Aid evaluation: Pursuing development as if evidence matters

Jan Willem Gunning*

Summary

■ Debates on what works in development and what does not are rarely evidence based but donors are increasingly interested in establishing rigorously whether the aid they provide to developing countries is effective. While much of the aid effectiveness literature uses a macro approach (cross country regressions) this paper proposes a bottom up approach whereby the impacts of general budget support or aid-supported sector programmes are assessed on the basis of statistical impact evaluation of the interventions affecting a representative sample of intended beneficiaries. This bridges the gap between the existing methods for evaluating impact (which are designed for projects) and the growing demand for impact evaluation of sector aid or general budget support. ■

JEL classification: O10, O22, I38.
Key words: Impact evaluation; budget support; heterogeneous treatment.

* *Jan Willem Gunning is Professor at the Free University, Amsterdam.*

Aid evaluation: Pursuing development as if evidence matters

Jan Willem Gunning*

There are striking similarities between the current public debate on development aid and the way medical interventions were discussed in the nineteenth century. For example, the question of how to deal with cholera, still one of the big killer diseases in Western Europe 150 years ago, was discussed in heated exchanges between participants who sometimes had no medical expertise whatsoever, but who invariably held strong opinions about the transmission of the disease, a topic on which science was still divided and on which there was as yet very little empirical evidence.

That debate was famously resolved by John Snow (1813-1858), a London physician. In 1854 he carefully plotted reported cholera cases in Soho on a street map, noted that differences between locations in cholera incidence were highly correlated with differences in the sources used for drinking water and correctly concluded that the disease was transmitted through the use of polluted drinking water. Snow then made his case in the most spectacular and convincing way imaginable: he had the handle of the water pump in a high incidence area, Broad Street, removed, thereby forcing the users to switch to a more distant but clean water source. Cholera incidence fell rapidly in the area, dramatically illustrating that Snow's hunch was correct. This evidence led to ambitious urban piped water and sewerage systems in many countries in the next few decades. By the end of the 19th century, cholera was no longer a serious public health menace in Europe.[1]

* *I am grateful to my discussant, Per-Åke Andersson, to an anonymous referee and to Arne Bigsten, Chris Elbers, Wendy Janssens, Henri Jorritsma, Antonie de Kemp, Stefan Molund, Jakob Svensson, Finn Tarp and Rita Tesselaar for many helpful discussions on this topic. In this paper, I draw on joint work with Arne Bigsten and Finn Tarp (Bigsten et al., 2006) and on ongoing work with Chris Elbers on evaluations of education in Zambia and water and sanitation activities in Tanzania.*

[1] *Dictionary of National Biography* and Porter (1997), pp. 412-4.

Snow's removal of the pump handle was one of the greatest triumphs of evidence-based medicine. Such successes of empiricism led to institutional changes: official recognition moved rigorous testing from the fringe to the centre of medical innovations. In the 20th century, the medical profession has been very successful in gaining wide acceptance for the position that drugs testing and public health interventions should be based on evidence. The principle of double blind testing with random assignment is now no longer seriously challenged.

This is a remarkable success for a field in which the value of professionalism has come to be recognized only very slowly and which continues to exert a magnetic attraction on quacks.

The analogies between the two fields are obvious. However, development is clearly still far behind medicine: it is much less evidence-based, professionals do not obviously enjoy more credibility than self-proclaimed experts such as rock star Bob Geldof and there is as yet at best only a lukewarm acceptance of rigorous testing. In that sense, modern day debates on development policy are much like public health discussions in the days before John Snow's famous pump experiment.

That it is feasible to rigorously test interventions in development, indeed very much like medical drugs, has been convincingly and eloquently argued by many authors, particularly in development economics. An excellent (and very entertaining) introduction in this field is Ravallion (2001) and a recent overview is given by Duflo (2005). I discuss this approach, i.e. statistical impact evaluation, in Section 1.

