Evaluation Department

Norad

The Challenge of Assessing Aid Impact:

A Review of Norwegian Evaluation Practice

Study 1/2008

Norad
Norwegian Agency for
Development Cooperation
P.O.Box 8034 Dep, NO-0030 Oslo
Ruseløkkveien 26, Oslo, Norway

Phone: +47 22 24 20 30
Fax: +47 22 24 20 31

Layout and Print: Lobo Media AS, Oslo
ISBN: 978-82-754-272-1

The Challenge of Assessing Aid Impact: A Review of Norwegian Evaluation Practice

Alf Morten Jerve and Espen Villanger
Chr. Michelsen Institute, Bergen

January 2008

Contents

Abstract 7

1. Introduction: A Call for Improved Analysis 8

2. What are the Main Challenges of Aid Impact Evaluation? 9
- 2.1 How can the aid component be isolated? 10
- 2.2 How can causality in planned interventions be established? 10
  - 2.2.1 Designed as an experiment 11
  - 2.2.2 Use of comparison 11
  - 2.2.3 Use of baselines and secondary data 12
  - 2.2.4 The problem of selection bias 12
  - 2.2.5 Spill-over effects 14
  - 2.2.6 Fungibility 15
- 2.3 How can the analysis be tailored to practical constraints? 15

3. The Methodology of this Review 17

4. What can We Learn from Norwegian Impact Evaluations? 19
- 4.1 Evaluation of the Debit Credit and Savings Institution (DECSI) in Tigray, Ethiopia 19
- 4.2 Marginalised groups and empowerment – a study of DECSI in Ethiopia 21
- 4.3 Evaluation of Norwegian support to psycho-social projects in Bosnia-Herzegovina and the Caucasus 23
- 4.4 Evaluation of Norwegian support to FORUT, Sri Lanka 25
- 4.5 Evaluation of Save the Children Norway (SCN) in Ethiopia 27
- 4.6 Evaluation of the Tanzania-Norway Development Cooperation 1994-1997 28
- 4.7 Evaluation of Norwegian support to FIFAMANOR, Madagascar 30

5. Practical Recommendations for Impact Evaluations 33

References 36

Abstract

The aim of this report is to contribute to a discussion on how to strengthen the evaluation practice of Norwegian aid. Norwegian aid authorities, in recent policy documents, emphasise the need for improved methods for assessing impacts of development aid. The issue is also one of setting sound and realistic objectives for aid. The report (1) reviews recent Norwegian aid evaluations with an explicit mandate to study impact, and assesses how the evaluators establish causal effects in their analyses, and (2) provides recommendations for how to improve the quality of aid impact evaluations.

The seven evaluation reports presented exemplifies great diversity in methodological approaches used – from econometric analysis based on survey data to qualitative assessments by the evaluators based on project documents. The analytical challenges encountered are threefold. One, that the commissioning agency asks for evidence of impact where this is not possible to identify, largely because the role of aid appears to be extremely marginal relative to the processes of societal change for which it is targeted. Second, the distinction between impacts of the aid element versus the totality of a development intervention is often blurred. Third, the methodological approaches used are either poorly developed or applied superficially because of resource constraints.

The main recommendations are: *First*, the Terms of References need to be internally consistent – i.e. reflecting a realistic scientific approach in terms of hypotheses of impact, the intended methodological approach, budget and time. *Second*, it is necessary to improve the analytical and empirical basis for conclusions on impacts. The use of data in the analysis was often ad hoc, or lacking. Most of the studies did not develop any logical chain on how an aid or development intervention would be expected to impact on the participants and non-participants. *Third*, the design of the evaluation should take into account unintended effects both for beneficiaries and for non-beneficiaries and former beneficiaries. *Fourth*, timing of the evaluation is critical, and should be based on carefully developed hypotheses about when one can expect specific impacts to emerge. *Finally*, the report emphasises that impact evaluation requires specific knowledge about evaluation methodologies in addition to advanced analytical skills. This must be taken into account when commissioning impact evaluations.

