The Effectiveness of Foreign Aid: Overview and an Evaluation Proposal

Arne Bigsten[1]
Jan Willem Gunning[2]
Finn Tarp[3]

February, 2006

[1] Department of Economics, Göteborg University, Box 640, SE 405 30 Göteborg, Sweden.
[2] Department of Development Economics, Free University of Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
[3] Department of Economics, University of Copenhagen, Studiestræde 6, DK-1455 Copenhagen K, Denmark.

1. Introduction

Over the last year there has been an extensive discussion within the DAC Evaluation Network about the possibilities of doing comprehensive aid impact evaluations at the country level. Edgren (2004) discusses ways in which total aid contributes to the achievement of desired development outcomes at the country level, and he elaborates on how the impact of the combined support from donors to a particular country might be evaluated. It is also suggested that the following activities are essential in a comprehensive approach: (i) evaluating the overall development cooperation system and associated development outcomes, (ii) assessing the combined ODA contribution,[4] and (iii) reviewing the contributions of individual agencies. It is finally noted that focus should be on process and strategy rather than on "success indicators". The key methodological challenge is to what extent it is feasible to separate out the effect of aid from the effects of other influences.

The Edgren proposal was discussed at a series of meetings, after which we were invited to reflect about what would in our view be doable and defensible. A first version of the current paper was prepared and presented at a workshop in Stockholm on 10-11 November 2005. A common view at the workshop was that our paper was too limited in scope, and that it should be extended to deal more with implementation and in particular with impact. On the basis of the discussions at the November workshop and on subsequent consultations with the steering group of the "Total ODA"-initiative, we revised our paper to the current version.

By way of introduction, we reiterate that there is indeed need for improved aid evaluation. Foreign aid has evolved considerably over the last five decades in response to a dramatically changing global political and economic context; and the allocation, implementation and impact of foreign aid remain contentious issues on both the donor

[4] Edgren suggested more specifically that the analysis should begin with a review of the allocation of resources from ODA between different sectors and lines of activity, along with an assessment of the role that privileged sectors and activities have played in achievements. He emphasized that indirect effects also need to be taken into account, and he identified three questions to consider: First, what is the value added by ODA, its positive contribution to recorded achievements, including (i) considering political economy aspects, and (ii) looking at the relationship between ODA and government revenue and domestic resource mobilisation, as well as non-material resources such as management. Second, the evaluator should take account of the negative effects of aid, for example in the form of increased transaction costs. Third, and finally, factors which enhance aid effectiveness should be reviewed.

and the recipient side of the aid relationship. Seen in this perspective, the initiative at hand is both interesting and associated with considerable potential. At the same time, a renewed aid evaluation effort needs to be based on sound academic principles to produce insights that are both scientifically valid and useful in practice. In this paper we therefore discuss the strengths and limitations of the case study approach, inherent in the Edgren proposal, and we note it is essential that the links between means and ends are demonstrated rather than assumed.

The need for a well identified and commonly understood counterfactual is the key technical issue addressed, and a call is made for insisting on analytical rigour as an essential principle in the evaluation effort. In this revised version of our paper we discuss again how the optimal design of a study of the country level effectiveness of ODA might look like. In response to the various demands made and given our understanding of existing technical constraints we propose an outline of a case study framework, and identify the added value of such an undertaking.

The paper is structured as follows: In the next section we briefly specify what is meant by the evaluation challenge. Section 3 reviews alternative approaches to aid evaluation, and this is followed in Section 4 by a discussion of some ongoing evaluations, which are both relevant and important in the present context. Section 5 motivates our overall proposal, which involves a two-pronged set of activities, and we also discuss what the value added would be. Section 6 presents in greater detail the first part of our proposal, which is focused on an evaluation of the impact of aid on policy choices and policy implementation in recipient countries. We present in turn in Section 7 how evaluation of the impact of total ODA in country cases could in our view be pursued. Finally, we outline in Section 8 the contents of intended country case studies, and concluding remarks are offered in Section 9.

2. The Evaluation Challenge

Before reviewing existing alternative analytical approaches to the evaluation of the impact of foreign aid, it is pertinent to identify three key questions, which are intrinsic to the evaluation challenge. They apply across the board to all evaluations, independent of their scope and other more specific characteristics.

First, it is necessary to clarify what is to be evaluated. A common practice in much of the aid system is to evaluate the delivery of inputs, and answer questions such as whether budgeted funds have been spent as planned. This is fundamentally different from evaluating the ultimate impact of the aid effort, or the socio-economic targets reached. Inputs may – even if they are delivered at the planned time and place – ultimately fail to have their desired effect. A main reason for doing evaluations is to conclude by proposing improvements to the aid process, so aid can become more effective in the future. We do not wish to downplay the importance of increased efficiency in input delivery, but there is in our view a very high potential payoff to be reaped from a better understanding of underlying causal relationships between aid delivery and aid impact. Moreover, the level of ambition in the proposed study is not only to emphasize final impact, observing apparent empirical, or casual, relationships, but to better understand why certain effects occur. We suggest in other words it is time to start addressing the challenge of opening up what is often referred to as "the black box" in aid evaluation. We believe our proposal would mark a step in the right direction, but the complete opening of the black box is realistically speaking not going to happen in the near future.

The second question is whose perspective should be relied on in evaluation efforts. In the Edgren proposal it is argued that a recipient perspective should be adopted throughout. While fully in agreement with furthering the general aim of increased efficiency and with the somewhat elusive goal of recipient "ownership", this is not wholly unproblematic analytically. Assume that by recipient is meant the government of an aid receiving country. How does one proceed when for example the assumption of a benevolent government does not hold, in full or in part? To put it more bluntly, what if the aid receiving government's main aim is in reality to enrich the ruling elite? Yet another potential problem in choosing the perspective of the aid receiving government in the medium- to long-term term is that this perspective may not be well defined. The government and both ideology and development strategy may have changed several times over. We acknowledge that an evaluation study should evidently relate to government targets, but we argue there is need to make sure that this is done with reference to a generally accepted broader set of welfare approaches and measures relevant to the poor or the whole population. What is commonly referred to as "donor driven" driven development and evaluation must in our view be avoided, on normative and on positive grounds. Yet, it should not be overlooked that identifying and defining the “correct” (or optimal) alternative is by no means straightforward. It is therefore in our view critically important to be as clear as possible about the actual set of criteria against which the evaluation exercise is undertaken, so independent replication becomes not only possible, but an established norm.

The third, partly overlapping question concerns how to evaluate the impact of aid in a scientifically defensible way. The issue of establishing an appropriate counterfactual is the key challenge here. Various approaches have been tried to deal with the problem of measuring “true” impact of aid. One is to compare targets with actual outcomes. The problem is that a target may be reached for reasons that have nothing to do with the provision of foreign aid. For example, if a given country is fortunate to reap a terms of trade gain that increases income, this does not mean that aid was effective, and vice versa. Another quite common approach is to rely on before/after comparisons, but also this approach suffers from the shortcoming that the analyst cannot attribute the effect to the treatment.

Thus, to be able to measure the effect of a treatment X (such as aid) of a person (group, country, ….), an evaluator must in principle be able to compare the value of a chosen indicator in two strictly independent situations: with and without treatment. Only in this way can the effect of X be calculated correctly by subtracting the outcome without X from the outcome with X. Yet, in reality a given person (group, country, ..) can only be observed in one of these two situations, and over time a lot of other circumstances may have changed, which influence the outcome. Accordingly, to establish a “true” measure of the impact of X, the importance of these circumstances needs to be accounted for. The evaluator must somehow try to find a way of establishing what would have happened if treatment X had not occurred, all else being equal, and why. Alternatively, if a group of selected people, groups or countries are compared at the same point in time (with and without treatment), the evaluator needs to account for the impact on the chosen indicator of the other differences that exist among the units of observation in reference. This is the fundamental evaluation problem, and we highlight that there are no ways in the social sciences of addressing this problem in a broadly acceptable way without making assumptions that are bound to be debatable, in theory and in practice.