In recent years donors have moved away from financing projects to sector aid and general budget support. In Section 2, I discuss the implications of this change for evaluation. Ironically, just when many donor agencies are becoming interested in statistical impact evaluation techniques (designed for narrowly defined projects), these methods are becoming less relevant for them as they move away from project finance. The key issue is whether evaluation should take place at the sectoral or national level or, alternatively, whether assessments at that level should be based on evaluations at a lower level of aggregation. I argue in favour of the latter, "bottom up" approach.

In Section 3, I show that the existing evaluation techniques can be modified to make them suitable for evaluating sector aid or budget support. Some experimental work is already under way in these areas and the paper outlines possible approaches. Section 4 considers externalities and interaction effects. Section 5 concludes.

1. Statistical impact evaluation[2]

The desirability of an evidence-based approach in development would seem self-evident. In fact, it is often considered as controversial or impractical.

In the Netherlands a high-level committee, chaired by a former deputy prime minister, recently advised the Dutch minister for development cooperation on criteria for judging Dutch development NGOs. The committee considered the use of impact evaluation to assess the effectiveness of these organizations and noted four points:[3]

- If impact variables are affected by many factors, it is difficult to establish what part of any change measured in these variables can be attributed to the project or programme one wishes to evaluate;
- One cannot aggregate over heterogeneous impact variables;
- There may be long lags between interventions and their ultimate impact;
- Data quality may be poor.

Clearly, these are valid points. The committee concluded (in a remarkable lapse of logic) that because of these "methodological problems", accountability in development cannot be based on impact evaluation. This is, of course, a *non sequitur*, but the example illustrates a reluctance to adopt rigorous evaluation which is very common in development. For example, while some parts of the World Bank (notably the research department) have produced excellent impact evaluations, the Bank's official evaluation arm, the OED, hardly makes any use of such techniques. Similarly, bilateral donor agencies are heavily engaged in evaluation activities but these typically focus on processes rather than impact and they do not rigorously test their attributions of results to interventions. However, this is changing: the debate on aid effectiveness has caused a surge of interest in better evidence and hence, in formal impact evaluation techniques.

It should be noted that the term impact evaluation is used in two different senses, denoting either the methodology used or the result

[2] This section draws on Bigsten et al. (2006).

[3] Dijkstal (2006), p. 17.

being evaluated. In the latter meaning, impact evaluation tries to assess to what extent an activity has had “impact”, i.e. whether it has succeeded in reaching ultimate development objectives such as reduced poverty, malnutrition or infant mortality. In this usage, impact is contrasted with inputs and intermediate results of donor-supported activities (in the jargon: outputs and outcomes). Many donor agencies have recently started to move away from their traditional focus on these intermediate results and are investigating to what extent they can evaluate interventions in terms of their ultimate impact. This new focus is at least in part a response to political pressure to establish “aid effectiveness” in a more convincing way.

Impact evaluation can also denote a formal (statistical) comparison of observed results with results for a counterfactual; the difference between the two is then attributed to the intervention. (In this case, there is no presumption that the results being evaluated are final (“impact”) rather than intermediate. For example, the analysis could focus on school enrolment, an intermediate result, rather than on literacy.) I will use the term statistical impact evaluation for this case.

Ideally, one would compare results for the same group with and without “treatment”. However, obviously, no group can be observed in both situations at the same time. This is the fundamental *evaluation problem*. It forces the evaluator to construct a control group in such a way that the results for this group can be used as the results for the hypothetical case when the “treatment group” would in fact have received no treatment. Rather than comparing the same group with and without treatment at the same time (which is desirable but impossible), he will then compare results for two different groups.[4]

The simplest application of this idea is the randomisation which is at the heart of experimental designs.[5] For example, in testing medical drugs participants are randomly assigned to treatment and control groups. Random assignment implies that there is no reason to suppose that there are any (statistically significant) differences between the two groups prior to the experiment. The control group therefore offers an appropriate basis for comparison: if any significant differences in results between the two groups are found, then these can confidently be attributed to the medicine.

[4] The hypothetical nature of the counterfactual is sometimes used as an argument against statistical impact evaluation. This is not well taken: the objection simply ignores the evaluation problem.