1. Introduction: A Call for Improved Analysis

Norad announced in 2006, as part of its new strategy, that it had decided to prepare an annual report on the results of Norwegian development aid.[1] The strategy argues that "aid yields results, but we know too little about how it works and the magnitude of the impacts". The first report on the results of Norwegian aid was published on 26 November 2007 and indeed it confirms the lack of knowledge of the precise impact of Norwegian aid. The report basically argues that one cannot talk of Norwegian results as such, if we refer to overall economic and political changes in a recipient country. Nevertheless, the report brings to the reader's attention a number of examples of positive results from Norway's engagement in more focused areas, such as national fisheries management in Namibia and hydropower development in Nepal.

The report amply illustrates the challenge facing results reporting on aid, despite the fact that there has been a range of evaluations of different aid projects and programmes during the last 20-30 years. The problem is twofold. It stems from the inherently difficult empirical task of linking cause and effect in studies of human society, which is made no easier with the growing ambitions set out for aid, and the corresponding demands for evidence of impact which we have witnessed in recent years. The discrepancy between the aims of foreign aid and its relative role in important outcomes has been increasing, particularly since the initiation of the Millennium Development Goals (MDGs). Many politicians, and some scholars, take the starting point that aid is to be the major instrument for achieving the MDGs. Similarly, the aid-based efforts to promote political development and peace and reconciliation have illustrated the gap between the realistic impacts of aid and what donors attempt to achieve.

Earlier, donors could be satisfied if a project they supported met its output targets, often measured by the physical actions undertaken in the project. Contemporary demands, however, require that donors specify the impacts of aid on the well-being of the recipient people or on political and economic processes in the recipients' society. Undoubtedly, this perspective is necessary for justifying aid strategies, but most aid evaluations tend to deal with such connections under the rubric of 'relevance' only, i.e. to make assumptions about impact rather than empirically verify impact. Moreover, the task of measuring the impacts of Norwegian aid is made even more challenging by the current trend towards redirecting aid from financing projects into new aid modalities such as sector and budget support – a point that is underlined in the Norad strategy.

The Norad strategy's emphasis on the need for the development of improved methods for results reporting and impact assessment is thus not only important to aid effectiveness in itself, but also for developing realistic and sound aid objectives and for giving an accurate basis for the political justification of aid. Thus, the aim of this report is to contribute to a discussion on how to strengthen the evaluation practice of Norwegian aid with these issues in mind. We shall:

- assess recent Norwegian aid evaluations with a mandate to assess impact, and see how the evaluators establish causal effects in their analyses, and
- use this assessment to provide practical recommendations for the evaluation of aid interventions.

The report is structured as follows: In section 2, specific challenges of impact evaluation are elaborated. Section 3 contains a description of the methodology used in this report and an overview of the evaluation studies that are scrutinised. This forms the basis for the analysis of the selected impact studies in section 4. The assessment, together with the main lessons from the literature on impact evaluations, serves as an input into recommendations for improving the evaluation practice of Norwegian aid in section 5.

1 The strategy focuses particularly on challenges to reporting on the results of Norwegian aid to sectoral programmes and budget support.

2. What are the Main Challenges of Aid Impact Evaluation?

The terminology of concepts used in aid evaluation is in itself a subject of discourse. In this report we use the following OECD/DAC definitions of terms used in aid effectiveness analysis:[2]

- **inputs** are the financial, human and material resources used for the development intervention,
- **outputs** are the products, capital goods and services which result from a development intervention; and may also include changes resulting from the intervention which are relevant to the achievement of outcomes,
- **outcomes** are the likely or achieved short-term and medium-term effects of an intervention's outputs, and
- **impacts** are the positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended.

It is evident that the distinction between outcome and impact may be blurred in practice. This is best resolved by applying a precise definition of the two concepts in the concrete evaluation and what they imply for the case under study. The term '**results**' is generally used quite broadly and may refer to outputs and outcomes as well as impacts.