On this background we review in Section 3 the approaches that have been attempted in the past to address this crucial challenge.

3. Approaches to the Analysis of the Impact of Aid

The past 50 years have witnessed a massive outpouring of studies on the impact of foreign aid. The topic has been a central and recurring theme with which both economists and specialists from other social sciences have grappled; and the question of whether aid works or not has been approached from different methodological and ideological perspectives. More specifically: (i) the impact of aid has been evaluated at both the micro- and macroeconomic level; (ii) cross-country as well as single-country case studies have been relied on; and (iii) aid effectiveness research includes broad surveys of a qualitative and inter-disciplinary nature as well as quantitative analyses. A comprehensive survey of the aid effectiveness literature is beyond the scope of this paper, but as a starting point for our discussion it is useful to review how the aid evaluation literature has evolved in general.

In the early years of the "modern" aid era in the 1950s and 1960s, foreign aid was mainly used to finance projects. Standard cost-benefit analysis (CBA) seemed, as discussed by Little and Mirrlees (1990), to be an appropriate tool for evaluation, and CBAs of stand-alone aid projects have by now been a major enterprise in the aid industry for more than fifty years. A mass of project evidence has been collected, and this includes, for example, that aid-financed projects in Africa often have high returns. The most rigorous project evaluations are done by the World Bank, and for the period 1993-2002 an average rate of return of 22% is referred to by the Economic Commission for Africa (ECA, 2005, p. 298). The ECA also highlights that projects have in recent years been characterised by improving sustainability and better institutional impact.

Turning next to what is typically referred in the evaluation literature as project impact evaluation, this approach focuses on the benefits caused by a project, e.g. changes in enrolment due to an education project. This type of evaluation can therefore be seen as a component of CBA, broadly speaking, but impact evaluation is regularly done in isolation. Moreover, in impact evaluation the analyst either relies on quasi-experimental techniques (if the treated were selected randomly) or, more commonly, statistical techniques are used (e.g. propensity score matching) to correct for differences, which are not due to the intervention, between the treated and a control group.

When it comes to evaluating "process projects" aiming at developing institutions, it seems clear that they are so far removed from the final development outcomes that it is hard to do quantitative analysis of impacts on final targets such as growth or poverty reduction. There is as yet no technique available that is broadly accepted by the profession which can effectively link institutional improvements to impacts.[5] The lack of commonly agreed techniques to measure such effects is a constraint on what we can propose, since a major requirement was that we should propose an approach that at least makes a first step in the direction of trying to deal with causal mechanisms.

So project evaluation continues (and should), but it seldom measures final impact, and even when impact is measured rigorously at the project level this does not provide a convincing answer to the macro aid effectiveness question. The total effect may differ from the sum of the parts, e.g. because aid has external (negative and/or positive) effects.

In the 1980s a shift took place in foreign aid towards relying more on adjustment lending as a leading aid modality. This led to asking whether adjustment aid and aid-induced policy changes work, in theory and practice. There were indications that this was not the case, and results seemed to depend on a whole range of country specific characteristics. These largely negative results were contrasted with the overwhelmingly positive results from CBA-analysis of projects. This led to a debate of the micro/macro paradox, identified by Mosley (1987), who suggested that aid seems to work at the micro but not at the macro level.[6] One issue used to explain the micro/macro paradox was that projects are affected by fungibility, which means that donors in effect finance other

---

[5] Some types of aid interventions are very hard to evaluate in quantitative terms. For example, it seems reasonable to argue that aid should be used to remove bottlenecks leading to aid absorption problems and inefficient administration. The analysis of the effects of such attempts on economic growth or poverty reduction remains controversial. Chauvet and Collier's (2004) have looked at the impact of aid on policy and institutional reform, but the analysis is tentative. Burnside and Dollar (2004) found evidence that institutional quality determines the effectiveness of aid. Yet, the possible endogeneity, not only of aid, but also of policies and institutions is addressed by Dalgaard, Hansen and Tarp (2004), who provide references to the debate on institutions and growth.

[6] See White (1992), Cassen *et al.* (1994) and Hansen and Tarp (2000) for further discussion and clarification.

projects than the aid project that receives the money. Fungibility seemed, it was argued, to undermine donor intentions, and in such a setting traditional CBA evaluation clearly failed to be convincing. It was also widely understood that simple “targets versus outcomes” and “before versus after” comparisons were lacking, so it became increasingly imperative to develop a well-specified counterfactual to sort out what the real effects of aid transfers are. In other words, the need to introduce a control group approach in aid evaluation based on modern econometric theory and new data became gradually better appreciated.

The dissatisfaction with the relevance of micro analyses of aid impacts as well as earlier cross country macro-work led on the above background to the emergence in the 1990s of an influential literature in which renewed attempts were made to analyse the aid-growth link at the macro-level. These studies broke novel ground in at least four areas. First, they made use of panel data for an increasing number of years and a large number of countries. Second, new growth theory inspired the analysis in distinct ways, providing a different analytical basis compared to previous work. Measures of economic policy and the institutional environment were included directly in the reduced form growth regressions alongside traditional macroeconomic variables. Third, endogeneity of aid and other variables was addressed explicitly in some studies. Finally, the aid-growth relationship was explicitly recognized as being non-linear. These were major methodological innovations, and they represented a distinct step forward in the empirical work on aid effectiveness.

One reason for the popularity in the past 10-15 years of the cross-country panel data approach is that it makes it possible to move beyond simplistic aid-growth correlation analysis, where the analysis of causal effects is indeed rather primitive. The analyst can in the modern aid-growth work control for the impact of a large range of variables, and in this way becomes able to move closer to the ideal of having a reliable counterfactual. It is therefore not really surprising that the cross-country aid-growth literature gradually became dominating from the mid-1990s in attempts at identifying conditional relationships between aid and growth.

Burnside and Dollar (2000) argues that aid is more effective in “good” policy countries, while its growth impact is negligible in “bad” policy environments, findings that were at the core of the influential "Assessing Aid" study by the World Bank (1998). The policy recommendation was that aid should be concentrated in good environments, rather than being allocated based on needs. There is however difficult modelling and econometric issues involved in the estimation of these effects, and Hansen and Tarp (2001) criticised some of the choices made by Burnside and Dollar and found in their own analysis that aid may have a positive impact on growth even in less favourable policy environments.[7] Also Easterly, Levine, and Roodman (2004) have concluded that the results in Burnside and Dollar are lacking in robustness, and Roodman (2004) can be consulted for a comparative analytical overview.

There is now a certain degree of agreement that aid has on average had a positive growth impact, but worries about constraints on absorptive capacity and diminishing returns, which mean that the positive effect of aid on growth peters out once its size exceeds a certain proportion of GDP, remains; and it has also been pointed out that Dutch Disease effects may hinder export growth. McGillivray (2005) found in his survey of the literature that various studies place the diminishing returns in the range of 15-45% of GDP.

