TI 2007-075/2
Tinbergen Institute Discussion Paper

Assessing Budget Support with Statistical Impact Evaluation

Chris Elbers
Jan Willem Gunning
Kobus de Hoop

VU University Amsterdam, Tinbergen Institute, and AIID.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam

Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam

Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Assessing Sector-Wide Programs with Statistical Impact Evaluation: a Methodological Proposal

Chris Elbers[i], Jan Willem Gunning[ii] and Kobus de Hoop[iii]

January 2008

forthcoming in *World Development*

[i] Chris Elbers, VU University Amsterdam, Faculty of Economics and Business Administration, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, phone +31 20 5986143, fax +31 20 5986004, e-mail: celbers@feweb.vu.nl

[ii] Jan Willem Gunning, VU University Amsterdam, Faculty of Economics and Business Administration, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, phone +31 20 598 6141, fax +31 20 5986004, e-mail: jgunning@feweb.vu.nl

[iii]Kobus de Hoop, VU University Amsterdam, Faculty of Economics and Business Administration, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, phone +31 20 5986145, fax +31 20 5986004, e-mail: jhoop@feweb.vu.nl

All three authors are affiliated with the Tinbergen Institute and the Amsterdam Institute for International Development.

Abstract

Donor agencies and recipient governments want to assess the effectiveness of aid-supported sector policies. Unfortunately, existing methods for impact evaluation are designed for the evaluation of homogeneous interventions (‘projects’) where those with and without ‘treatment’ can be compared. The lack of a methodology for evaluations of sector-wide programs is a serious constraint in the debate on aid effectiveness.

We propose a method of statistical impact evaluation in situations with heterogeneous interventions, an extension of the double differencing method often used in project evaluations. We illustrate its feasibility with an example from the education sector in Zambia.

Keywords: Impact Evaluation, Sector-Wide Programs, Aid Effectiveness, Education, Africa, Zambia

Acknowledgements

We are grateful to Jean-Louis Arcand, Arne Bigsten, Antonie de Kemp, Jean-Philippe Platteau, Jacob Svensson, Finn Tarp and Rita Tesselaar for many helpful discussions on this topic.

Assessing Sector-Wide Programs with Statistical Impact Evaluation: a Methodological Proposal

1. Introduction

For many decades discussions on what works and what does not in world development have been characterized more by ideology and arm-chair theorizing than by appeal to evidence. Many public sector interventions in developing countries are not evidence-based and in policy debates in donor countries professionals do not enjoy noticeably more credibility than self-proclaimed development experts such as rock star Bono. However, this is beginning to change: concerns about aid effectiveness have led to a demand for higher standards in evaluations of aid-supported interventions (Duflo, 2005; Tarp, 2006; Gunning 2006) and the enormous improvement in the availability of both macro and micro datasets has made it feasible to meet this demand.

That it is feasible to test interventions in development rigorously, much like medical drugs are tested, has been argued convincingly and eloquently by many authors. An excellent (and very entertaining) introduction to this field is Ravallion (2001) and very useful recent overviews are given by Duflo (2005) and by World Bank (2006).

Statistical impact evaluation methods are designed for ‘projects’, where the intervention (‘treatment’ in the jargon) is homogeneous: it is well-defined and identical for all members of the ‘treatment group’. This makes it feasible and sensible to infer the impact of the intervention from a comparison of a treatment and a control group. However, nowadays the evaluation question is often quite different. Donors have started to move away from project finance in favor of sector aid and general budget support. As a result, ironically, donor agencies are becoming interested in statistical impact evaluation techniques (designed for narrowly defined projects) at the very time when their evaluation demands have shifted, making these existing techniques less suitable. This has led to methodological confusion. Donors want to assess the effectiveness of aid at the sector or even national level but it is not clear how this should be done.[1]

In this paper we address this dilemma for the case of sector-wide programs. We argue that existing statistical impact evaluation techniques can be modified in such a way that they become suitable for evaluating such programs. The methodology we propose requires “intervention histories” for a representative sample of the target population. For example, in an education sector evaluation one would need to have data at the level of schools on the nature and timing of government controlled school and teacher characteristics, e.g. the availability of textbooks and the level of training of the teachers. In many developing countries education ministries already maintain data bases with this type of information. The intervention histories have to be complemented with impact measures at the level of schools, e.g. the quality of schooling as measured by exam scores or standardized national assessments. The proposed methodology then involves a regression of exam scores on the intervention history variables. Such regression results can be used (but this is outside the scope of the paper) to obtain an estimate of the aggregate impact of all the various schooling interventions. That estimate is needed to addresses the question (relevant for both donors and recipient governments) whether the money spent on, say, education in a particular period was well spent in the sense that it achieved a significant and substantial improvement in terms of exam results or enrollment.