[5] See e.g., Duflo (2005).

Quasi-experimental methods have a long history in policy evaluation. For example, while traditional evaluations of employment policies relied heavily on before/after comparisons (did a group of unemployed workers succeed in finding jobs after participating in a training programme?), such comparisons clearly suffer from a selection effect. If candidates self-selected themselves into the programme, then their finding jobs need not reflect the impact of the training: those who signed up for the programme might have characteristics that made them more likely than others to find jobs in the absence of the programmes. Clearly, a traditional (before/after) evaluation would then be meaningless. Labour market research established a strong tradition of rigorous statistical impact evaluation to construct convincing counterfactuals for such cases.

In development, the use of such evaluation methods is more recent, but the last decade has seen numerous applications in evaluations of social safety nets, schooling programmes targeted at the poor, health interventions and even rural empowerment programmes. As in the case of labour market evaluations, work in this area has moved from its initial research focus to practical applications. Both NGOs and bilateral and multilateral donor agencies are beginning to experiment with such methods. (One of the most famous papers in this field, Miguel and Kremer (2004), describes an evaluation of primary schooling activities in Kenya and this was initiated by a small NGO.)

Where implementation of an intervention is gradual (e.g., 25 per cent coverage of the villages concerned in the first year, 50 per cent in the second year and so on) there is a strong case for using random assignment of villages to the various rounds of implementation and some policy makers are beginning to realize this.[6] In the absence of randomization, a central issue in such evaluations is the availability of baseline data. With baseline data, one can address the fundamental problem of unobserved differences between the treatment and control groups. (Under randomization, the problem does not arise: any such differences are then non-systematic.) Rather than measuring differences at time *t* (after "treatment") between the two groups, one can measure changes over time for both groups. Impact can then be assessed as the difference between the two groups in those changes

[6] Since the implementation of the intervention is gradual in any case, the usual moral objection to randomisation does not apply. If one is not going to instantaneously extend the treatment to the entire target group anyway, then random assignment of the initial beneficiaries will seem to be equitable.

over time ("differences in differences" or "double differencing"). In Section 3, I discuss how this method can be extended to a more realistic, multi-period setting.

While policy makers are understandably reluctant to invest in the collection of baseline data, there is a growing awareness that without either randomization or baseline data it is quite difficult to assess the results of an intervention. There is an alternative (discussed in Section 3) but instead of baseline data, it requires information on interventions at various points in time. Those data are often not recorded in a suitable form so that an evaluation is possible only if one engages in *ex post* data collection.

Statistical impact evaluation presupposes that both the treatment and its possible effects are well defined. For example, the treatment might be a project offering cash transfers to poor households conditional on the (continued) school enrolment of their children.[7] Given the project's objective, its impact is then obviously to be measured in terms of enrolment of children in the target group. Many development interventions fall into this category of specific activities with obvious success indicators. If donors support such activities, then they can use statistical impact evaluation. (But, of course, there may be fungibility: the project evaluated may not be what the donor in fact financed.)

However, donors are increasingly moving from project aid to sector support or general budget support. This shifts the evaluation question to a much higher level of aggregation, a level for which the techniques of statistical impact evaluation have not been designed.

2. From project finance to budget support

One approach is to measure the impact of aid through cross-country growth regressions. Inter-country variance is then used to estimate the impact (in terms of changes in poverty, income or economic growth) of total aid (or its various components). Implicitly, the experience of other countries is then used to construct a counterfactual whereby one controls as much as possible for inter-country differences other than those in aid receipts. This is an active (and somewhat controversial) area of research.[8] The results are far from settled and

[7] An example of such an evaluation is discussed at length in Ravallion (2001).

[8] The father of growth theory, Robert Solow, provides a thoughtful critique of growth regressions in Solow (2002). He is critical of the assumption that the same

much of the work in this area fails to pass tests of robustness.[9] In addition to econometric weaknesses, this approach has the disadvantage of generating no information on the relative effectiveness of the various aid-supported activities, information which both donors and recipient governments hope to obtain from an evaluation.