So the aid-growth cross-country regression literature did devise a way to deal with the counterfactual issue, but it suffers from the problem that it assumes that all countries have the same underlying structure, once the impact of the explanatory variables has been accounted for.[8] Another problem is that the aid processes that mediated the outcomes were not dealt with explicitly. They were left in the black box. It is therefore hard to extract lessons about donor practices from the results of this literature.

Another strand of the aid-impact literature makes use of Computable General Equilibrium (CGE) models, where the analyst can effectively hold everything else but a specific policy intervention X constant and then simulate what impact X has. These kinds of models have the advantage that they have an inbuilt counterfactual and take indirect economy-wide effects into account. However, they also suffer from limitations of their

---

[7] Reviews of the debate are provided by Kanbur (2003), Collier and Dollar (2004), Dalgaard, Hansen, and Tarp (2004) and Roodman (2004). Recently, Clemens, Radelet, and Bhavnani (2004) addressed the issue by defining "short impact aid" to distinguish aid for which a growth response may be expected. A recent paper by Dalgaard and Hansen (2005) reports high returns to foreign aid in a cross-country analysis comparable to the micro level evidence referred to above.

[8] See Solow (2001) for a thought provoking and very critical reference.

own if the aim is to address the question what aid inflows do for growth. The CGE models typically optimize within given constraints, but aid inflows lift these constraints and thereby tend to lead to higher output. Different results may emerge if distortions are incorporated in the model, but the CGE models are in any case stylized versions of the world that cannot take all relevant aspects into account. Thus, it is hardly a credible proof in practice that "aid works" if it can be shown within a CGE model that aid leads to higher output. Instead, these models may be more useful in testing different types of policies and comparing outcomes, but this approach is hardly what is called for in our case.

We have noted that the analyst needs a well-defined counterfactual to get quality results, and new techniques have emerged to identify credible control groups such as impact randomization and propensity score matching (Ravallion, 2001). The problem with these techniques is that they are appropriate only if the impact of aid can be quantified, e.g. as an increase in enrolment of children from a clearly defined group. This condition is often not satisfied for policy reforms or programme aid, technical assistance or institutional development. At present there are attempts under way to extend these techniques to (at least a partial) evaluation of sector aid or general budget support. We will build on those in our proposal below.

Econometric time-series country studies of individual countries have not been used extensively in the recent aid evaluation literature. To do this there is need for long time series data to run unit root tests, trying different specifications and estimation techniques. The risk of getting unusable or insignificant results is at present very high, the basic problem with this type of country specific studies being constraints on the degrees of freedom and difficulties in finding cointegrating relationships. It might be worthwhile to test whether such analyses can produce usable results, but we are a priori doubtful, so this is not what we are aiming for at present.

More comprehensive, but less formal, country evaluation studies have also been carried out over the years. For example, a set of studies was done by SASDA in the 1990s (Bigsten *et al.* 1994a, 1994b, on Tanzania and Zambia) and another set of country studies was carried out by White (1994) on the macroeconomic impacts of aid. Other studies have looked at the impact of structural adjustment, here meaning policy impacts in addition to the amount of aid and how it is spent. White (1999) reports on a series of studies, which were done to evaluate the impact of programme aid funds. They tried to trace how programme aid affected macroeconomic aggregates such as imports and government spending, and from this an effort was made to infer the impact on growth. It is generally acknowledged it is very hard to establish an appropriate counterfactual at the country level. For example, with regard to the analysis of the impact of funds and policy White (1999, p. 93) states: "…econometric estimation of meaningful economic relationships at the required level of disaggregation is virtually impossible". What is applied is an "ad hoc approach which starts by identifying aid in a consistent set of internal (government finance) and external accounts (balance of payments). From that starting point a "no aid" counterfactual set of accounts is constructed." From this set of data one then attempts to get some idea about aid impacts. These impacts refer to factors that are taken to be important for growth, but one cannot identify the direct impact on growth or poverty reduction. While clearly of interest to the evaluation literature these types of studies all suffer from the problem of leaving unaddressed the challenge of clearly defining a counterfactual.

Still, White (2005) provides two arguments for country case studies.[9] The first is that although they abandon being representative they allow for more for depth. A case study can go deeper into a variety of issues that cannot be addressed effectively in a desk-based portfolio review. The second argument is that the approach used by for example the OED of the World Bank is not based on case studies alone. Instead it combines a number of approaches, and as such is a mixed approach. This kind of approach is probably what Edgren *et al.* had in mind. We find such studies interesting, but are concerned whether they will produce novel results, and results that measure aid impacts with enough (causal) precision to have decisive impact in the aid evaluation debate. Still, in the next section we will look at some interesting and ambitious comprehensive studies to see what can be learnt.

[9] White (2005) also argues that there is scope for meta studies, which aggregate performance from different studies. He notes the such analysis involves six steps, namely: (1) define the problem; (2) collect the relevant studies to be reviewed; (3) screen the studies for quality; (4) standardize the indicators and code them; (5) produce a synthesis measure; and (6) present results.

4. Comprehensive Evaluations

Comprehensive in-house evaluations are carried out by all individual donor agencies, and the amount of this kind of work is massive. It is therefore beyond the scope of this paper to try to summarize it here. Instead we will focus on two types of evaluations which contain elements that seem particularly relevant in our search for methods for evaluation of the impact of total ODA, namely World Bank Country Assessments, and the ongoing multi-donor evaluation of general budget support. We highlight that the choice of these two examples, which are "donor driven", and the way in which they are described, is not meant to ignore issues of who should ideally "drive" the evaluations to make them as pertinent and insightful as possible. This is as clearly illustrated by the comments made by participants at the Stockholm seminar on 10-11 November 2005. It is a difficult topic on which there is at present no consensus. We merely use the two examples below as illustrations of the kinds of concerns, which any comprehensive evaluation must address one way or the other.

The World Bank Country assessments are comprehensive, but they do not evaluate the recipient's overall development process. They start by defining a set of objectives, typically a sub-set of the Client's development objectives. Then it is assumed that the outcome of the assistance programme is determined by the *joint* impact of four agents: (a) the Client; (b) the Bank; (c) partners and other stakeholders; and (d) exogenous forces (e.g., events of nature, international economic shocks, etc.). The Bank's evaluations have three dimensions (triangulation) in their evaluation to check for consistency in the results. They are "(a) a *Products and Services Dimension*, involving a "bottom-up" analysis of major programme inputs-loans and aid coordination; (b) a *Development Impact Dimension*, involving a "top-down" analysis of the principal programme objectives for relevance, efficacy, outcome, sustainability, and institutional impact; and (c) an *Attribution Dimension*, in which the evaluator assigns responsibility for the programme outcome to the four categories of actors." In evaluating the outcome (expected development impact) of an assistance programme, the OED gauges the extent to which major strategic objectives were relevant and achieved. OED utilizes six rating categories for outcome, ranging from highly satisfactory to highly unsatisfactory. It also considers some dimensions of institutional development impact and sustainability.

This type of analysis is a sort of mixed approach as discussed above, and this could be one possible model from which to start developing a framework for the evaluation of total ODA. Yet, the reservations we had above still apply.

Apart from the World Bank Country assessments, a consortium of donors is in the process of completing one of the most ambitious evaluations that have so far been attempted. This is the evaluation of the general budget support (GBS) programmes.[10] The purpose of the evaluation is according to its Terms of Reference to "evaluate to what extent, and under what circumstances (in what country contexts) GBS is relevant, efficient and effective of achieving sustainable impact on poverty reduction and growth". The aim is thus to establish a link from the aid to the impacts. It is therefore in the present context relevant to uncover how the analysis was one in the total of seven country cases studied.