In this paper we focus on the first part of such an effectiveness study: the use of regressions to obtain estimates of the effect of all the heterogeneous interventions in a sector. The feasibility of this approach has now been established in a number of evaluation studies.[2]

It should be emphasized that the method provides an *ex post* assessment unlike an *ex ante* evaluation where one is concerned about the future impact of current allocations to the sector. The results of *ex post* evaluations can, of course, inform *ex ante* evaluations but it is useful to keep the distinction in mind. For example, investment in education may have been highly successful in the past but because of diminishing returns an *ex ante* evaluation may indicate that continuing the same types of investment will have much less impact.

The structure of the paper is as follows. In section 2 we discuss recent developments in statistical impact evaluation and the shift in donor demands towards evaluations at a much higher level than that of individual projects. We present and discuss our methodological proposal in section 3. In section 4 we use some results from a recent evaluation of primary education in Zambia to illustrate how the method can be used. Section 5 concludes.

2. Statistical Impact Evaluation[3]

There are few public sector activities which are as often and as intensively evaluated as development aid. Nevertheless there still is remarkably little systematic evidence on what does and does not work in development. The apparent contradiction is easily resolved. The vast majority of development evaluations are focused on process rather than on impact and on recording changes rather than on attribution of changes to interventions. Consultants who specialize in evaluations of development activities are usually very good in establishing what happened and why. They report, of course, to what extent targets were achieved but typically they do not attempt to establish rigorously whether observed changes can be attributed to the intervention.[4]

As a result the fundamental evaluation question: what and how much was achieved *as a result of this intervention*? usually remains unanswered. This is changing rapidly: the debate on aid effectiveness has caused a surge of interest in better evidence and hence in formal impact evaluation techniques.[5] Often these techniques can indicate not only whether the intervention had an effect but also the size of that effect. They therefore provide a quantitative assessment which can be used in a cost-benefit analysis.

Impact evaluation relies on comparing groups with and without ‘treatment’. However, obviously, no group can be observed at the same time in both situations. 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 not have received treatment. Rather than comparing the same group with and without treatment at the same time (which is impossible) one compares results for two different groups. (The hypothetical nature of the counterfactual is sometimes used as an argument against statistical impact evaluation: the methodology is then dismissed because it requires estimates of what would have happened in a hypothetical situation. This objection simply ignores the evaluation problem.)

Ideally, impact evaluation involves a comparison of two randomly selected groups, a treatment and a control group. This is the experimental design familiar from, for example, the testing of medical drugs. In this setup the control group provides the counterfactual: since participants in the experiment have been assigned randomly to the two groups, there is no reason to suppose that there are any (statistically significant) differences between the two groups other than that one group is exposed to treatment while the other one is not. The control group can therefore be used to infer what would have happened to the members of the treatment group in the hypothetical case when they would have received no treatment. Any significant differences in results between the two groups can therefore be attributed to the treatment.

Random assignment is often not feasible but if it is (e.g. because an intervention is implemented sequentially so that there is scope for randomization in the order in which, say, different locations are given treatment) then it certainly should be used (Duflo, 2005).

In policy evaluations one often has to accept non-random assignment. Consider the case of an evaluation of an employment promotion policy, say a training program. A traditional evaluation would simply rely on before-and-after comparisons: did a group of unemployed workers succeed in finding jobs after participating in a training program? Such comparisons clearly suffer from a selection effect. If candidates self-selected themselves into the program then their success in finding a job need not reflect the impact of the training: those who signed up for the program might have (unobserved) characteristics that made them more likely than others to find jobs in the absence of the programs. Clearly, a before-and-after evaluation is then meaningless. If the evaluator is not allowed to assign workers randomly to the two groups then he has to correct for selection effects. Labor market research has a strong tradition of using rigorous statistical impact evaluation to construct convincing counterfactuals for such cases (Heckman *et al.*, 1999).

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 (e.g. Newman *et al.*, 2002), schooling programs targeted at the poor (Sadoulet *et al.*, 2001), health interventions (Pradhan *et al.*, 2007) and even rural empowerment programs (Janssens, 2007). Thorbecke (2007) has described impact evaluation as “arguably the most important contribution to development doctrine” in the present decade. As in the case of labor 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 now experimenting with such methods. Indeed, even quite small donor agencies have started to use these techniques: one of the best-known evaluations (Miguel and Kremer, 2004) describes an evaluation of primary schooling in Kenya which was initiated by a small NGO, ICS Africa.

Thorbecke (2007) also mentions weaknesses of impact evaluation. Most importantly, impact evaluation answers an important but narrow question: does an intervention have an effect? It does not explain why or how the effect occurred, only that it occurred. In addition impact evaluation usually ignores general equilibrium effects.

In the absence of random assignment there may be systematic differences between the two groups. One can often correct for the resulting bias in the evaluation (with methods such as propensity score matching, see e.g. Rosenbaum and Rubin, 1983) if the differences are measured but, of course, there are likely to be unobserved differences. The availability of baseline data is then of crucial importance. If baseline data are available then one can measure changes over time for both groups. Impact can then be assessed as the difference between the two groups in those changes over time (“differences in differences” or “double differencing”). The method can easily be extended to a multi-period context. This is important since in many practical situations the target group is affected not just by current interventions but also by previous interventions. Such lagged effects could turn out to be important.