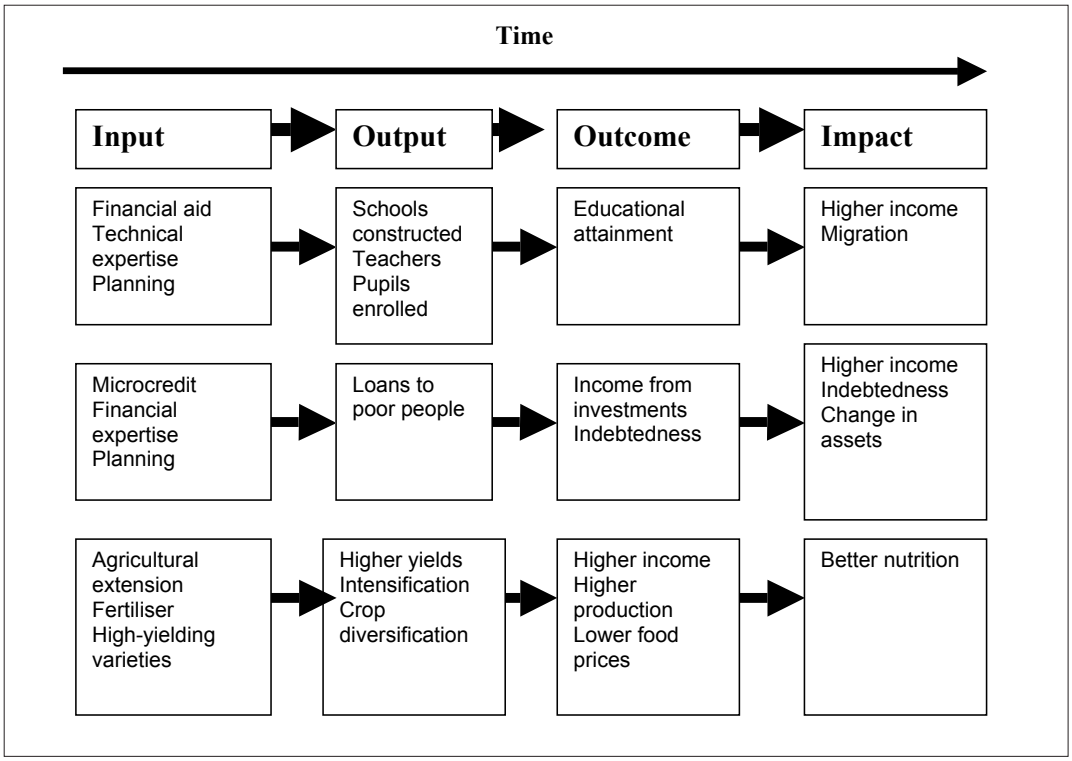
The relationships between these concepts, including some practical examples, are illustrated in Figure 1. It is important to note that the chains of events resulting from the inputs are in most cases not the same as those most easily observed by the evaluator. Many events and processes not related to projects usually intermingle with the process from input to impact as illustrated in Figure 1.

It is thus one of the evaluator's main tasks to disentangle the effects of the programme inputs from all other events and processes occurring simultaneously. In so doing, it is useful to distinguish between three types of challenge:

- how to isolate the aid component;
- how to establish causality in planned interventions; and
- how to tailor the analysis to practical constraints.

2 See http://www.oecd.org/dataoecd/29/21/2754804.pdf

Figure 1. Following the results chain



2.1 How can the aid component be isolated?

Aid is rarely the only source of financing for a public or private development intervention and most often it is not even the largest source. But the aid element is not only an issue of finance, since the donor agency in most cases takes an active organisational role in influencing the process of planning and implementing the interventions to which it is contributing. Furthermore, aid comes in different forms, the main one being budgetary contributions, but aid in the form of personnel (i.e. transfer of knowledge) and in kind also plays an important role.

Aid evaluations frequently blur the distinction between the role and impact of the aid component versus the programmes being supported. How is it possible to differentiate analytically between the impacts of a micro-credit programme and the aid provided to the implementing organisation? The objectives of the former, and hence appropriate results indicators, relate to improvements in household-level incomes, while the objectives underpinning the aid involvement might be of a different kind. While it obviously matters to the donor whether the programme achieves its ultimate impact objectives, the aid component as an add-on may have its particular justifications, such as expansion of operations, experimenting with new approaches, skills development etc.