The study uses an evaluation methodology (Lawson and Booth, 2004) which was especially developed for the study. In accordance with the assessment towards which we are inclined, the framework paper acknowledges that it is generally not possible to identify the contribution of individual donors. The framework applies the standard DAC criteria of relevance, effectiveness, efficiency, impacts, and sustainability. A logical framework is used to analyse the sequence of effects from inputs to immediate effects (activities) outputs, outcomes, and finally impacts. So this study attempts to deal with the whole chain from inputs of aid resources to ultimate impacts. GBS is considered to have three types of effects, namely flow-of-funds effects (macroeconomic effects, budgetary effects), institutional effects (changes in ownership, planning and budgetary processes), and policy effects (changes in macro-policies, sector policies, and cross-cutting policies). The evaluators aim to disaggregate the GBS inputs as well as the poverty impacts. The interrelationships involved are shown in intricate causality maps.

The studies first describe the aid context and the evolution of the GBS in the countries concerned. Then the evaluators discuss a set of problem areas, including the relevance and sustainability of GBS in addition to the effects on:

- Harmonisation and alignment.

[10] In what follows, the abbreviation PGBS, which is sometimes used, has consistently been shortened to GBS.

- Performance of public expenditures.
- Planning and budget systems.
- Policies and policy processes.
- Macroeconomic performances.
- Delivery of public services.
- Poverty reduction.

They, finally, provide a set of recommendations for the individual country.

Under each heading a set of relevant facts is presented first. This is followed by assessments against predetermined evaluation criteria, and in each of these assessments information from various secondary sources and informants was used. Answers are ranked with regard to the strength of the effects and the degree of confidence in them. Finally, there is a discussion of principal causality chains and counterfactuals.

We have above drawn attention to the counterfactual problem, which we see as both critical and fundamental in any evaluation exercise. Also the budget support study framework discusses at length the counterfactual problem, and it is acknowledged that it is hard to address effectively. What is done in the actual country analyses is typically to assume a continuation of older structural adjustment type of programmes or assume concentration on project aid as counterfactual scenarios. The counterfactual results are largely based on broad assessments by knowledgeable people, and not on quantitative estimates of impacts. It was very hard in the study to directly attribute observed changes to the GBS. It is argued that the most important impact to uncover would ideally have been on poverty reduction, but this is at the same the area in which it is hardest to separate causal effects out and attribute output changes to GBS or other factors. To formulate a counterfactual for the impact on targets like empowerment is also very difficult.

Like in the World Bank assessments, the GBS team refers to its analytical approach as triangulation. Varied data sources are used to identify effects. The time span covered is three years. Overall, the key problem with this set of evaluations has been to define a plausible counterfactual, and the analysis of the counterfactual was restricted to considering the plausibility of well-constructed case histories. It seems clear that the evaluators cannot say anything in quantitative terms as to how the GBS affected poverty.

Accordingly, the preliminary synthesis report acknowledges that "we cannot confidently track distinct (separately identifiable) PGBS effects to the poverty impact level in most countries" (IDD et al., 2006, p. 9).

The evaluators also discuss other problems in relation to the measurement of impact. One question raised with regard to institutional impact is whether the GBS was introduced because the country had set up a reasonably good system or the system was set up because of the GBS. It is argued that it is usually possible to make a reasoned judgement about what GBS financed, but we find it is hard to argue that the GBS study really manages to deal effectively with the counterfactual issue and the fungibility problem.

Again, we would like to reiterate that we do not conclude that this analysis is uninteresting. In fact, this is an important contribution to the work of donor agencies in the evaluation area. A lot of useful insights are generated in many dimensions and this should not be overlooked within the broad set of aims pursued by carrying out aid monitoring and evaluation. All we want to bring across is that no really "hard" quantitative results on impact emerge from the case studies. The technique for measuring the link from GBS to impacts is analytically weak, and the link from total ODA to final impacts is tenuous.

5. Motivation of the Proposed Study

The original idea behind the proposal to do an analysis of the impact of total ODA was, as we understand it, to do an analysis at the country level which would look (i) at all aid inputs in its many different forms and origins and (ii) analyse the process through which these aid inputs operate, all the way to final impacts. Against the background of our discussion above of available research methodologies we find that such a full-scale approach is somewhat overambitious, if the identification of the causal aid-impact chain is assigned the key role to which we have also alluded. In addition, we believe it would generally be hard to analyse or maybe better identify convincingly the impacts of individual donors; and even if this were feasible this is arguably not central to the stated objectives. We understand these as being to analyse what works, rather than to find out who was responsible.

Taking a general view at how aid may affect outcomes there are several ways of looking at this and various mechanisms on which the analyst can focus. (a) Aid provides financing. This continues of course to be an important potential channel of influence of aid on growth and poverty reduction and it was at the very core of the original rationale for aid. This area is however by now well-trodden ground. Much has been done in terms of understanding what aid can and cannot do in this perspective. (b) Aid also interacts with recipient country policies as well as with policy implementation and institutions. Themes that have featured in the debate in this regard are for example the effects of attempts at institution building, the impact on institutions of donor proliferation (Bigsten, 2005), and the risks identified by Platteau and Gaspart (2003) of elite capture. We believe that the highest return in this area of investigation would be from concentrating on the different ways in which aid and policy may interact and on how aid interplays with policy implementation. Consequently, this will be one of the themes we suggest for the evaluation (Section 6). Secondly, the ultimate aim of aid is its impact on welfare in its various dimensions. We have noted that there are distinct methodological limits on what can be done, but we also acknowledge that this is an area where much needs to be done. We therefore propose a second stage in the analysis, where an attempt is made to apply statistical impact evaluation techniques in a more comprehensive fashion than what has been done before (Section 7). This will not cover all aspects, and issues of causality keep looming in the background, but it may extend the use of available statistical techniques beyond what has so far been done in comprehensive evaluations. This would be a challenging undertaking, and one that could potentially provide novel evidence on the impact of aid.

The proposed evaluation, unlike most other evaluation studies including the GBS study, will take a long-run perspective. Many of the effects assumed to be caused by aid will only emerge over a long period of time. Another extension relative to earlier work, such as the GBS evaluation, is that we open up for the evaluation of the impact of total ODA, including all types of donor-recipient relationships.

6. The Effect of Aid on Policies

Whether and to what extent donors influence economic policies in aid recipient countries and how donors and recipients interact remain open questions. Kanbur (2000) provides a telling overview of the difficulties of pursuing policy change through donor conditionality, and the issue appeared to be settled when Burnside and Dollar (2000) showed econometrically that there was *no* effect of aid on policies. This conclusion quickly entered into conventional wisdom, but the robustness of the Burnside and Dollar results has since been questioned by many commentators. The critique has reopened the possibility that donors do in various ways manage to affect policies in a way considered desirable by donors.

Such influence need not be tied directly to aid in the financial sense of the word. Indeed, many studies of conditionality, *e.g.* the *Aid and Reform in Africa* study edited by Devarajan *et al.* (2001), have argued that in cases where donors were influential it was typically by supporting reformers in the early phases of domestic policy debates. In those cases aid mattered, but the contribution was intellectual rather than financial. This is likely to bias the statistical analyses of conditionality: inevitably such studies use (net) disbursements as the appropriate measure of aid. They may thereby miss part (possibly the most important part) of the impact of donors on policies.

It might seem that the issue could be easily settled by investigating whether in the period considered policies were indeed changed in the direction advocated by donors.[11] However, for aid to have influenced policies this is only a necessary condition. It is by no means a sufficient one, as many other changes have been going on. This is best seen by noting that if a donor tries to affect policy reform there are in principle four conceivable outcomes:[12]

- The reforms were (independent of their intrinsic characteristics) not adopted, so donor efforts have obviously failed.