An ambitious evaluation of general budget support (GBS) was recently completed. This evaluation used case studies rather than a statistical approach. Counterfactual analysis remained informal: the evaluators considered the plausibility of various alternative scenarios. As a result, they could not say anything in quantitative terms on the ultimate question: the impact of GBS on poverty. The synthesis report is very clear on this: "Study teams could not confidently track distinct PGBS effects to the poverty impact level in most countries".[10]

A third possibility is to apply statistical impact evaluation, but in such a way that conclusions can be drawn at a higher level of aggregation than that of the individual project. It should be emphasized that this is largely virgin territory. While the methodology for statistical impact evaluation at the project level is well established, there have as yet been no attempts to aggregate the results.[11]

The key idea is to select a sample of households (representative of the intended beneficiaries of the policies to be evaluated), to identify the interventions they have been subjected to in a particular period and then apply statistical impact evaluation to each of those interventions. On the basis of the evaluations of the individual interventions, one then arrives at an assessment of policy impact at the aggregate level.

The second step, identifying the relevant interventions, is in itself a major exercise. It requires a detailed description of the (aid-supported) activities in the sector concerned in the period considered. Where donors have switched to a common pool approach, it no longer makes sense to evaluate the impact of the aid provided by a particular donor. Rather, the activities to be evaluated should in that case include *all* interventions undertaken by the Ministry responsible

specification applies to all countries, so that differences in growth rates can only be explained by differences across countries in the values of the regressors used.

[9] See Bigsten et al. (2006) for discussion and references.

[10] IDD and associates (2006, p. S7).

[11] The evaluation agency of the Dutch Ministry of Foreign Affairs (IOB) has started a series of such evaluation studies to test the feasibility of this approach.

for the sector concerned or (in the case of general budget support) all activities of the Government.

The second step is to apply statistical impact evaluation to each of the activities in the sample. Here, there are two key questions. First, one needs to consider whether the way in which the activity has been implemented allows one to choose a convincing control group. As noted above, this is straightforward when randomisation has been used (either intentionally or by accident, the case of "natural experiments"), but unfortunately this is rare. More commonly, the programme has been implemented sequentially or partially so that treatment and control groups can be identified. However, non-randomness implies that the two groups may well differ systematically in other ways than in having received treatment or not. Whether one can adequately control for this (with methods such as propensity score matching) depends on the answer to the second question, that of data availability. Where baseline data have been collected, these may need to be complemented with new (post-intervention) data. In some cases, it may be possible to use existing census or survey data if these allow the identification of treatment and control groups.

The end result is a statement of the form "public spending (possibly limited to particular sectors) in country X reduced poverty by so much".[12] It is important to note that this statement is not aid-specific. Under general budget support (or under the sector approach), the only sensible way of estimating the contribution of aid is to apply the share of aid in total public expenditure to the estimated total impact. It should also be noted that this methodology does not distinguish between the effect of aid through conditionality-induced policy changes on the one hand and the financing role of aid (aid enables more activities under unchanged policies) on the other hand. The evaluation will pick up the total effect of aid without being able to disentangle the contribution of these two channels.

Applying statistical impact evaluation to a large number of activities can be described as a bottom up approach. A point to note is that this approach to evaluation will reveal differences in returns between various government activities. For example, some types of schooling programmes may turn out to be much more effective than others.

[12] To establish effectiveness, the estimated impacts should be related to inputs, e.g., to total educational spending in the period considered. Note that this will measure both the extent to which the funds actually reach the intended activities (the issue addressed by tracking surveys) and the extent to which those activities succeed.

The evaluation is then informative not only on the average return on educational spending, but also on whether the portfolio of activities within the sector is efficient. This is important: if efficiency is rejected, then there is scope for raising effectiveness by expanding some activities at the expense of others. The same applies to differences in returns across (rather than within) sectors.

