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 Budget Support with Statistical Impact Evaluation: a Methodological Proposal

Chris Elbers, Jan Willem Gunning and Kobus de Hoop
Free University, Amsterdam, Tinbergen Institute and AIID

Revised September 17 2007

Keywords: Impact Evaluation, Budget Support, Aid Effectiveness, Education, Africa, Zambia

Abstract

Increasingly both donor agencies and recipient governments want to assess the effectiveness of aid. Unfortunately, existing methods for impact evaluation are designed for the evaluation of homogeneous interventions ('projects') where those with and without 'treatment' can be compared. However, when assessing the effectiveness of sector aid or general budget support one is concerned about the impact of numerous heterogeneous interventions; there is then no obvious control group. The lack of a credible methodology for such high level evaluations 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 for the education sector in Zambia.

Assessing Budget Support with Statistical Impact Evaluation[1]

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. Public sector interventions in developing countries are rarely evidence-based and in policy debates 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 a recent overview is given by Duflo (2005).

[1] 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.

Unfortunately, existing evaluation techniques do not meet current demands. 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 unsuitable. This has led to methodological confusion. Donors want to assess the effectiveness of aid at the sector or national level but it is not clear how this should be done.

In this paper we address this dilemma. We argue that existing statistical impact evaluation techniques can be modified in such a way that they become suitable for sector or general budget support evaluations. 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. The regression results can be used to obtain an estimate of the *aggregate* impact of all the various schooling interventions. 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: it 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. This is different from the issue in an *ex ante* evaluation where one is concerned about the future impact of current allocations to the sector. The results of *ex ante* evaluations can, of course, inform *ex post* 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

[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 in Yemen.

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 so often and so 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 observed 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. 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.[4] Often these techniques can indicate not only whether the intervention had an effect but also the size of that effect. They

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

[4] 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 then is no presumption that these effects are measured in terms of ultimate objectives. In principle statistical impact evaluation could focus on results in terms of outputs or outcomes.

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 the 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). 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.

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 there may well 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 rather than measuring differences at time *t* (after "treatment") between the two 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 need to be taken into account.

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.[5]

When statistical impact evaluation is used at the project level treatment is well defined and the same for all members of the control group. 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 enrolment of their children then this intervention is the same for all households in the target group.[6] Given the project's objective its impact should obviously 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.)

[5] 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.
[6] An example of such an evaluation is discussed at length in Ravallion (2001).

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 can 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 total aid (or 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.[7] Results are far from settled and much of the work in this area fails to pass tests of robustness.[8] 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),

[7] 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.
[8] See Bigsten *et al.* (2006) and Tarp (2006) for discussion and references.

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 (Joint Evaluation, 2005). In this massive study counterfactual analysis remained informal: the evaluators used their judgment in assessing the plausibility of various alternative scenarios. As a result there is no hope of achieving a quantitative estimate of the impact of general budget support on poverty. The preliminary synthesis report recognizes this: "we cannot confidently track distinct [budget support] effects to this [poverty impact] level in most countries".[9]

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 support.[10]

How can it be done? Our proposal (discussed at greater length in the next section) involves three steps. First, a random sample is drawn, representative of the population. For example, in the case of education one might draw a sample of schools and make

[9] Joint evaluation (2005, p. 16) as quoted in Bigsten *et al.* (2006).
[10] 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.

the probability of being included in the sample proportional to the population size of a school's catchment area. In the second step, intervention histories are collected for the sample schools. This is a record of policy-induced changes at the school level: availability of text books, the number of toilet facilities for girls, the number of class rooms, and the level of training of the head of the school and so on. Sometimes these data will have to be collected at the schools but often they will already be available in data bases at the Ministry of Education or at district or provincial government centers. In addition school level data need to be collected on results (e.g. the exam scores of the pupils of the sample schools) and of variables which may have influences these results (other than the policy variables). The final step then involves a regression of changes in exam scores on changes in policy variables (as identified in the intervention histories) and on changes in other explanatory variables. To take an example from another sector: if one is interested in the health effects of a water supply program, one would want to 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.[11]

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.

[11] 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.

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.)

It should be noted that under a "common pool" approach the interventions evaluated cannot be associated with any particular donor. What we propose will provide an assessment of the effectiveness of a set of interventions (possibly *all* interventions in a particular sector and period). If donors have supported these activities by contributing to, say, the budget of a particular ministry it would be sensible to attribute the effect of those interventions to donors in proportion to their contribution.

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 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. There are two reasons to prefer the reduced form approach which we advocate over the log frame approach. First, the log frame amounts to estimating a structural model. It assumes implicitly that one is certain about which variables appear in each of the structural equations and, conversely, about the exclusion restrictions. In effect the model is exactly identified. 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. 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. 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.[12]

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 an exhaustive set of objectives. The

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

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. This requires a dataset representative of interventions and outcomes at the level of the ultimate beneficiary of policy: the individual. For example, in an evaluation of primary education we look at what happens at the level of the individual pupil: what is the size of classes, how many text books are available for each pupil, etc. Hence, it is natural to sample schools and collect data on policy variables affecting pupils as well as outcomes such as enrolment or test scores.

