one option, doing our best to avoid pitfalls such as those experienced by the HIV/AIDS Alliance in Andhra Pradesh. Particularly for more well researched areas, we can also be more effective if we pursue interventions whose effectiveness has already been demonstrated and, conversely, stay away from those whose effectiveness is suspect. Making better use of our existing evaluation resources can further make a difference. In particular, we must get it out of our heads that outcome indicator tracking is a credible strategy for understanding and demonstrating our impact. Doing fewer but more rigorous quantitative, qualitative, and mixed-method evaluations and even using such resources to simply strengthen implementation is sure to deliver better "value-for-money." Finally, we must invest more in strengthening the technical capacities of our organisations in impact evaluation. Such capacity is necessary to not only effectively support programme staff to pursue innovative, yet credible, impact evaluation designs and effectively identify and manage consultants but also ensure we have a more substantive say as a sector in shaping the current debate, particularly to avoid having things imposed that are unworkable and/or do not add value.

ANNEX 1: Oxfam GB's global outcome indicators

Thematic Area	Outcome Indicator
Humanitarian Support	• % of people who received humanitarian support from responses meeting established standards for excellence, disaggregated by sex
Disaster Risk Reduction/Climate Change Adaptation	• % of targeted households indicating positive ability to minimise risk from shocks and adapt to emerging trends & uncertainty
Livelihoods Support	• % of targeted households living on more than £1.00 per day per capita
Women's empowerment	• % of supported women meaningfully involved in household decision-making and influencing affairs at the community level
Popular Mobilisation (Citizen's Voice)	• % of targeted state institutions and other actors that have modified their practices in response to engagement with supported citizens , community based organisations/civil society organisations
Policy Influencing	• % of policy objectives/outcomes successfully achieved, disaggregated by thematic area

It is worth acknowledging here that the humanitarian support indicator is technically not an *outcome* indicator, as it is focused on adherence to quality standards. While the aim of providing humanitarian support is arguably to reduce morbidity, mortality, and other forms of suffering, estimating the extent that Oxfam GB supported responses have done or even contributed to this would be considerably challenging, given the inherent limitations of identifying suitable counterfactuals. This does not mean that possibilities do not exist, e.g. exploiting "natural experiments" where people are not supported for quasi-random reasons or regression discontinuity designs where people who just fall within or outside of official targeting criteria are compared (Angrist and Pischke 2009). However, taking a critical look at the extent to which targeted populations are provided with support that meets recognised standards, e.g. Sphere guidelines, was considered good enough to serve as a pseudo outcome indicator for this thematic area of work.

A few additional points on the other indicators are worth mentioning. First, one may question how reliable data can be accessed on the second indicator pertaining to disaster risk reduction and climate change adaptation. There is, however, an approach that underlies the apparent madness. In particular, following John Twigg (2009), we hypothesise that households possess particular context specific characteristics – e.g. the degree of reliance on climate dependent livelihood activities and access to climate prediction information – that influence their vulnerability to hazards and/or ability to adapt to climate change. The approach scores household's in relation to these characteristics. The fourth indicator, on women's empowerment, may also appear confounded by intrinsic measurement challenges. The associated instrument involves asking women questions pertaining to both the breadth and depth of their involvement in household decision-making. Several Likert scales are further employed to measure their perceived ability to influence affairs outside the home. Finally, perhaps to the horror of some, the final two indicators involve the quantification of qualitative information. In particular, external evaluators will be asked to assign "contribution scores," the value of which will depend on the extent there is evidence that links the popular mobilisation and policy influencing interventions in question to any expected and/or unexpected policy-related outcomes.

ANNEX 2: Effectiveness audit pilot summary – Tanzania agricultural scale-up

THE PROGAMME:
Oxfam is working with local partners in four districts of Shinyanga Region, Tanzania, to support over 4,000 small-holder farmers (54% of whom are women) to enhance their production and marketing of local chicken and rice. To promote group cohesion and solidarity, the producers are encouraged to form themselves into savings and internal lending communities. They are also provided with specialised training and marketing supporting, including forming linkages with buyers through the establishment of collection centres.

THE INDICATORS:
Two of Oxfam GB's global outcome indicators that are part of its Global Performance Framework were piloted under this programme. These include:

- **% of targeted households living on more than £1.00 per day per capita**
  The tool that captures data on this indicator is an adapted version of the World Bank's Living Standards Measurement Survey instrument. Household representatives are asked to provide information on the amount they have consumed and spent in relation to food and non-food items. The assumption is that more wealthy households consume and spend more than those who are poorer.
- **% of supported women are meaningfully involved in household decision-making and are able to influence affairs in their communities**
  Obtaining data on this indicator involves asking women about the breadth and depth of their involvement in 24 household decision-making areas and the extent they believe they are able to influence decisions and governance processes in their communities.