Policy makers are understandably reluctant to invest in the collection of baseline data but there is a growing awareness that without such data it is quite difficult to assess the results of an intervention in a convincing way. Also, policy makers increasingly accept that where implementation of an intervention is gradual (e.g. 25% coverage of the villages concerned in the first year, 50% in the second year and so on) there is a strong case for using random assignment of villages to the various rounds of implementation.[6]

When statistical impact evaluation is used at the project level treatment is usually well defined and the same for all members of the control group.[7] Also, it is clear from the project’s objectives how success is to be defined. For example, if the project involves offering cash transfers to poor households conditional on the (continued) school enrollment of their children then this intervention is the same for all households in the target group.[8] Given the project’s objective its impact should obviously be measured in terms of enrollment 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, in recent years donors have moved away from project aid. Increasingly aid is given as sector support or general budget support. This is problematic for assessing aid effectiveness: the evaluation question must now be considered at a higher level of aggregation, a level for which the techniques of statistical impact evaluation have not been designed. This has contributed to methodological confusion. NGOs and donor agencies are under great pressure to demonstrate the effectiveness of their work but they are not sure how sector aid or general budget support should be evaluated.

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 aid (and its various components) on economic growth. 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.[9] Results are far from settled and much of the work in this area fails to pass tests of robustness.[10] In addition to econometric weaknesses this approach has the disadvantage that it generates very limited information. Most importantly, it does not indicate the relative effectiveness of the various aid-supported activities (e.g. education versus water supply), information which both donors and recipient governments hope to obtain from an evaluation.

An alternative to cross-country regressions is to rely on case studies. This was the approach adopted in a recent ambitious evaluation of general budget support (IDD and Associates, 2006). In this massive study, supported by 18 bilateral donors and a host of multilateral institutions, counterfactual analysis remained informal: the evaluators used their judgment in assessing the plausibility of various alternative scenarios, for example the magnitude and composition of aid in the absence of a switch to general budget support. Hence no “hard” conclusions were possible and the synthesis report is in fact extremely careful: “Study teams could not confidently track distinct (separately identifiable) PBGS [Partnership General Budget Support] effects to the poverty impact level in most countries. This applies more particularly to income poverty and empowerment dimensions. There are some clear links from PGBS to improved basic services, through funding and through a collective commitment of donors and governments to service delivery targets. …. This somewhat agnostic finding largely reflects the difficulty of data, time-scale and methodology .. ”.[11]

Cross-county regressions and case studies therefore have severe limitations. In this paper we propose an alternative: to apply statistical impact evaluation but in such a way that conclusions can be drawn at a higher level than that of the individual project. This is still largely virgin territory. The methodology for statistical impact evaluation at the project level is well established but such methods have only just started to be used to assess sector-wide policies and sector support.[12]

Our proposal (discussed at greater length in the next section) involves three steps. First, a sample (of households or locations) is drawn which is representative for the target population. Secondly, for each unit in the sample data are collected on impact variables (e.g. poverty) and on the possible determinants of these variables, including policy variables. This requires the collection of “intervention histories”. Finally, the variables are used in a reduced form regression. For example, if one is interested in the health effects of a water supply program, one would regress changes in a location-specific health measure (e.g. incidence of a water-related disease such as cholera) on changes in all possible determinants of that incidence, including the location’s water supply characteristics.[13] In this paper we will go no further but in practice one will often want to use the regression results in an analysis of the effectiveness of public spending. For example, the regression coefficients can be combined with the intervention histories recorded for the sample to arrive at an estimate of the total impact of the various interventions. (Basically this would amount to multiplying the change observed for an intervention variable for a sample location by the corresponding estimated coefficients and summing this product over locations and intervention variables). If the sample is representative of the target population the result is an estimate of the total impact. This can be combined with cost data to estimate, say, the cost of reducing poverty by a given number. In this sense "impact evaluation can feed into a full *ex post* cost-benefit analysis" (World Bank, 2006, p. 12).

Regression-based approaches have long been used in impact evaluation, mainly to allow for variables which may affect an impact variable but which do not reflect an intervention. More recently, such approaches have been used when the activities to be evaluated are heterogeneous.[14] We take this one step further by applying the approach to *all* interventions in a particular sector (or a number of sectors) or geographical area. This makes it possible to use the evaluation to assess the effectiveness of, say, all public sector interventions in primary education and hence the effectiveness of aid provided as educational sector support. In traditional project evaluations one implicitly assumes that there is no issue of fungibility: the donor-financed project would not have been undertaken in the absence of the aid. That this is unrealistic has long been known. Our approach takes a position closer to the opposite extreme: we assume that activities are fully fungible up to the sector level. The implication is that it makes no sense to trace aid to specific activities. Instead aid is assumed to have increased total public spending in the sector *without affecting its intrasectoral allocation*. It then is appropriate to attribute the total impact of activities in the sector to a donor in proportion to the donor's contribution to sector spending.