There are also disadvantages. First, precisely because the approach attempts to correct for all factors which might have influenced the observed outcomes, it is data intensive. (However, in some cases, the necessary data will already have been collected for other purposes, e.g., for poverty assessments.) Second, the evaluation establishes whether (and to what extent) an intervention was effective, but not why. In terms of the drugs testing analogy, it will indicate that the drug is effective, but it does not identify the active ingredient. (I return to this "black box" critique in the next section.)

3. Heterogeneity of "treatment": Beyond binary evaluation

Statistical impact evaluation is designed for binary situations: situations where treatment is homogeneous so that it is clear whether a policy intervention applied to a particular participant or not. Then, there is no ambiguity as to who belongs to the treatment group and to the control group, respectively. To take an example from the education sector, the intervention to be evaluated might be a conditional cash transfer programme (active for a limited period) and the treatment group would consist of the households which received transfers. The evaluation methods discussed in the previous section are designed for such "binary" interventions.

Unfortunately, support for sector programmes or general budget support cannot be evaluated in this way. Aid for the education sector might have been used to fund many different interventions: construction of schools, provision of teaching materials, training of teachers, or cash transfers to increase enrolment. Any school might have benefited from several of these interventions. Schools will differ both in *what* they benefited from and *when*. The implication of such heterogeneity of treatment is clear: there is no longer an obvious distinction between treatment and control groups.

In this situation, an evaluation can obviously not be based on a comparison of a treatment and a control group. However, treatment heterogeneity implies variance which can be exploited to estimate the effect of various interventions. The idea is simple: if panel data are available for all possible determinants of impact (including variables measuring "treatment"), then regression analysis can be used to estimate the impact of interventions.

As an example, consider how a complex programme of water and sanitation activities might be evaluated.[13] Here, a sensible unit of observation would be a location (perhaps a village, but possibly only a part of a village) with a common source of water (e.g., a particular type of well) and a common history of training in, say, latrine construction and hygienic practices. Then, locations do not only differ in *whether* they have a well but if so, also in when that well was installed, what type it is, whether it has been rehabilitated, how far away it is located, whether the villagers received training in hygiene, the nearness of latrines and in many other ways.

Ideally, data on these location-specific "treatment histories" have already been collected prior to the evaluation. Unfortunately, this is rarely the case. Usually the details of a particular government programme, what was undertaken, when and where, are no longer available or were never adequately recorded. The evaluator is then expected to do the impossible: to assess the impact of a set of interventions which are not clearly known. However, it may well be possible to collect such data retroactively, *i.e.* by relying on recall data for major events such as the construction of a well. For each type of treatment, this would generate time series, either as a dummy variable (e.g., indicating whether the location received a particular type of training in a certain year) or in continuous form. Ironically, collecting "treatment histories" is likely to be the major task in an evaluation. Once these data have been collected, the impact evaluation itself is a relatively simple exercise.[14]

For the regression analysis, we also need observations for an impact variable, e.g., the incidence of a waterborne disease such as chol-

[13] I here draw on Elbers and Gunning (2006) who describe such an evaluation of series of activities in water and sanitation in one region in Tanzania, Shinyanga, over a 35-year period.

[14] It is, of course, possible that data availability is better for relatively successful projects, so that these may disproportionately qualify for inclusion in an evaluation. This selection effect may be more serious for the proposed, data-intensive approach than for less formal evaluation methods.

era.[15] Changes over time in the impact variable can then be regressed on changes in explanatory variables. The regressors will include the treatment variables for water and sanitation and also any non-programme variables which may have affected the impact variable. (As always, one will have to confront endogeneity issues, e.g., selection effects as a result of the non-random assignment of interventions to locations.)

By using changes rather than levels, unobserved (and time-invariant) differences between locations will be filtered out. Such a fixed-effect panel approach is an extension of the well-known double differencing method to our case where treatment is no longer a binary variable and impact is measured more than twice. Double differencing uses data for two periods, a baseline ($t = 0$) and a period ($t = 1$) when members of the treatment group participate in the programme ($P_i = 1$). In our case, P is no longer a single binary variable but a set of (possible continuous) variables describing how a household or a location has been affected by the various interventions in the programme.