THE PROCESS:
Supported by an external consultant, a household survey (n=457) and a women's questionnaire (n=446) were administered to randomly selected chicken and rice producers in 86 intervention and matched comparison villages in the four targeted districts. These instruments not only included questions relevant to the above indicators but also important information on household characteristics and other outcome variables discussed further below. In order to compare like with like, statistical analysis was undertaken using propensity score matching (PSM) with exact matching by product type to control for observable differences.

THE RESULTS:

- Overall, there was no statistically significant difference in the proportion of households living above £1.00 per day per capita. However, when producers are examined by product type, a different picture is revealed. Households supported through the chicken value chain intervention are significantly better off, with the proportion being 63% versus 48% for the intervention and comparison groups, respectively (p-value < 0.05). However, this did not translate into a notable difference in their food security. It should also be noted that both a serious chicken disease and drought have hit the Shinyanga Region, which have negatively affected both groups of producers.
- Interestingly, women in the rice producing groups were assessed as having greater household decision-making power: 37% scored positively, against 17% in the

comparison group (p-value < 0.05). The overall difference for both rice and chicken producers was not statistically significant. However, women in both groups scored better than their comparators in relation to the international self-efficacy scale (p-value < 0.01) and were found to be more likely to own productive assets – 45% versus 32% (p-value < 0.05).

ANNEX 3: Effectiveness Audit Pilot Summary – Policy Influence in Africa

THE PROJECT:

The *Fair Play* campaign works to amplify the voices of African citizens to demand their right to universal access to health and HIV/AIDS services. At a national level, the campaign acts as a facilitator, bringing together over 200 civil society partners to work in coalition on existing national campaigns united under the *Fair Play* goals, promoting cohesion and clarity of purpose amongst these actors. *Fair Play* also works to directly engage with policy- and decision-makers, including governments and parliaments at the regional level.

THE INDICATOR:

Drawing on theory-based evaluation approaches, Oxfam has defined a robust qualitative research protocol, 'Process Tracing', to enable assessment of a) the extent to which intended policy objectives, or interim outcomes that signal progress towards these objectives were successfully achieved, and b) the extent to which the intervention contributed to these changes. c) constructing the campaign's theory of change with key stakeholders, the approach identifies the interim and final outcomes the campaign sought to achieve. The evaluator then seeks evidence for the extent to which these outcomes have materialised; identifies plausible causal explanations for those outcomes (including but not limited to the campaign itself); and assesses the extent to which each of the explanations are, or are not, supported by the available evidence.

THE PROCESS:

Using the above methodology, the evaluator identified and assessed four key outcome areas:

1. African Union Member States meet and aim to exceed the Abuja Commitment to allocate 15 per cent of their national budgets to health (**Regional**)
2. *Fair Play* country governments take accelerated action by investing in health as per the campaign 'asks' (**National**)
3. Strongly linked civil society organisations work to improve health for all Africans (**Civil Society**)
4. Africans have a strong collective voice which they use to demand their right to health (**Community**)

THE RESULTS (by outcome):

1. **Regional:** There has been little significant change in the direction or pace of investment in health towards the Abuja Declaration target. Nevertheless, *Fair* Play has contributed to some clear successes on interim outcomes. For example, the evaluation found it reasonable to conclude that the decision by African Union Finance Ministers to hold to health budgetary commitments made in the Abuja Declaration had a causal relationship with *Fair Play*.
2. **National:** *Fair Play* has contributed to raising the priority among African Union Member States endorsing a health-MDG accelerated action plan. However, there is little evidence of country governments taking accelerated action to invest in health, and, in a crowded policy space, it remains inconclusive as to whether actions by *Fair Play* have contributed to those changes that have materialised.
3. **Civil Society:** There is good evidence to suggest that *Fair Play* was successful in increasing civil society organisation knowledge and advocacy capacity in relation to global health issues, and linking these organisations under the campaign's aim of *'health for all'* through creative campaign tactics, clear campaign messages and popular branding.
4. **Community:** While there is evidence that *Fair Play* activities have supported African citizens to increase their knowledge on health issues and have brought some recognition that communities have a role to play, this has been on a very small scale and there is insufficient evidence to suggest that *Fair Play for Africa* has empowered African citizens to demand their rights to health. Key informants agreed that this has been the weakest element of the campaign overall, though the evaluation noted that it will take time for the campaign to generate impact at the citizen level.

References

Angrist, J. and Pischke, J., 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Bamberger, M., 2009. Strengthening the evaluation of programme effectiveness through reconstructing baseline data. *The Journal of Development Effectiveness*, 1 (1), 37-59.