Applying statistical impact evaluation to a whole set of activities can be described as a bottom up approach: impact is measured at the level of the ultimate beneficiaries.

An important advantage of this approach is that it will reveal differences in returns between various government activities. For example, some types of schooling programs may turn out to be much more effective than others. 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. Information on these differences can be used to raise the aggregate return by changing the allocation of resources across activities. (This is analogous to the approach in the aid allocation literature where differences in aid effectiveness between countries are used to raise aggregate effectiveness; Collier and Dollar, 2002.)

Many evaluations follow a log frame approach where inputs are seen as leading to impact via the intermediate outputs and outcomes. It is appealing to follow this logical sequence in the evaluation. Instead, our approach bypasses most of the relations in the log frame. In effect we estimate a reduced form rather than a structural model. There are two reasons to prefer the reduced form. First, a log frame amounts to imposing a particular structural form. This is convenient but implies that there is no room left for testing the assumptions on the variables to be included or excluded. While the theory summarized in the log frame may be plausible, situations where there is *no* doubt as to exclusion restrictions must be extremely rare.[15] We therefore prefer to estimate a reduced form without committing ourselves to whether all the regressors considered belong in the equation, let alone to restrictions which would enable us to recover all the structural coefficients. Obviously, the approach can and should be theory-based in the sense that theory gives guidance as to what variables to include in the regression.[16] Secondly, and related, there may simply not be enough instruments available to deal with endogeneity in each of the structural equations. In that case the log frame is a useful device for organizing one's thoughts but no more: estimating each of the structural relations identified in the log frame is simply not possible.[17]

3. Heterogeneity of "Treatment": Beyond Binary Evaluation

The basic idea of our proposal is to evaluate sector-wide policy by linking an exhaustive set of sector-related interventions to a set of objectives. The term 'intervention' should be interpreted here in a broad sense: it does not only consist of special projects, but includes regular policy, inputs and procedures. Typically interventions are not uniformly applied in a sector and they will change over time. The way to identify the impact of overall policy and of policy components is to compare differences in interventions across the sector as well as changes over time to differences and changes in outcomes.

Looking only at the policy variables that are observable at the level of the ultimate beneficiary necessarily excludes some interventions that might well be very effective. A sector-wide administrative reform could boost the effectiveness of teaching without being directly observable at the pupil level. The effect of the reform would be detected along at least two channels. First, it could also affect pupils in some way, e.g. in the form of better-trained or motivated teachers, less teacher absenteeism, etc. Thus the impact of the administrative reform could be inferred from the impact of teacher training and the total improvement of teacher qualification etc. Second, it could affect the sector by reducing the cost of education, thus improving the benefit/cost ratio of the sector. In this paper we do not discuss this second channel.

A regression model might be specified as follows. Let outcome variable Y_{it} depend on a vector of policy variables P_{it}, some control variables X_{it} not related to policy and a 'disturbance' term $\mu_i + \varepsilon_{it}$ explained below:

$$Y_{it} = a + bP_{it} + cX_{it} + \mu_i + \varepsilon_{it}. \tag{1}$$

Here *i* denotes the unit of the analysis (the school, or the pupil), and *t* the time of observation. Say there are two observations for each unit, denoted $t = 0$ and $t = 1$. A good measure for the impact of policy variables is the coefficient vector *b*, so the evaluation problem is reduced to estimating *b*.[18] Typically, the coefficient vector *b* cannot be estimated by means of simple OLS[19] regression. The disturbance term $\mu_i + \varepsilon_{it}$, representing all variables omitted from the analysis, allows for a 'fixed' (i.e., constant over time) effect μ_i, reflecting the possibility that units differ in outcomes even if they do not differ in *P* or *X*. Such fixed effects are known to invalidate the results of simple regression techniques, in particular when they are correlated with intervention variables.[20] One way to deal with fixed effects is to 'difference' the regression equation:[21]

$$Y_{i1} - Y_{i0} = a + b(P_{i1} - P_{i0}) + c(X_{i1} - X_{i0}) + (\varepsilon_{i1} - \varepsilon_{i0}), \tag{2}$$

so that the fixed effect drops out of the equation.[22] In principle, this can be repeated for every outcome variable *Y* of interest. The vector of impact coefficients b can now be estimated consistently if *P* and *X* (or rather their change) are uncorrelated to the (change) in the disturbance term ε. An alternative sufficient condition for consistent estimation of *b* is that *P* reflects truly exogenous policy.[23]

Equation (2) is formally similar to the familiar ‘difference-in-differences’ estimator of more conventional policy evaluation. However, there are important differences. Statistical impact evaluation is designed for binary situations: for every individual in the sample it is clear whether she was in the treatment or in the control group. Moreover, care is often taken to make sure that treatment is the same for all treated individuals. To take an example from the education sector, the intervention to be evaluated might be a conditional cash transfer program (active for a limited period) and the treatment group would consist of the households receiving transfers. Many of the evaluation methods discussed in the previous section are designed for such “binary” interventions. (Dose-response models of course allow for continuous effects.) In terms of the regression equation above, the ‘vector’ of policy variables P_{it} would be a binary number, equaling 1 for treated and 0 for non-treated individuals.