[11] We highlight that in making this statement we do not take sides in the ongoing debate about whether polices pursued through conditionality were indeed the ”correct” ones in order to promote development or not. It is however a fact that aid interacts with policy formulation and implementation and we believe it is important to uncover this interaction more completely than done before.
[12] This section draws on Gunning (2000, 2001).

- The reforms *were* adopted but this would also have happened without aid. In this case donor efforts had in reality no effect because the government would have adopted the policy changes desired by donors anyway.
- The reforms were adopted and would not have been adopted in the counterfactual case (without aid or donor pressure), but the effect is only temporary: the reforms are ultimately reversed. This is the outcome predicted by time inconsistency models of conditionality: the recipient government has no incentive to maintain the reforms when the aid runs out.
- The reforms are adopted and sustained and this would not have happened without aid.

Clearly, the donor is only truly effective in the last case in affecting policy change in the direction desired by the donor. Aid achieves nothing in this dimension in the first and the second case (although only in the first case is this obviously so), while in the third case aid only impacts temporarily. Regression analysis runs into an obvious limitation here: it can establish whether aid flows are accompanied by policy reforms, and it may be able to establish whether reforms are reversed. But this approach cannot convincingly distinguish between situations where aid was instrumental or only incidental to policy reforms. Regression evidence of a link between aid and policy at best shows that a necessary condition for effective conditionality is satisfied.

Case studies have sometimes attempted to overcome this problem by using in-depth, country-specific information to construct a credible counterfactual. However, here too there is a methodological problem: donors have an incentive to exaggerate their influence over policy changes while the government typically (but not always) has the opposite tendency. Hence case studies may be biased in either direction, depending on which policy accounts they rely on.

Many governments simply do not adopt the reforms favoured by donors (the first possible outcome suggested above). For example, in the final phase of the Kaunda regime in Zambia, the government reached an agreement with the IMF on a programme "designed to create a diversified and market-oriented economy" (internal IMF document, as quoted by Botchwey *et al.*, 1998, p. 95). The Kaunda government had no intention of creating such an economy. Indeed, price controls were formally abolished but continued to be enforced, and efforts to accord a larger role to the private sector were resisted effectively (Botchwey *et al.* 1998). More recently, again in Zambia, the government long resisted donor pressure to privatize copper mines. There are also many examples of aid, far from inducing reforms, actually delaying them. Devarajan *et al.* (2000), document this for the Democratic Republic of Congo, Kenya, Nigeria, and Tanzania.

There is also evidence of reforms being adopted but not because of donor pressure (the second possible outcome suggested above). In Vietnam donors made an important contribution by informing the policy debate but were apparently virtually powerless to force through reforms against the wishes of the authorities. As van Donge, White, and Nghia (1999, p. 33) stress: "The pace and direction of reform is determined by Vietnamese politics."

The case studies literature also documents many cases where reforms were "owned" by the government, so donor pressure was not essential. At one extreme, reforms were even undertaken without donor involvement, as was the case in the early phases of liberalization in Burkina Faso, Eritrea, Mozambique, Uganda, and Vietnam. In Uganda key reforms (abolishing price controls, liberalizing the foreign exchange market, privatization) were all undertaken on government initiative. While donors, understandably, like to take credit for the success of reforms in Uganda, there have been very few cases of substantial disagreement on policy issues between donors and the government (Ddumba-Ssentamu and Dijkstra 1999, pp. 91–92). Kasekende and Atingi-Ego (1999) argue convincingly that Uganda's decisions in the past cannot be attributed to donor pressure, and much of the recent developments in Uganda would seem to reinforce this view.

This is not to say that the involvement of donors in these countries was not important and/or potentially helpful; it often was. But the contribution of donors was to help build the case for reform through policy dialogue rather than to buy reform with aid. In addition, donor support eased the implementation of reform programmes.

The third possible outcome, involving policy reversals, has been especially common in Africa. "Programme interruptions" (in most cases involving policy reversals) were the main concern in the IMF internal evaluation of its Extended Structural Adjustment Facility (IMF 1997). A well-known case is the liberalization of maize marketing in Kenya, a reform that was repeatedly undertaken, with each attempt ending in a policy reversal. Oyejide *et al.* (1999) document that trade liberalizations have been reversed in seven out of 10 African countries, in many cases (including Kenya) more than once.

In the fourth possible outcome conditionality is effective in the sense that donor pressure was essential and the reform was sustained. This case seems fairly rare. In an exhaustive evaluation of Swedish programme aid one of the few examples is the liberalization of Mozambique's cashew market—a reform that donors effected despite strong government objections (White 1999; Gunning 2000, McMillan, Roderick, and Welch 2002).

When analyzing the impact of aid on policy we should not be confined to looking at policy formulation but also consider the policy implementation aspect. We cannot confine the analysis to considering whether policies are put on the books or not, but we must also investigate how they are implemented and affect public-sector management. Donors affect recipient governments via the transfer of resources, policy dialogue, conditionality, and the like. Aid can facilitate implementation by releasing the government from revenue constraints, enabling it to strengthen domestic institutions by paying higher salaries to civil servants. It can also provide technical assistance and training and be used to build legal systems and accounting offices. The character of the aid relationships will influence the extent to which policies are successfully implemented.

Bigsten (2005) reviews the available evidence there is about how the aspect of donor coordination affects outcomes and concludes that it is hard to isolate the effect of donor coordination policy management. It is difficult to estimate the contribution of one factor in a very complex institutional set-up, especially within a single country, which is what we are considering in the proposed total ODA-study.[13] The common approach to the problem of identifying a counterfactual is cross-country regressions. Knack and Rahman (2004) did an empirical cross-country analysis of the impact of donors on recipient administrative quality. The econometric results support their hypothesis that aid

[13] Johnson and Subramanian (2005) say that "there are considerable uncertainties in our knowledge about how institutional change can be promoted, if at all" and Kraay (2005) notes that we know relatively little about how to tackle the governance issues in Africa.

undermined the quality of government bureaucracy more severely in recipient countries where aid was fragmented among many donors. This country-study cannot use such an approach, but will have to find other ways.

The study we propose will use a case studies approach. The alternative of the statistical methodology of impact evaluation is more rigorous, but is not realistic in the present context. When the supposed effects of aid are economy-wide it is clearly impossible to construct a strict counterfactual: there is no control group left. In the absence of a statistical methodology the evaluation depends critically on the quality of the informal methodology used for constructing a counterfactual. We propose that political scientists play an important role in establishing what might have happened in the absence of aid. In the case studies literature the basis for the counterfactual is often not more than a series of assertions. This is clearly unsatisfactory. The case studies will have to make extensive use of policy documents and interviews with key participants to build a convincing counterfactual for the policy and policy implementation analysis. This would be one of the two distinguishing characteristics of the study.

The second characteristic is the use of the 4-class taxonomy. Whereas much of the literature focuses on the rather crude question "does aid work?" the country authors will have to place their study within this taxonomy. This will give a much richer view of the way aid works (or does not) than has been common in the aid effectiveness literature. Finally, inspiration in preparing the study could be derived from a recent study where conditionality is revised, including a review of concepts, experiences and lessons (see Koeberle et al. 2005).