Hence, the impacts of the aid component are not synonymous with programme impacts. Obviously, it is difficult to discern the aid influence factors in an ex-post assessment of a particular programme or intervention. How much and in what way can aid be partly credited for the final outputs and impacts? The same can be said for failures or non-achievement of objectives. It is important, nevertheless, that the terms of reference (analytical design) of aid evaluations address this distinction by formulating hypotheses about possible effects of the aid relationship on the process of implementation and the institutional actors involved. For instance, the aid component is often thought of as a catalytic factor or a capacity boosting factor. Did such factors make a difference for the longer-term programme impacts?

2.2 How can causality in planned interventions be established?

What are the most frequent methodological pitfalls when establishing causal effects between (1) *input* and *output* indicators of programmes – understood as planned public or private interventions – targeting certain beneficiaries, whether individuals, households, villages or larger geographical areas, and (2) indicators of social, economic and institutional change

(*outcome* and *impact*)? Logically speaking, the ultimate evidence of impact is the counterfactual, i.e. evidence of what the situation would have been or is without the specific intervention under study. It is of course inherently difficult to assess a programme's performance against an explicitly defined counterfactual. It is not possible for people simultaneously to benefit and not benefit from a programme. Moreover, the counterfactual is usually quite difficult to establish because programmes seldom address beneficiaries in a random manner. However, the literature on what is labelled '*programme evaluation*'[3] contains a range of solutions to these problems, and in the following we shall present the main methodological issues raised in this literature that are of relevance to most evaluations of aid interventions (see for example Ravallion 2001).

There is a wide range of methodologies for assessing the effects of other types of intervention (aid-funded or not) where the application of statistical/econometric methods is impossible, typically where particular beneficiaries are not identified. In such cases impact analysis will rely on other types of data that make it possible to study change over time, based on the analysis of historical processes, institutional change, life stories and informants' perceptions. Essential to all types of impact evaluation, however, is that conclusions are based on a systematic and objective analysis of empirical information.

One approach, which is often found in current evaluations, is to compile output figures and impact indicators for programme participants only, undertake interviews with participants and stakeholders and summarise secondary data, and then apply theory and prior knowledge to discuss causality and likely programme effects. Although important insights can be generated, this evaluation design does not qualify as a sound impact evaluation design as it cannot be considered as producing rigorous quantitative estimates of project impact (World Bank 2006). In practice, it may be challenging to construct a proper counterfactual, or even impossible. However, it is important for the evaluator to think through what would be a good counterfactual since this by itself will give important inputs to the evaluation and prevent erroneous conclusions from being drawn.

2.2.1 Designed as an experiment

The ideal setting for an evaluation is that it is included in the project design from the outset. One approach for designing high-quality evaluations of new aid programmes is to sequence their implementation randomly between different areas or based on some observable characteristics of the participants/beneficiaries. Then, one would measure impact indicators (e.g. income, assets or health status) before the programme starts both for randomly selected participants and for randomly selected people in the areas where the programme is not implemented in the first phase.

Measuring the impact indicators for the same people after the programme has ended and comparing the change over time for the participants with the changes for the non-participants would most likely yield the true effect of the programme if no other major events or processes affect one or the other group. An example of such an event could be a natural disaster occurring only in one of the areas during the programme period, i.e. only participants or non-participants would have been affected. An example of a process having a similar asymmetrical effect is where the proportion of migrants from the programme area is considerably higher than from the comparison area, or vice versa. However, if no such events or processes affect the indicator of interest, then there is sound analytical evidence of programme impact which can be used for further extension of the programme.

2.2.2 Use of comparison

Although most aid programmes are not suited for randomised delivery of services, this need not impose large problems on the application of programme evaluation techniques. In the case where the programme is targeted at participants based on their characteristics, for example towards the poor, one can find the approximate programme effect by *comparing participants with non-participants* that share characteristics that may influence the impact of the programme. Take an income-generating scheme, for example. If those who are chosen for participation are those with little education *and* the level of education may influence how

3 For a recent overview of the main issues in the programme evaluation literature, see Ravallion (2006) in the Handbook of Development Economics.

much income they will be able to generate through the programme, then the comparison group must have the same composition of people when it comes to education to get a good estimate for the difference between participants and non-participants. This approach can be useful if the aid programme is targeted towards a particular area or group.