Unfortunately, a set-up like this cannot be used to evaluate sector-wide activities. For instance, an educational policy package contains many interventions such as construction of schools, provision of teaching materials, training of teachers, cash transfers to increase enrollment, affecting the ultimate beneficiary – the pupil – in many ways and in different degrees. In principle one could imagine doing a separate evaluation for each policy intervention and add up the results of each to determine the impact of a policy package. However, results for individual interventions are bound to be affected by the presence and intensity of other policy interventions as well. A more promising evaluation strategy is therefore to exploit policy heterogeneity: schools will differ both in what they benefited from and when and this can be the basis for determining the effectiveness of individual interventions by means of a regression equation such as equation (2).

Of course, estimating the impact of policy in this way breaks down if a policy instrument is the same for all observation units. For instance, national legislation that affects enrollment in schools is the same for all schools. Therefore the impact of the legislation cannot be separated from the effect of the constant *a* in equation (2). A somewhat different difficulty arises if a policy affects several outcome variables and one would like to assess the impact of a policy on an outcome net of the effect on other outcomes. For instance, an increase in the number of teachers in a school could be expected to increase both enrollments (because parents expect better education for their children) and improve exam results (through a decline in the pupil-to-teacher ratio). However, the impact on enrollment counteracts the decline in the pupil-to-teacher ratio leading to a reduced (or even perverse) effect of the increase in the number of teachers on exam results. Clearly, this ultimate effect is the proper one for evaluating sector policy, but one might still want to know what the impact of an increase in teachers is when the effect on enrollment is controlled for. In our example we will take this into account.

4. Example: Education in Zambia

As an example we consider the effect of educational inputs on schooling achievements (English exam scores) in primary education in Zambia.[24] In Zambia the Ministry of Education has data for all primary schools in the country. These cover school characteristics (number of classrooms, toilet facilities, availability of textbooks etc.) as well as teacher characteristics (education, professional training, experience). These data indicate enormous heterogeneity in terms of school characteristics. From a research point of view this is highly attractive: the differences between schools allow us to identify the effect of policy interventions. The Ministry data have been linked at school level to data from the Exam Council of Zambia for grade 7 pupils taking exams in English and mathematics. We consider the exam scores as our measure of impact and the question is to what extent these can be explained by school and teacher characteristics.[25]

Most of our data are for 2003 ($t = 0$) and 2006 ($t = 1$). School characteristics (but not exam scores) are also available for 2002. In line with equation (2) we regress changes in English exam scores (2003-6) on changes in the log of: the number of English textbooks, the number of classrooms, the number of teachers and, in addition, on changes in an index of the professional quality of the heads teacher and changes in a dummy indicating the availability of flush toilets.[26] To capture possible lagged effects we include level variables for 2002 as well.[27] (Table 1).

(Table 1 here)

.

The results of this initial regression are quite disappointing: the fit is poor and at the 5% level only one of the variables is statistically significant.

Recall that we have chosen a reduced form specification. This implies that the effect of the number of teachers (treated as an exogenous policy variable) is the *total* effect. It includes not only the direct effect (with more teachers pupils presumably get more attention and therefore achieve better exam scores) but also the indirect effect: a higher number of teachers may make the school more attractive to parents and therefore increase enrollment. However, enrollment will (controlling for the number of teachers and other school inputs) have a negative effect on the quality of teaching. Our reduced form estimate therefore measures the net effect of two opposing forces. We cannot even be sure of the sign of the net effect. (In the Table 1 regression it is negative but insignificant.) A reduced-form regression of enrollment is reported in Table 2 below. It shows a strong positive trend and significant effects of books, classes and teachers, consistent with the interpretation that policy variables have led to an increase in enrollment but that enrollment has affected exam scores negatively, defeating any positive direct effects of policy. Whether the overall assessment of policy turns out positive or negative will therefore depend on the relative weight attached to enrollment and exam scores.

(Table 2 here)

Tables 1 and 2 show what reduced-form regressions can tell us. If in addition we want to estimate the direct and indirect effect *separately* we must add enrollment as a regressor in the exam score regression but take into account that that this variable is likely to be endogenous. We would therefore want to instrument for enrollment. However, this requires imposing structural assumptions on the exam scores equation. We will assume (for illustrative purposes) that the level variables for 2002 do not directly affect exam scores, but only operate through the impact of changes in enrollment. Table 3 gives the results.

(Table 3 here)

In Table 3 we regress exam scores on the same policy variables as in Table 1 but now in addition on enrollment (where we use the values predicted by the Table 2 regression). This dramatically changes the results.