7. Statistical Impact Evaluation

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

When used in the second way impact evaluation involves a formal comparison of results for a “treatment group” and a “control group”; the difference between these results is attributed to the intervention. There is in this case no presumption that the results being evaluated are final (“impact”) rather than intermediate. For example, the analysis could focus on school enrolment, an intermediate result, rather than literacy. To avoid confusion we will use the term statistical impact evaluation for this case, and we note that if the focus is on ultimate results then the evaluation is also impact evaluation in the first meaning of the term.

Ideally, one would compare results for the same group with and without “treatment”. However, no group can as already discussed above be observed at the same time in both situations. This is the fundamental *evaluation problem* referred to in Section 2. 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 where the “treatment group” would in fact have received no treatment. Rather than comparing the same group with and without treatment at the same time (which is desirable but impossible) the evaluator compares results for two different groups.[14]

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

[14] The hypothetical nature of the counterfactual is sometimes used as an argument against statistical impact evaluation. This is not well taken; as Mrs. Thatcher used to say: there is no alternative. The suggestion (implicit in this critique) that one can do without such a hypothetical construct ignores the evaluation problem.

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

In development the use of such evaluation methods is more recent, but the last decade has seen numerous applications in evaluations of social safety nets, schooling programmes targeted at the poor, health interventions and even rural empowerment programmes.[15] As in the case of labour market evaluations, work in this area has moved from its initial research focus to practical applications. One of the most famous papers in this field describes an evaluation of primary schooling activities in Kenya and this was initiated by a small NGO. NGOs and bilateral and multilateral donor agencies as well as aid receiving countries are beginning to experiment with such methods.

A central issue in such evaluations is the availability of baseline data. When such data are available one can address the fundamental problem of unobserved differences between the treatment and control groups. Rather than measuring differences at time t (after "treatment") between the two groups one can measure for both groups changes over time (differences between the situation at time t and an the time of the baseline observation). Impact can then be assessed as the difference between the two groups in those changes over time ("differences in differences" or "double differencing"). While policy makers are understandably reluctant to invest in the collection of baseline data there is a growing awareness that without such investment prior to a policy intervention assessing the effect achieved will be very difficult. Similarly, policy makers are increasingly aware 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)

[15] An excellent (and very entertaining) introduction is Ravallion (2001).

there is a strong case for using random assignment. Note that since intervention is gradual in any case, the usual moral objection to randomisation does not apply: if one is not going to extend the treatment to the entire target group instantaneously anyway one may as well assign the initial beneficiaries randomly.

Statistical impact evaluation presupposes that both the treatment and its possible effects are well defined. For example, the treatment might be a project offering cash transfers to poor households conditional on the (continued) school enrolment of their children. Given the project's objective its impact is then obviously to be measured in terms of enrolment of children in the target group. Clearly, development interventions usually 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, increasingly donors are moving from project aid to sector support or general budget support. This shifts the evaluation question to a much higher level of aggregation, a level for which the techniques of statistical impact evaluation have not been designed. This is particularly relevant in the present context since the DAC concern is to establish the impact of *total* aid.

This raises the question whether statistical impact evaluation can be employed at a higher aggregation level than that of an individual project. The key issue here is the enormous heterogeneity of the activities undertaken by a recipient government with donor support. One approach is to measure the impact of aid through cross-country growth regressions. Inter-country variance is then used to estimate the impact (in terms of changes in poverty, income or economic growth) of total aid (or its various components). Implicitly, the experience of other countries is then used to construct a counterfactual whereby one controls as much as possible for inter-country differences other than those in aid receipts.

This is an active (and somewhat controversial) area of research. Results are far from settled; indeed much of the work in this area fails to pass tests of robustness.

The alternative is to apply statistical impact evaluation but in such a way that conclusions can be drawn at a higher level of aggregation than that of the individual project. This is the approach we suggest be pursued for the purposes of the evaluation activity in reference.

It should be emphasized that this is largely virgin territory. While the methodology for statistical impact evaluation at the project level is reasonably well established there have yet been no attempts to aggregate the results. The Dutch evaluation agency IOB is initiating a number of feasibility studies at the sector level. The current proposal is an extension of that initiative. The key idea is, first, to select a sample of activities representative for the aid-supported activities to be evaluated; secondly, to apply statistical impact evaluation to each of the sample activities; and, thirdly, to aggregate the results to arrive at an assessment of impact at the aggregate level.

The first step, drawing a representative sample, is in itself a major exercise. It requires a detailed description of the aid-supported activities in the period considered. One of the IOB feasibility studies focuses on Dutch support for the education sector in Zambia. This will involve a detailed desk study to establish what Dutch aid in the education sector was used for.

It should be noted that a key assumption underpinning this approach is that the effects of individual activities are additive. This is, obviously, a strong assumption since it rules out non-linearities and in particular interaction effects and these are likely to be important in practice. For example, preventive health measures (such as vaccination programmes) are typically highly non-linear, with sharply declining marginal returns when coverage is extended to a larger a larger part of the population. Similarly, the extent to which farm households can benefit from price increases for the crops they sell depends on the extent of market integration and transport costs. As a result, the impact of pricing policies on rural poverty will depend on infrastructure and competition policies so that there are interaction effects. There is no inherent reason why non-linearities and interaction effects cannot be accommodated but there is a practical limitation: the sample size would have to increase. We would therefore propose to start with the simplest (linear) specification. That this involves an important limitation should be kept in mind.[16]

[16] This limitation is not specific to what we propose: it applies to most evaluation methods. An exception is the cross-country regression methodology which - by aggregating over different activities - in principle allows for interaction effects between them.

This first step involves a choice of aggregation level. The logic of the focus on total aid of the DAC initiative would seem to imply that the appropriate level is that of the country. Whether this is indeed appropriate is an important topic for discussion at the Paris meeting. In many cases (e.g. when aid is given as general budget support) evaluating all aid amounts to evaluating the impact of all government activities in the country concerned. To establish such impact with any confidence clearly would require a gigantic effort. One might prefer a series of sector-level evaluations (extending the work initiated by IOB) in a number of different countries over an economy-wide (multi-sector) evaluation (which would have the advantage of covering total aid but might have to remain limited to a single country because of resource constraints). This is a trade-off which needs to be considered at the Paris meeting.

The second step is to apply statistical impact evaluation to each of the activities in the sample. Here there two key questions. First, one needs to consider whether the way the activity has been implemented allows one to choose a convincing control group. As noted above, this is straightforward when randomisation has been used (either intentionally or by accident, the case of “natural experiments”), but this is rare. More commonly, the programme has been implemented sequentially or partially so that treatment and control groups can be identified. However, non-randomness implies that the two groups may differ systematically in other ways than in having received treatment or not. Whether one can adequately control for this (with methods such as propensity score matching) depends on the answer to the second question, that of data availability.

Where baseline data have been collected these may need to be complemented with new (post-intervention) data. In some cases it may be possible to use existing survey data if these allow the identification of treatment and control groups. An evaluation would start with a pilot phase focusing on data availability issues so as to establish the feasibility of statistical impact evaluation. The exercise might have to be aborted after this phase, if it turned out that an evaluation was impossible without massive data collection efforts.

Project-level statistical impact evaluation typically focuses on the question whether the project has had a (statistically significant) impact. To make the results of different project evaluations commensurate they have to be comparable both in terms of inputs and of outputs. The former implies that costs must be measured, the second that the evaluators must attempt a sensible aggregation over possibly quite heterogeneous impact measures such as poverty or malnutrition.

Given the results of the second step, the final step, aggregating the results, is relatively straightforward. This reflects the linearity assumption discussed previously. The end result is a statement of the form "public spending (possibly limited to particular sectors) in country X reduced poverty by so much". It is important to note that this statement is not aid-specific. Under general budget (or sector approach) the contribution of aid to this impact is simply given by its share in total public expenditure.