An obvious implication is that as compared to the standard case, we need more observations since there are now more coefficients to estimate, one for each element of P. It is important to note that programme effects are not identified by comparison of a treatment and a control group, but by the differences in intervention histories between locations. This strategy would seem applicable in many situations: there is often enormous variance in intervention histories which is an advantage from the econometric point of view.

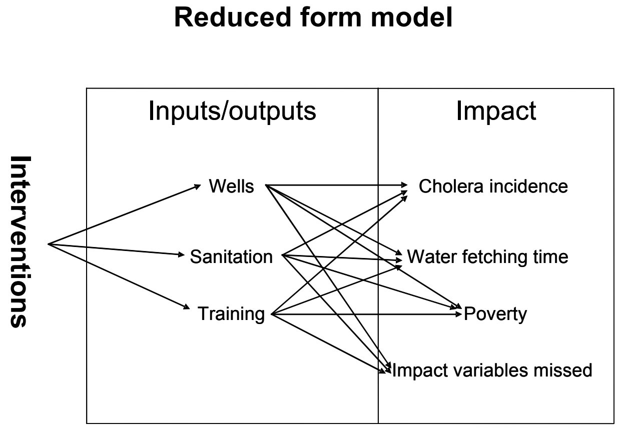
In many evaluations, inputs are seen as leading to impact via the intermediate outputs and outcomes. It is therefore appealing to follow this logical sequence in the evaluation. Instead, our approach directly relates impact variables to inputs, thereby bypassing the output and outcome variables. Statistically, this amounts to estimating a reduced form rather than a structural model.

This is illustrated in Figure 1. The interventions might have numerous impacts, some of which would be missed in the evaluation. In this example, three impact variables are considered: changes in cholera incidence, changes in the time used by household members (typically women and girls) to fetch water and changes in poverty. In the fixed-effect regression these impact variables are directly regressed on

[15] In Tanzania such data are collected at dispensaries.

measures of the number of wells and latrines constructed and the training (e.g., in hygienic practices) provided.

Figure 1. Impact analysis in the water and sanitation sector: Reduced form estimation

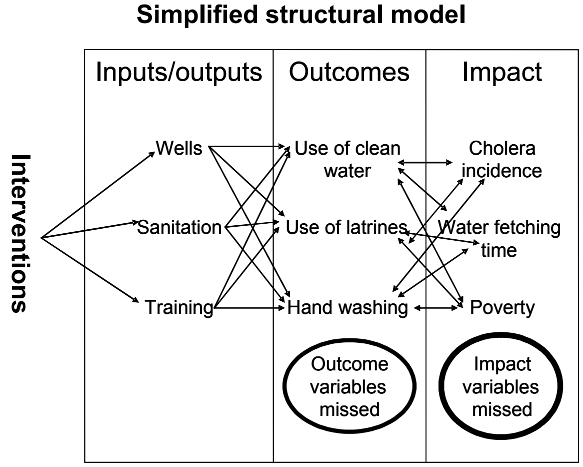


Source: Elbers and Gunning (2006).

Two aspects of this simple regression design are worth noting. First, interaction effects are likely to be important and these can easily be incorporated. For example, the extent to which the availability of a well will offer protection against cholera obviously depends on its proximity to latrines. An interaction variable will pick this up. The coefficient of the interaction term will then provide very useful evidence on whether the balance between the two types of interventions was appropriate. For example, it could show that there was overinvestment in wells in the sense that the return to improved sanitation should have been higher.

Second, as Figure 1 makes clear, one can allow for multiple impacts of the same intervention and, conversely, for various interventions affecting the same impact variable. This is likely to be important in practice. Being able to accommodate them is an important advantage over other (non-regression) impact evaluation methods.

Figure 2. Impact analysis in the water and sanitation sector: Structural form estimation



Source: Elbers and Gunning (2006).