The results indicate that exam scores are positively related to availability of textbooks and to the number of classrooms and negatively to school enrollment. These three effects are significant. Head teacher quality, the number of teachers and toilet availability have no significant effect on exam scores. In the case of the number of teachers this implies that the direct effect is quite weak, unlike the indirect effect: in

Table 2 the variable has a t-score of 5.7. Hence in Zambia the number of teachers matters for enrollment, but a direct impact on exam scores cannot be demonstrated.

It may be noted that the effect of the policy instruments is quite small. For example, (since the mean exam score is about 30) the coefficient on books amounts to an elasticity of only 0.01. (This is not to say that textbook availability is unimportant but rather that schools are very heterogeneous in terms of the use they make of available books.) Similar small effects have been reported in the literature, e.g. Hanushek (1995). The most striking finding is the very large (negative) effect of enrollment on exam scores, corresponding to an elasticity of about one third.

5. Conclusion

Increasingly donors are expected to demonstrate the effectiveness of aid. Researchers have responded 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.[28] In this paper we have proposed a bottom-up approach whereby (aid-supported) sector programmes are evaluated on the basis of their impact on a representative sample of the target group.

The proposed methodology can be used to estimate impact parameters for various types of interventions. These can be used in cost-benefit analyses of sector policies or in calculating the impact of aid provided as sector report. They can also be used to study the relative effectiveness of different types of interventions in the same sector. The methodology is backward looking and is therefore suitable for estimating the effect of past interventions. Whether such *ex post* assessments can be used for *ex ante* evaluations has to be decided in each individual case.[29]

We have presented estimates for primary education in Zambia to illustrate the feasibility of the approach. We found that the number of teachers has no significant direct effect on quality (as measured by exam scores), that the effect of the number of classrooms and the availability of textbook *is* significant (but quite weak) and, most strikingly, that enrollment has a strong (negative) effect on educational quality.

[1] Even if were feasible to evaluate all projects interaction effects preclude simply adding up the project effects. See below for a further discussion.

[2] The authors are involved in evaluations of the Dutch Ministry of Foreign Affairs of water and sanitation in Tanzania, education in Zambia and water supply and sanitation in Yemen and Egypt.

[3] This section draws on Gunning (2006).

[4] Indeed, causal relationships are often simply assumed to apply by imposing a "log frame" linking interventions to outcomes.

[5] There is some terminological confusion here since in the evaluation literature the term impact is used in two different senses. It sometimes denotes the effect of an intervention in terms of ultimate objectives such as poverty alleviation or improved literacy. (If used in this sense it is contrasted with inputs or intermediate results which in the jargon are designated as outputs or outcomes.) Alternatively, in the statistical literature impact evaluation refers to any statistical assessment of the effects of an intervention: there is no presumption that these effects are measured in terms of ultimate objectives.

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

[7] There are exceptions. For example, the World Bank Independent Evaluation Group evaluated four educational projects in Ghana which supported "a range of activities" (World Bank, 2006, pp. 24-27). This differs from our case of heterogeneous treatment in considering a narrow subset rather than most or all interventions in the sector. This approach may therefore suffer from omitted variable bias.

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

[9] 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 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.

[10] See Bigsten *et al.* (2006) and Tarp (2006) for discussion and references.

[11] IDD and Associates (2006, p. 72).

[12] 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.

[13] Just as in statistical impact evaluation at the project level one will have to deal with the non-random assignment of the treatment variables. This may involve, for instance, using the Heckman method to model the selection effect. Whether such a correction is needed depends on the purpose of the evaluation. If the question is whether the money allocated to the sector was well spent *taking as given the political processes which might bias the allocation of that money across interventions and across locations* then a correction would be inappropriate. For a technical discussion of this point see Elbers and Gunning (2007). This is a situation which often arises in practice: the donor can shift money between sectors but is powerless to influence the within-sector allocation processes.

[14] For example, as noted before, the Independent Evaluation Group (IEG) of the World Bank used this approach to evaluate heterogeneous educational projects in Ghana. As in this paper the IEG is concerned with counterfactuals and with "final welfare outcomes". However, the IEG does not aim at impact evaluation of "all interventions within a given sector or geographical area" (World Bank, 2006, pp. 1-2).

[15] In the Joint Evaluation of General Budget Support, the approach is illustrated with a log frame (Figure 3, p. 22) which is entirely recursive. Obviously, this is convenient but highly restrictive.

[16] The approach therefore to some extent addresses Thorbecke's (2007) concern that impact evaluation does not explain "the underlying mechanism". Note that if the model is exactly identified it is possible to recover structural coefficients from reduced form estimates.

[17] Elbers and Gunning (2006) provide an example of this for an evaluation of the health effects of water supply and sanitation programs.

[18] The total effect of the policy in period *t* is then given by $\hat{b}\Sigma_i P_{it}$. In a cost-benefit analysis this would have to be converted to a monetary value and compared with the cost of the policy.

[19] Ordinary Least Squares.

[20] For a technical discussion of fixed effects, see e.g. Verbeek (2000, chapter 10).

[21] Besley and Burgess (2000) use a reduced form equation similar to equation (1). They have data for 30 years and are therefore able to estimate fixed effects at the level of the primary sampling unit so that there is no need for differencing. In sector evaluations time series are often quite short necessitating the differencing method we adopt in equation (2).