A corollary is that the information generated by such an evaluation should be of great interest to the government of the recipient country. After all, if the sample is truly representative what is evaluated is effectiveness of the whole set of government activities. This should in our view increase the possibility to ensure real collaboration and avoid the perception that the evaluation is "donor driven". In line with this we propose that rather than donors selecting countries as case studies, recipient countries should up front indicate whether they are interested in commissioning such an evaluation. The final selection of case studies from the group of countries expressing such interest would be made in such a way as to make the sample as representative as possible for total aid. In effect, recipient countries would be invited to sign up for a statistical impact evaluation of their (aid supported) development activities.

Statistical impact evaluation has strengths and weaknesses. Its main strength is its use of a rigorously constructed counterfactual and the statistical testing of many of the underlying assumptions. Traditional impact evaluation methods can be equally "quantitative", but evaluators then apply lower standards when deciding what qualifies as counterfactual. For example, quite often the (implicit) counterfactual is the treatment group prior to the treatment. Clearly, observed changes over time can be due to many factors other than the treatment considered. This is especially pertinent when long periods are considered. If one does not control for, say, changes in the country's external environment then attribution becomes highly problematic. Applying statistical impact evaluation to a whole sample of activities rather than evaluating a single activity has an additional advantage: it will reveal differences in returns between various government activities. For example, some types of schooling programmes may turn out to be much more effective than others. 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.

There are also disadvantages. First, precisely because the approach attempts to correct for all factors which might have influenced the observed outcomes, it is data intensive. However, in some cases the necessary data have already been collected for other purposes, e.g. for poverty assessments. Secondly, the evaluation establishes whether (and to what extent) an intervention was effective, but not why. This will have to be investigated separately, with different methodologies. In terms of the drugs testing analogy, it will indicate that the drug is effective, but it does not identify the active ingredient. This must be brought out separately, but we believe careful reflection based on the above results could be added, which would help bring the aid debate forward.

The fungibility critique implied a fundamental criticism of evaluation at the project level. This was one of the factors that led to increased programme lending at the sector of national level. Donor agencies are well aware that at its logical extreme (when an individual donor has no influence over the activities undertaken by the recipient government other than by relaxing the budget constraint) budget support implies that an evaluation can no longer be donor-specific. This calls for collective action. Donor collaboration in evaluation, as envisaged by the DAC, therefore makes eminent sense.

8. Structure of the Proposed Study

Summing-up, we propose a study of the impact of total ODA with two parts. As the first part we propose an analysis of policy choice and implementation, and as the second part a broad-based application of statistical impact evaluation. The steps in each case study would be as follows:

Part 1

1) The study starts with a brief description of major events in the period considered (as long a time period as possible, preferably all the way back to

the introduction of aid into the country) and of the key outcomes, such as changes in poverty.

2) The study should include a discussion of policy changes, how policies were decided, the nature and extent of accountability, and how policies actually implemented compared to stated objectives. This will be largely descriptive.

3) An assessment of the extent to which donors and policy makers agreed or disagreed (by period) is also required.

4) When donors and government are in full agreement (and act accordingly) the role of aid is essentially financing (not trivial: good policy countries have imperfect access to capital markets). The impact of such aid can still be low, e.g. because of cumbersome donor procedures.

5) If there is not full agreement the study has to indicate where the case fits in the taxonomy of the previous section.

There are no simple rules for how to build the counterfactual for this part of the analysis, but as pointed out above the researchers carrying out the study would have to rely on policy documents, donor files and interviews with key participants. This can be useful only if it is based on a careful analysis of the ideology, perceptions and incentives (and changes therein) of donors and recipients.

If the case study can argue convincingly which of these four cases applies it thereby answers the question: was aid effective in changing policies and implementation? It should be noted that this does not establish the impact on welfare: the policies may have been bad and we do not know precisely how growth translates into poverty reduction etc.

The case studies should recognize that the way aid affects policies and implementation depends on the full set of aid modalities. For example, channelling aid through a multitude of projects (as opposed to for example budget support) may put too many demands on poorly staffed institutions, thereby undermining policy reform, implementation or both.

Part 2:

1) First, a representative sample of government activities is drawn to be evaluated. The selection of these activities is a major exercise and requires a careful investigation into the availability of useable data.

2) Application of statistical impact evaluation techniques to all the activities in the sample.

3) Aggregation, discussion and reasoned extension of the results with a view to generalization and identification as much as possible of potential causal effects that have been at work.

9. Concluding Remarks

It is a widely agreed aim in the aid evaluation literature and practice that it is desirable to come up with sensible and academically sound and convincing responses to questions such as:

- Does aid work?
- Does aid help promote growth?
- Does aid reduce poverty?
- Is aid reducing inequality (in its many dimensions)?
- Does aid delay or promote economic reform?
- Etc.

There is at present no simple way of coming up with clear cut answers to these questions, which are going to be broadly acceptable, and this is so whether the analyst operates within a cross country or a single case study context.

First, it is extremely difficult to evaluate aid across countries and in specific contexts. This is related to (at least) three sets of interlinked and highly complex – and as yet still contentious issues: (i) growth and poverty in developing countries, (ii) governments and institutions in aid receiving countries, and (iii) institutional characteristics and (real) motives of aid donors. Professional people differ in their reading and understanding of these relationships and adjacent policy implications. Accordingly, evaluations of aid impact and how they are interpreted more broadly are likely to differ as well.

Second, ideological preferences are particularly strong in the aid area. Think for a moment about all the differences and discussions that exist about economic policy in developed countries. These differences are no less significant when aid to poor countries is in focus.

So, in real life (including the present situation) choices have to be made between (i) the academic approach, which can/will invariably be "shot down" as "unrealistic", "too complicated", "black box" etc. due to the simplifying assumptions that are necessarily required; or (ii) the generalist approach, which can on its side always be "shot down" due to the issue of "the counterfactual/causality", and its non-scientific methods. We honestly believe there is no easy answer here. In practice, one or the other approach is chosen in aid evaluations, depending on circumstances. Choices vary over time and sometimes a balancing act between the two is pursued. Sometimes real insight emerge in the process and the studies pursued come up with something new and convincing based on the strictest scientific principles; but many times so-called policy-influence of research has little to do with academic rigorousness and the way in which the analysis is carried out. At times, the evaluator is simply fortunate to come up with a message that fits well with what "people like to hear" or happen to "believe" is right (at a particular point in time).

In sum, it needs to be made clear up-front whether the desire to have a study of aid and aid impact in a case is broadly meant to produce a study that is "a really good and intuitively stimulating full dimensional story", from which something can hopefully be learnt one way or the other in the aid community; or what is desired is "a more cautious scientific piece of work", which may be less stimulating to the broad community than to a more narrow group of specialists. These two approaches often go hand-in-hand; but there is no way of promising this up-front. This is in the very nature of the research process. So, a strategic choice is needed.

We have suggested in this paper to take an approach that is admittedly somewhat narrower than what was originally proposed by Edgren (2004). In this sense we are proposing the second route identified above. This reflects our belief that ultimately the challenge in the aid evaluation literature is to come up with analyses that adhere as closely as possible to established academic principles. This does not however mean that this is the only way to go forward in this area. A multi-facetted approach where different pieces of a puzzle, which is not going to be solved in the short run, are looked for is clearly required. We conclude by noting that the two areas identified in this proposal for further work are in our assessment two critically important and potentially promising pieces of the aid evaluation puzzle.