Evaluators tend to think in terms of a "log frame" (logical framework) and hence, a logical progression from interventions through inputs, outputs and outcomes to impacts. In such a framework, estimation of a structural model (as in Figure 2) would seem more attractive than the reduced form estimation shown in Figure 1. Estimating a structural model would involve relating impact variables to outcome variables, outcome variables to output variables and so on. Unfortunately, this attractive approach is riddled with estimation problems, since some of the regressors are bound to be endogenous.

For example, if "use of latrines" was used as an explanatory variable in the regression for cholera incidence while "hand washing" was not, then there would be an endogeneity problem: the availability of wells affects cholera incidence through both these variables and the omission of the "hand washing" variable would therefore lead to biased estimates. Omission of an explanatory variable in structural form estimation is likely.

There are technical solutions for such endogeneity problems, notably instrumental variable estimation. For example, the interventions could be used as instruments for outputs or outcomes. The problem is that this will not produce a sufficient number of instruments in a situation such as the one depicted in Figure 2.

This is a reason to prefer reduced form estimation, which will produce estimates of the effect of inputs (e.g., shallow well construction) on impact variables (e.g., cholera incidence). Such estimates are very useful in themselves, since they allow an assessment of the effectiveness of the intervention, presumably the main objective of the evaluation.

However, it may be objected (as noted in the previous section) that this is a black box approach since it does not explain *why* cholera responded to the construction of the well.[16] If this is seen as a problem, one could complement the statistical exercise with more informal, descriptive methods to assess whether outcome variables which are known to intermediate the effect of water availability on cholera incidence (e.g., the amount of contaminated water consumed) have also indeed improved. This would make an estimated effect of the well on cholera incidence all the more credible.

However, we can go further in the case of multiple interventions. In Figures 1 and 2, the impact variables are affected by more than one policy. For example, cholera incidence is affected by three different inputs: interventions aimed at providing clean water, installing improved sanitation facilities and hygiene training. An attractive feature of the proposed approach is that we can account for the impact of each of these interventions. Far from having to accept an impact estimate as a black box result, one would be able to attribute it to these various interventions: one could calculate how much of the decline in cholera can be attributed to each of the interventions and their interaction.

This is extremely useful for policy design: the evaluation may indicate that building wells was effective in itself, but that the return would have been higher if more had been invested in sanitation facilities. In this sense, the approach will be much more informative than the traditional randomized evaluation designed for a single policy intervention.

Figures 1 and 2 illustrate a complex situation. It is worth stressing that this does not preclude rigorous evaluation, provided that all relevant interventions and impact variables can be measured. Indeed, in

[16] This is a general objection, not specific to our example with heterogeneity of treatment.

an important respect the method is simpler than alternative approaches: reduced form estimation implies that there is no need for estimating the intermediate effects, e.g., from outputs to impacts.

4. Externalities and non-linearities

Non-linearities in impact are likely to be important in practice. For example, preventive health measures (such as vaccination programmes) are typically highly non-linear, with sharply declining marginal returns when coverage is extended to a larger part of the population. Similarly, the extent to which farm households can benefit from price increases for the crops they sell, depends on the extent of market integration and transport costs. As a result, the impact of pricing policies on rural poverty will depend on infrastructure and competition policies so that there are interaction effects.

There is no inherent reason why non-linearities (including interaction effects) cannot be accommodated in the regression equation. In many cases (such as the interaction between water and sanitation policies in affecting cholera incidence), it would be essential to allow for non-linearities. However, to reliably identify non-linearities, one will usually have to estimate a larger number of parameters and this may require an increase in sample size.[17]

A separate concern in the evaluation literature is whether unintended effects of interventions are properly accounted for. Externalities (in this sense of the word rather than what economists understand as an externality) arise for example when an intervention does not only affect the intended beneficiaries but others as well. Janssens (2005) evaluated a women empowerment programme in Bihar (India) and found strong evidence of such externalities.