Bamberger, M., Rugh, J., et al., 2006. *RealWorld Evaluation: Working Under Budget, Time, Data, and Political Constraints*. Thousand Oaks: Sage Publications.

Bennett, A., 2008. Process Tracing: A Bayesian Perspective. *In*: J. M. Box-Steffensmeier, H. E. Brady and D. Collier, eds. *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press**,** 702-720.

Brady, H., 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield Publishers, Inc.

Chen, H., 2005. *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness*. Thousand Oaks: Sage Publications.

Cook, T. D., Shadish, W. R., et al., 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27 (4), 724-750.

Cracknell, B. E., 2000. *Evaluating Development Aid: Issues, Problems and Solutions*. New Delhi: Sage Publications.

Cronbach, L. J., 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Dobson, D. and Cook, T. J., 1980. Avoiding type III error in program evaluation : Results from a field experiment. *Evaluation and Program Planning*, 3 (4), 269-276.

Duflo, E., Glennerster, R., et al., 2008. Using Randomization in Development Economics Research: A Toolkit. *In*: T. P. Schultz and J. Strauss: *Handbook of development economics*. Amsterdam: Elsevier, 4, 3895-3962.

George, A. and Bennett, A., 2005. *Case Studies and Theory Development in the Social Sciences.* Cambridge: MIT Press.

Greene, J. C., Benjamin, L., et al., 2001. The Merits of Mixing Methods in Evaluation. *Evaluation* 7(1), 25-44.

Greenhalgh, T., 1997. How to read a paper: Papers that summarise other papers (systematic reviews and meta-analyses). *BMJ*, 315 (7109), 672-675.

Guba, E. G. and Lincoln, Y. S., 1989. *Fourth Generation Evaluation.* Newbury Park: Sage Publications.

Handa, S. and Maluccio, John A., 2010. Matching the Gold Standard: Comparing Experimental and Nonexperimental Evaluation Techniques for a Geographically Targeted Program. *Economic Development and Cultural Change*, 58 (3), 415-447.

Hedström, P., 2008. Studying Mechanisms to Strengthen Causal Inferences In Quantitative Research. *In*: J. M. Box-Steffensmeier, H. E. Brady and D. Collier, eds. *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Higgins, J., Green, S., et al., 2005. Cochrane Handbook for Systematic Reviews of Interventions . *The Cochrane Library* (3).

Imbens, G. W. and Wooldridge, J. M., 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47 (1), 5-86.

Karlan, D. and Appel, J., 2011. *More than Good Intentions: How a New Economics is Helping to Solve Global Poverty*. New York: Dutton.

Khandker, S., Koolwal, G., et al., 2010. *Handbook on Impact Evaluation*: The World Bank.

Lay, M. and Papadopoulos, I., 2007. An Exploration of Fourth Generation Evaluation in Practice. *Evaluation*, 13 (4), 495-504.

Morgan, S. and Winship, C., 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press.

Nelson, J. L., 2008. Are we ready for RCTs? *Monday Developments*, InterAction.

Newman, J., Rawlings, L., et al., 1994. Using Randomized Control Design in Evaluating Social Sector Programs in Developing Countries. *The World Bank Research Observer*, 9 (2), 181-201.

Noar, S. M., Palmgreen, P., et al., 2009. A 10-Year Systematic Review of HIV/AIDS Mass Communication Campaigns: Have We Made Progress? *Journal of Health Communication*, 14 (1), 15-42.

Patton, M., 2008a. Advocacy Impact Evaluation. *Journal of MultiDisciplinary Evaluation*, 5 (9), 1-10.

Patton, M., 2008b. *Utilization Focused Evaluation, Fourth Edition*. Los Angeles: Sage.

Patton, M., 2011. *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: The Guilford Press.

Picciotto, R., 2003. International Trends and Development Evaluation: The Need for Ideas. *American Journal of Evaluation*, 24, 227-234.

Plautz, A. and Meekers, D., 2007. Evaluation of the reach and impact of the 100% Jeune youth social marketing program in Cameroon: findings from three cross-sectional surveys. *Reproductive Health*, 4 (1), 1-15.

Ravallion, M., 2009. Should the Randomistas Rule? *The Economists' Voice*, 6 (2).

Reilly, R. C., 2010. Process Tracing. *In*: A. J. Mills, G. Durepos and E. Wiebe, ed. *Encyclopedia of Case Study Research*. Thousand Oaks, Sage Publications, Inc.

Reynolds, A. J., 1998. Confirmatory program evaluation: A method for strengthening causal inference. *American Journal of Evaluation*, 19 (2), 203-221.