10. References

Bigsten, A. (2005), "Donor coordination and the uses of aid", paper presented at the AFD/EUDN conference: Financing Development: What Are the Challenges in Expanding Aid Flows?, Paris December 14.

Bigsten, A. *et al.* (1994a), *Evaluation of Swedish Development Co-Operation with Tanzania: A Report for the Secretariat for Analysis of Swedish Development Assistance*, SASDA, Ds 1994:113.

Bigsten, A. *et al.* (1994b), *Evaluation of Swedish Development Co-Operation with Zambia: A Report for the Secretariat for Analysis of Swedish Development Assistance*, SASDA, Ds 1994:114.

Botchwey, K., Collier, P., Gunning, J.W., Hamada, K. (1998), *Report by a Group of Independent Persons: External Evaluation of the ESAF*. International Monetary Fund, Washington, D.C.

Burnside, C., Dollar, D. (2000), "Aid, Policies, and Growth", *American Economic Review* 90: 847-68.

Burnside, C., Dollar, D. (2004), "Aid, Policies and Growth: Revisiting the Evidence", World Bank Policy Research Working Paper No 3251, Washington DC.

Cassen, R. and associates (1994), *Does Aid Work?* Oxford University Press, New York.

Chauvet, L., Collier, P. (2004), "Development Effectiveness in Fragile States: Spillovers and Turnarounds", Centre for the Study of African Economies, Oxford University.

Collier, P., Dollar, D. (2004), "Development Effectiveness: What have we learnt?", *Economic Journal* 114: F2244-271.

Clemens, M.A., Radelet, S. Bhavnani, R. (2004), "Counting Chickens when they Hatch: The short-term of aid on growth", Working Paper no. 44, Center for Global Development, Washington DC.

Dalgaard, C.-J., Hansen, H. (2005). "The Return to Foreign Aid". University of Copenhagen, Discussion Paper 05-04.

Dalgaard, C-J:, Hansen. H., Tarp, F. (2004), "On the Empirics of Foreign Aid and Growth", *Economic Journal* 114: F191-F216.

Ddumba-Ssentamu, J., Dijkstra, G. (1999), "Swedish Programme Aid to Uganda: An Evaluation." SIDA Evaluation 1999/17:6. Swedish International Development Authority, Stockholm.

Devarajan, S., D.R. Dollar and T. Holmgren (2001), *Aid and Reform in Africa*, World Bank, Washington, DC.

Easterly, W., Levine, R., Roodman, D. (2004), "New Data, New Doubts: A comment on Burnside and Dollar's "Aid, Policies and growth (2000)", *American Economic Review* 94(3): 781-84.

Economic Commission for Africa (2005), *Our Common Future*, March.

Edgren, G. (2004), A*ssessing the Contribution of ODA to National Development Results: Proposal for a Joint Study*, mimeo, Stockholm.

Gunning, J.W. (2000), "The Reform of Aid: Conditionality, Selectivity and Ownership", in: *Conference on Aid and Development*, *20-21 January 2000. Proceedings and Discussion Papers*, Swedish International Development Cooperation Agency, pp. 14-23, Stockholm.

Gunning, J.W. (2001), "Rethinking Aid", in: B. Pleskovic and N. Stern (eds.), *Annual World Bank Conference on Development Economics 2000*, World Bank, pp. 125-144, Washington, DC.

Hansen, H., Tarp, F. (2000), "Aid effectiveness disputed" *Journal of International Development* 12: 375-98.

Hansen, H., Tarp, F. (2001), "Aid and growth regressions" *Journal of Development Economics* 64(2): 547-70.

IDD and Associates (2006), *Joint Evaluation of General Budget Support 1994-2004. Draft Synthesis Report. Executive Summary*, University of Birmingham..

IMF (International Monetary Fund) (1997), *The ESAF at Ten Years: Economic Adjustment and Reform in Low-Income Countries*. Washington, D.C.

Johnson, S. and Subramanian, A. (2005), "Aid, Governance, and the Political Economy: Growth and Institutions", paper presented at the seminar on Foreign Aid and Macroeconomic Management, Maputo, March.

Kanbur, R. (2003), "The Economics of International Aid". Working paper 39, Department of Applied Economics and Management, Cornell University.

Kanbur, R. (2000), "Aid, Conditionality and Debt in Africa". In F. Tarp (ed.) *Foreign Aid and Development: Lessons Learnt and Directions for the Future.* London and New York: Routledge.

Kasekende, L.A., Atingi-Ego, M. (1999), "Uganda's Experience with Aid." *Journal of African Economies* 8 (4): 617–49.

Knack, A. and Rahman, A. (2004), Donor Fragmentation and Bureaucratic Quality in Aid Recipients, Background Paper to World Development Report 2004, Washington DC.

Kraay, A (2005), Aid, Growth, and Poverty, paper presented at the seminar on Foreign Aid and Macroeconomic Management: Maputo, March.

Koeberle, S.G. et al. (eds.) (2005) *Conditionality Revisited: Concepts , Experiences, and Lessons*. World Bank, Washington.

Lawson, A., Booth, D. (2004), E*valuation Framework. Report to Management Group for the Joint Evaluation of General Budget Support*, ODI, London.

Little, I. M. D. og J. A. Mirrlees (1990). "Project Appraisal and Planning Twenty Years On". In *Proceedings of the World Bank Annual Conference on Development Economics 1990*, eds. Stanley Fisher, Dennis de Tray og Shekhar Shah, pp. 351-382.

McGillivray, M. (2005), "Is Aid Effective?", mimeo, WIDER, Helsinki.

McMillan, M., Rodrik, D. and Welch, K. H. (2002). "When Economic Reform Goes Wrong: Cashews in Mozambique," CEPR Discussion Papers 3519.

Mosley, P. (1987), *Overseas Development Aid: Its Defence and Reform*, Wheatsheaf, Brighton.

Oyejide, A., Ndulu, B. Gunning, J.W. (1999). "Introduction and Overview." In Ademola Oyejide, Benno Ndulu, and Jan Willem Gunning, eds., *Regional Integration and Trade Liberalization in Sub-Saharan Africa. Volume 2: Country Case-Studies,* Macmillan, London.

Platteau, J.P. and Gaspart, F. (2003) "The Risk of Resource Misappropriation in Community-Driven Development" *World Development* 31(10): 1687-1703.

Ravallion, M. (2001), "The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation", *World Bank Economic Review* 15(1): 15-40.

Roodman, D. (2004), "The Anarchy of Numbers: Aid, Development, and Cross-country Empirics". Working Paper 32, Center for Global Development, Washington DC.

Solow, R. (2001). "Applying Growth Theory Across Countries". *World Bank Economic Review* 15(2): 283-88.

Van Donge, J.K., White, H., Nghia, L.X. (1999), *Fostering High Growth in a Low Income Country: Programme Aid to Vietnam*, Sida, Stockholm.

White, H. (1994), *The Macroeconomics of Aid. Case studies of four countries*, SASDA, Ds 1994:115.

White, H. (1999), *Dollars, Dialogue, and Development: An Evaluation of Swedish Programme Aid*, Sida Evaluation Report, Sida, Stockholm.

White, H. (2005), "Challenges in Evaluating Development Effectiveness", in Pitman, G.K., Feinstein, O.N., Ingram, G.K. (2005), *Evaluating Development Effectiveness*, World Bank Series in Evaluation and Development Vol. 7, Washington DC.

World Bank (1998), *Assessing Aid. What Works, What Doesn't, and Why*, Oxford University Press, Oxford.