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 could be traced along 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 incorporating these ideas looks 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:

$$Y_{it} = a + bP_{it} + cX_{it} + \mu_i + \varepsilon_{it}. \quad (1)$$

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*.[13] Typically, the coefficient vector *b* cannot be estimated by means of simple OLS 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.[14] One way to deal with fixed effects is to 'difference' the regression equation:[15]

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

[13] 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.

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

[15] 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).

so that the fixed effect drops out of the equation.[16] 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.[17]

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 support for sector programmes or general budget support. For instance, an educational policy package

[16] This can be generalized to the case of more than two observations per unit.
[17] 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.

contains many interventions such as construction of schools, provision of teaching materials, training of teachers, cash transfers to increase enrolment, 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.

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. 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.

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 (Table 1).

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

	Coefficient	Rob SE	t	P>|t|
Log of English Books (06-03)	0.243	0.146	1.67	0.096
Log of Classes (06-03)	0.382	0.276	1.39	0.166
Log of Teachers (06-03)	-0.531	0.386	-1.38	0.168
Professional Quality of Head Teacher (06-03)	0.275	0.307	0.90	0.371
Toilets Available (06-03)	0.305	0.509	0.60	0.550
Trend (06-03)	0.129	0.057	2.27	0.023

Dependent variable: changes in exam scores (2003-6). R-square = 0.005. Number of observations: 2699. Robust standard errors are denoted Rob SE. Changes in the period 2003-2006 are denoted 06-03.

The results of this initial regression are quite disappointing: the fit is poor and at the 5% level none of the policy variables are statistically significant. Only the time trend is significant. This is in itself an encouraging result, indicating that (controlling for the observed changes in educational inputs included in the regression) exam scores are improving over time.

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. This includes not only the direct effect of the number of teachers (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 attractive to parents and therefore increase enrolment. However, enrolment 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 effects. We cannot even be sure of the sign of the net effect. (In the Table 1 regression it is negative.)

If we want to estimate the direct and indirect effect *separately* we must add enrollment as a regressor but take into account that that this variable is likely to be endogenous. We therefore instrument for enrolment. Table 2 shows a first stage regression with the change in log enrollment in the period 2003-2006 as the dependent variable.[18] Here we treat variables such as the availability of school books as exogenous policy variables and variables measured in 2002 as predetermined.[19] (The use of level variables to explain changes over time is similar to the use of initial conditions in empirical growth analysis.) We find a highly significant effect of policy variables on enrolment, notably of the number of teachers and textbook availability.

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	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

[18] The regressors include both level variables for 2002 (denoted 02) and changes over time (denoted by 03-06).
[19] Econometrically the 2002 variables provide identification since they are not included in the second stage regression.

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.

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 enrolment. 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 enrolment, but not (directly) for exam scores.

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.

The Zambian results are interesting in themselves. In the period considered (2003-6) primary school enrolment grew enormously: the gross rate went from 8 to 109%, the net rate from 78 to 96%. Our results indicate that this generated (as expected) a reduction in quality, as measured by exam scores. Nevertheless, the increase in enrolment in such a short period is in itself a very impressive achievement. Here our purpose is simply to show that it is feasible to use statistical impact evaluation to assess the impact of heterogeneous interventions, as in the case of an evaluation of the education sector.

5. Conclusion

Increasingly donors are expected to demonstrate the effectiveness of aid. They 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.[20] In this paper we have proposed a bottom-up approach whereby the impacts of general budget support or aid-supported sector programmes (or indeed sector policies more generally, whether aid-supported or not) are assessed on the basis of its 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. 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.[21]

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

[20] Examples are the many papers by World Bank or IMF staff, e.g. Burnside and Dollar (2000) and Rajan and Subramanian (2005).
[21] 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.

classrooms and the availability of textbook *is* significant (but quite weak) and, most strikingly, that enrolment has a strong (negative) effect on educational quality.

References

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

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

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

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

Elbers, Chris and Jan Willem Gunning (2007), 'Impact Evaluation or Sector-Wide Programs: Is Correcting for Self-Selection Always Desirable?', mimeo, Tinbergen Institute, Amsterdam.

Gunning, Jan Willem (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., R. Lalonde and J. Smith (1999), 'The Economics and Econometrics of Active Labor Market Programs', in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, vol. 3c, Amsterdam: North-Holland.

Jalan, Jyotsna and Martin Ravallion (2003), 'Does Piped Water Reduce Diarrhea for Children in Rural India', *Journal of Econometrics*, vol. 112, pp. 153-173.

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

Joint Evaluation of General Budget Support (2005), *Synthesis Note. Findings and Issues from Country Studies*, Working Paper.

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

Newman, J., M. Pradhan, L.R. Rawlings, G. Ridder, R. Coa and J.L. Evia (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., F. Saadah and R. Sparrow, (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. and A. Subramanian (2005), `Aid and Growth: What Does the Cross-Country Evidence Really Show?', IMF Working Paper WP/05/127.

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

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

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

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

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

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