[22] This can be generalized to the case of more than two observations per unit. A disadvantage of differencing is that measurement error (which we ignore in this paper) and the standard error of the regression are likely to increase. Measurement error is likely to bias the coefficients to zero while their standard error will typically increase. For both reasons coefficients may become statistically insignificant even if there is a true effect.

[23] The policy variables in *P* are not likely to be exogenous in the regression unless they contain essentially all relevant policy interventions affecting the ultimate beneficiaries of policy. Leaving out an important policy variable will lead to omitted variable bias on coefficients of variables that *are* included in the regression.

[24] The example is partly based on work the authors did in the context of a Dutch-Zambian evaluation study, IOB (2008). We are very grateful to Antonie de Kemp who led that study for his comments and encouragement and to the Zambian Ministry of Education for their assistance in using the Ministry's data base.

[25] We treat educational inputs in Zambia as exogenous because educational policy is still mainly initiated at levels well above the target population (schools). Note that if schools would be left completely free in their choice of educational inputs (e.g., subject to an overall budget constraint) it could still be argued that the

resulting regression coefficients represent the *ex post* impact of educational inputs as long as the particular choice of inputs (or rather their change over time) is uncorrelated with the error term in equation (2).

[26] Das *et al.* (2007) find evidence that when government grants increase the effect on test scores is offset by an induced reduction in parental contributions. We can restrict our regression to government schools where e.g. school books are almost exclusively supplied by the government. This reduces our sample by about one third but has very little impact on the estimated coefficients (although standard errors increase). However, where such substitution effects are important one could try to estimate the effect of educational policies on household spending (in addition to its impact on enrollment and exam scores). This would pick up the effect of government spending on parental contributions. This is, of course, beyond the scope of the present paper in which the Zambian results merely serve as an illustration.

[27] The observations included are determined by the IV regression reported in Table 3. The use of level variables to explain changes over time is also similar to the use of initial conditions in empirical growth analysis.

[28] Examples are the many papers by World Bank or IMF staff, e.g. Burnside and Dollar (2000) and Rajan and Subramanian (2005).

[29] For example, if cohort effects in education are important, the marginal effect of educational resources may be below the average effect picked up in an *ex post* evaluation.

References

Besley, T., & Burgess, R. (2000). Land Reform, Poverty Reduction, and Growth: Evidence from India. *Quarterly Journal of Economics, vol. 115,* pp. 389-430.

Burnside, C., & Dollar, D. (2000). Aid, Policies and Growth. *American Economic Review, vol. 90,* pp. 847-868.

Collier, P., & Dollar, D. (2002). Aid Allocation and Poverty Reduction. *European Economic Review, vol. 46,* pp. 1475-1500.

Das, J., Dercon, S., Habyarimana, J., & Krishnan, P. (2007). *When Can School Inputs Improve Test-Scores?* Oxford: CSAE, University of Oxford.

Duflo, E. (2005). *Evaluating the Impact of Development Aid Programs: the Role of Randomized Evaluation*. Paris: paper presented at the third AFD-EUDN Conference.

Elbers, C., & Gunning, J.W. (2007). *Impact Evaluation or Sector-Wide Programs: Is Correcting for Self-Selection Always Desirable?* mimeo, Department of Economics, VU University Amsterdam.

Gunning, J.W. (2006). Aid Evaluation: Pursuing Development as if Evidence Matters. *Swedish Economic Policy Review, vol. 13,* pp. 145-163.

Hanushek, E.A. (1995). Interpreting Recent Research on Schooling in Developing Countries. *World Bank Research Observer, vol. 10*, pp. 227-246.

Heckman, J., Lalonde, R., & Smith, J. (1999). The Economics and Econometrics of Active Labor Market Programs. In O. Ashenfelter, & D. Card (Eds), *Handbook of Labor Economics, vol. 3c.* Amsterdam: North-Holland.

Jalan, J., & Ravallion, M. (2003). Does Piped Water Reduce Diarrhea for Children in Rural India. *Journal of Econometrics, vol. 112*, pp. 153-173.

Janssens, W. (2007). *Social Capital and Cooperation: an Impact Evaluation of a Women's Empowerment Programme in Rural India.* PhD thesis, VU University, Amsterdam; Tinbergen Institute Thesis no. 401.

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

IDD, & Associates (2007). *Evaluation of General Budget Support – Note on Approach and Methods.* Birmingham: University of Birmingham, International Development Department, March.

IOB (2008). *Impact Evaluation of Primary Education in Zambia.* The Hague: Ministry of Foreign Affairs, Independent Evaluation Department.

Miguel, E., & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica, vol. 72,* pp. 159-217.

Newman, J., Pradhan, M., Rawlings, L.R., Ridder, G., Coa, R., & Evia, J.L. (2002). An Impact Evaluation of Education, Health and Water Supply Investments by the Bolivian Social Investment Fund. *World Bank Economic Review, vol. 16,* pp. 241-274.