Roetman, E., 2011. A can of worms? Implications of rigorous impact evaluations for development agencies. *3ie Working Paper (11)*.

Rossi, P. H., Lipsey, M. W., et al., 2004. *Evaluation: A Systematic Approach (Seventh Edition)*. Thousand Oaks: Sage Publications.

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

Shadish, W. R., Cook, T. D., et al., 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Shroff, R. and Stevenson, J., 2008. *Indicator Feasibility Study: Final Report for International Division Senior Management Team*. Oxford: Oxfam GB.

Stewart, R., van Rooyen, C., et al., 2010. *What is the impact of microfinance on poor people? A systematic review of evidence from sub-Saharan Africa*. London: EPPI-Centre, Social Science Research Unit, University of London.

Trochim, W. M. K., 1989. Outcome pattern matching and program theory. *Evaluation and Program Planning*, 12 (4), 355-366.

Twigg, J., 2009. Characteristics of a Disaster Resilient Community: A Guidance Note Version 2. London.

Valadez, J. and Bamberger, M., 1994. *Monitoring and Evaluating Social Programs in Developing Countries: A Handbook for Policymakers, Managers, and Researchers*. Washington, D.C.: The World Bank.

Veerman, J. W. and van Yperen, T. A., 2007. Degrees of freedom and degrees of certainty: A developmental model for the establishment of evidence-based youth care. *Evaluation and Program Planning*, 30 (2), 212-221.

Voils, C. I., Sandelowski, M., et al., 2008. Making Sense of Qualitative and Quantitative Findings in Mixed Research Synthesis Studies. *Field Methods*, 20 (1), 3-25.

Waddington, H. and Snilstveit, B., 2009. Effectiveness and sustainability of water, sanitation, and hygiene interventions in combating diarrhoea. *Journal of Development Effectiveness*, 1 (3), 295-335.

Weiss, C. H., 1998. *Evaluation: Second Edition*. Upper Saddle River: Prentice Hall.

White, H., 2005. The road to nowhere? Results-based management in international cooperation. *In*: S. Cummings, ed. *Why Did the Chicken Cross the Road? And other stories on development evaluation...* Amsterdam: KIT Publishers, 71-76.

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

White, H., 2010. A Contribution to Current Debates in Impact Evaluation. *Evaluation*, 16 (2), 153-164.

Williams, E., 2007. Extensive reading in Malawi: inadequate implementation or inappropriate innovation? *Journal of Research in Reading*, 30 (1), 59-79.

Wynn, B. O., Dutta, A., et al., 2005. *Challenges in Program Evaluation of Health Interventions in Developing Countries*. Santa Monica: Rand Corporation.

Zhu, S., 1999. A Method to Obtain a Randomized Control Group Where It Seems Impossible. *Evaluation Review*, 23 (4), 363-377.

Other publications in the 3ie Working Paper series

Sound expectations: from impact evaluations to policy change by Vanessa Weyrauch and Gala Díaz Langou, 3ie Working Paper 12, April 2011

A can of worms? Implications of rigorous impact evaluations for development agencies by Eric Roetman, 3ie Working Paper 11, March 2011

Conducting influential impact evaluations in China: the experience of the Rural Education Action Project by Mathew Boswell, Scott Rozelle, Linxiu Zhang, Chengfang Liu, Renfu Luo, Yaojiang Shi, 3ie Working Paper 10, February 2011

An introduction to the use of randomized control trials to evaluate development interventions by Howard White, 3ie Working Paper 9, February 2011

Institutionalisation of government evaluation: balancing trade-Offs by Marie Gaarder and Bertha Briceno, 3ie Working Paper 8, July 2010

Impact Evaluation and interventions to address climate change: a scoping study by Martin Prowse and Birte Snilstveit, 3ie Working Paper 7, March 2010

A checklist for the reporting of randomized control trials of social and economic policy interventions in developing countries by Ron Bose, 3ie working paper 6, January 2010

Impact evaluation in the post-disaster setting by Alison Buttenheim, 3ie Working Paper 5, December 2009

Designing impact evaluations: different perspectives, contributions from Robert Chambers, Dean Karlan, Martin Ravallion, and Patricia Rogers, 3ie Working Paper 4, July 2009. Also available in Spanish, French and Chinese

Theory-based impact evaluation by Howard White, 3ie Working Paper 3, June 2009. Also available in French and Chinese.

Better evidence for a better world edited by Mark W. Lipsey University and Eamonn Noonan, 3ie & The Campbell Collaboration, 3ie Working Paper 2, April 2009

Some reflections on current debates in impact evaluation by Howard White, 3ie Working Paper 1, April 2009