In this particular case, villages had been selected randomly for inclusion in the programme. However, within a programme village women could freely choose whether to participate or not. Rather than simply comparing participating women in the programme villages with a control group in the non-programme villages, Janssens considered a third group: the non-participating women in the programme

[17] This limitation is not specific to what we propose: it applies to most evaluation methods. An exception is the cross-country regression methodology which—by aggregating over different activities—in principle allows for interaction effects between them. The price one pays for this is that the individual contributions can no longer be identified.

villages. She could convincingly estimate how these would have behaved in the absence of the programme by matching them with women in the non-programme villages. (Here the random assignment of villages to the programme was of course a great help.) She found a large effect of the programme on participants but she also found a smaller but still quite large effect of the programme on the non-participants in the programme villages. This would have been missed in a conventional evaluation design, resulting in a substantial underestimate of the programme's impact.

What is the implication of such externalities for the procedure advocated in the previous section where sector or economy-wide impacts are estimated on the basis of a sample? First, possible externalities have to be considered when choosing the unit of observation and the stratification of the sample. For example, if the unit of observation is the household, then one should sample households randomly in the village (or whatever the area where the externality is supposed to work). Conversely, if the unit of analysis is a village (or another geographical area) then using village level data, one will automatically pick up any externalities operating within that area.

5. Conclusion

Unless one accepts Bob Geldof's view of development ("Something must be done. Anything. Whether it works or not.") there is scope for making policy debates more evidence-based. While many people hold strong positions, there is not yet a great deal of evidence on what works in development (at least beyond the great issues such as openness versus inward-looking strategies) and much scope for experimentation (Easterly, 2006) and rigorous evaluation.

Donors have responded to questions about aid effectiveness with evaluations at a high level of aggregation, e.g., using cross-country growth regressions or country case studies to assess the impact of aid on economic growth or poverty. In this paper, we have argued in favour of a bottom up approach whereby the impacts of general budget support or aid-supported sector programmes are assessed on the basis of statistical impact evaluation of the interventions affecting a representative sample of intended beneficiaries. Heterogeneity of "treatment" (in the nature and timing of activities) suggests a fixed-effect panel data approach to estimate impact. We have argued that rigid adherence to a logical framework is likely to run into serious endogeneity problems and that reduced form estimation is more appropriate.

The proposed approach can be used in situations with multiple interventions (as in our water and sanitation example) and multiple impacts. In such situations, it will generate information on the relative effectiveness of various activities.

References

Bigsten, A., Gunning, J.W. and Tarp, F. (2006), The effectiveness of total ODA: An evaluation proposal, paper for the DAC Evaluation Network, OECD.

Dijkstal, H. (2006), Vertrouwen in een kwetsbare sector?, (Trust in a vulnerable sector?'; in Dutch), Report of the committee on public support in The Netherlands for development cooperation in relation to its effectiveness (the "Dijkstal Committee"), April.

Duflo, E. (2005), Evaluating the impact of development aid programs: The role of randomized evaluation, Paper presented at the third AFD-EUDN Conference, Paris.

Easterly, W. (2006), The White Man's Burden, Penguin Press, New York.

Elbers, C. and Gunning J.W. (2006), Evaluation of Dutch development assistance in water and sanitation, Shinyanga Region, Tanzania, 1990-2006, Proposal to the Tanzanian Ministry of Water and the Dutch Ministry of Foreign Affairs, April.

IDD and Associates (2006), Evaluation of General Budget Support: Synthesis Report, International Development Department, University of Birmingham.

Janssens, W. (2005), Measuring externalities in program evaluation, Discussion Paper 05-017/2, Tinbergen Institute, Amsterdam.

Miguel, E. and Kremer, M. (2004), Worms: Identifying impact on education and health in the presence of treatment externalities, Econometrica 72, 159-217.

Porter, R. (1997), The Greatest Benefit to Mankind: A Medical History of Humanity from Antiquity to the Present, HarperCollins, London.

Ravallion, M. (2001), The mystery of the vanishing benefits: An introduction to impact evaluation, World Bank Economic Review 15, 15-40.

Solow, R. (2001), Applying growth theory across countries, World Bank Economic Review 15, 283-88.