Pradhan, M., Saadah, F., & Sparrow, R. (2007). Did the Health Card Program Ensure Access to Medical Care for the Poor during Indonesia's Economic Crisis? *World Bank Economic Review, vol. 21,* pp. 125-150.

Rajan, R., & Subramanian, A. (2005). *Aid and Growth: What Does the Cross-Country Evidence Really Show?* IMF Working Paper WP/05/127.

Ravallion, M. (2001). The Mystery of the Vanishing Benefits: an Introduction to Impact Evaluation. *World Bank Economic Review, vol. 15,* pp. 115-140.

Rosenbaum, P., & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effect. *Biometrika, vol. 70,* pp. 41-55.

Sadoulet, E., de Janvry, A., & Davis, B. (2001). Cash Transfer Programs and Income Multipliers: PROCAMP in Mexico. *World Development, vol. 29,* pp. 1043-1056.

Solow, R. (2001). Applying Growth Theory Across Countries. *World Bank Economic Review, vol. 15,* pp. 283-88.

Tarp, F. (2006). Aid and Development. *Swedish Economic Policy Review, vol. 13,* pp. 9-61.

Thorbecke, E. (2007). The Evolution of the Development Doctrine, 1950-2005. In G. Matrovas and T. Shorrocks (Eds.), *Advancing Development: Core Themes in Global Economics.* Basingstoke: Palgrave Macmillan.

Verbeek, M. (2000). *A Guide to Modern Econometrics.* Chichester: John Wiley.

World Bank (2006). *Impact Evaluation: the Experience of the Independent Evaluation Group of the World Bank.* Washington, DC: World Bank.

Table 1: Determinants of English Exam Scores (reduced form regression).

	Coefficient	Rob SE	t	P>|t|
Log of English Books (06-03)	0.274	0.152	1.81	0.071
Log of Classes (06-03)	0.344	0.292	1.18	0.238
Log of Teachers (06-03)	-0.199	0.425	-0.47	0.640
Professional Quality of Head Teacher (06-03)	0.251	0.398	0.63	0.528
Toilets Available (06-03)	0.216	0.541	0.40	0.690
Trend	0.096	0.754	0.13	0.898
Log of English Books (02)	0.068	0.137	0.50	0.617
Log of Classes (02)	-0.234	0.373	-0.63	0.532
Log of Pupils Enrolled (02)	-0.237	0.437	-0.54	0.588
Log of Teachers (02)	0.909	0.345	2.63	0.009
Professional Quality of Head Teacher (02)	-0.130	0.468	-0.28	0.780
Toilets Available (02)	0.568	0.490	1.16	0.246

Dependent variable: changes in exam scores (2003-6). R-square = 0.005. Number of observations: 2495. Robust standard errors are denoted Rob SE. Changes in the period 2003-2006 are denoted 06-03, levels in 2002 by 02.

Table 2: Determinants of Enrollment

	Coefficient	Rob SE	t	P>|t|
Log of English Books (06-03)	0.009	0.005	1.890	0.059
Log of Classes (06-03)	0.030	0.010	3.090	0.002
Log of Teachers (06-03)	0.067	0.012	5.730	0.000
Professional Quality of Head Teacher (06-03)	0.002	0.012	0.170	0.868
Toilets Available (06-03)	-0.012	0.027	-0.430	0.666
Trend (06-03)	0.180	0.023	7.860	0.000
Log of English Books (02)	0.002	0.004	0.400	0.686
Log of Classes (02)	-0.006	0.012	-0.490	0.625
Log of Pupils Enrolled (02)	-0.075	0.014	-5.540	0.000
Log of Teachers (02)	0.025	0.010	2.410	0.016
Professional Quality of Head Teacher (02)	0.004	0.014	0.290	0.770
Toilets Available (02)	-0.029	0.020	-1.430	0.152

Dependent variable: change in log enrollment (2003-6). R-square = 0.32. Number of observations: 2495. Robust standard errors are denoted Rob SE. Changes in the period 2003-2006 are denoted 06-03, level variables for 2002 by 02.

Table 3: Determinants of English Exam Scores (IV regression).

	Coefficient	Rob SE	t	P>|t|
Log of English Books (06-03)	0.308	0.157	1.960	0.050
Log of Classes (06-03)	0.740	0.331	2.240	0.025
Log of Pupils Enrolled (06-03)	-10.974	3.957	-2.770	0.006
Log of Teachers (06-03)	0.237	0.540	0.440	0.661
Professional Quality of Head Teacher (06-03)	0.336	0.328	1.020	0.306
Toilets Available (06-03)	0.190	0.850	0.220	0.823
Trend (06-03)	0.670	0.178	3.770	0.000

Dependent variable: the change (2003-6) in English exam scores (school averages). Number of observations: 2495. Robust standard errors are denoted Rob SE. Changes in the period 2003-2006 are denoted 06-03. IV-regression: enrollment as predicted by the Table 2 regression.