One may also make use of programme evaluation techniques to produce rapid assessments where the evaluator compares the programme area with *similar areas*, and assesses differences over time. The challenge is to find areas that are similar to the programme area (economic level, central location or remoteness, infrastructure, economic activities, institutions, service provision etc.). Although shallower than more comparison groups, such approaches will usually yield useful information on the impacts of aid programmes.

If it is difficult to find a good comparison group that matches the target group, then one may consider econometric programme evaluation techniques where one first identifies factors that affect the likelihood of being in the target group, and then use the results from the first estimation to correct any bias due to non-random selection of the target group in a second stage where the programme effect is estimated. While simple matching of comparison groups does not require many resources, more advanced matching techniques and two-stage estimations require quality data and much research effort, which may be too costly for many evaluation studies.

2.2.3 Use of baselines and secondary data

Although the programme evaluation techniques discussed above ideally compare pre- and post-intervention situations, it is important to acknowledge that the evaluator can come a long way even if only post-intervention data is available. Hence, the *absence of baseline data*, which is usually the case for aid projects, need not impede the use of programme evaluation. Again, the important issues are that the data need to reflect the characteristics that are used to assign the programme to beneficiaries, and that we believe unobserved characteristics do not play an important role in programme participation and individual outcomes. Hence, the more that is known about the project context and participant selection process, the more confidence can be attributed to the results, as one will be better able to construct a good comparison group, and thus identify the true programme effect. As data collection often amounts to half the cost of the evaluation (Baker 2000), this approach of relying solely on post-intervention data reduces costs substantially from comparing data collected before and after, which often entails tracking people who moved during the period. For an overview of different low-cost designs of impact evaluations and a case study example, see World Bank (2006). The important point is that economical designs can reduce the cost of data collection by up to 50%.

The use of secondary data can be a costless and rapid approach to evaluating impacts, but should always be triangulated, i.e. by comparing the results from such estimations with results from other approaches such as direct observation, other secondary sources, key informants, stakeholder surveys, Participatory Rapid Assessments, photographs and newspaper articles (World Bank 2006). For an elaboration of the challenges of using secondary data in quality evaluations, see Bamberger et al. (2006), and for a checklist useful for assessing the potential weaknesses in secondary data, see World Bank (2006).

2.2.4 The problem of selection bias

Establishing the causal relationship between the output and impact indicators is a more challenging task than the simple approach of reporting the degree of co-variation between them. One major challenge for assessing the impacts of a programme occurs when the programme area is chosen based on characteristics of the population or area that in themselves would affect the success indicators chosen for the programme independently of its interventions:

Is the selection of participants determined by characteristics that also influence the impacts of the programme?

This selection problem is often encountered in standard consultancy approaches towards assessing the impact of micro-credit programmes, when the approach is to count the number of entrepreneurs that started a business after they got a micro-loan. However, if participation

in the micro-credit scheme is open to all, then many of the people who get finance are most likely clever entrepreneurs who would probably have started a business anyway. If this is the case, then the effect of the credit programme would be greatly exaggerated if measured by the success of the participants. An estimate of the actual programme effect could have been obtained by comparing two similar communities, one where the micro-credit scheme was implemented, and one without such a programme. Then one could compare the number of entrepreneurs starting business in the two communities, and the difference between the increases in entrepreneurs could be ascribed to the micro-credit programme. Hence, if some entrepreneurs would anyway start businesses even without the micro-credit programme, then the true impact of this programme will be lower than would be reported by the consultancy approach. It is evident that very few of the many evaluations of micro-credit programmes take such selection problems into account (see Hatlebakk 2007 for an overview of studies of micro-credit that do account for this problem).

Another example of selection bias can be found in electrification programmes. When building new extensions of the electricity grid, even in rural areas, it is often the aim that new customers are to contribute significantly to the economic viability of the project through connection fees and user rates. Hence, new lines will often be built first to the growth centres and more advantaged areas since this is where the probability of cost recovery is the highest. If we want to assess the impact of electrification on the incomes of the people who are connected, the common approach would be to measure incomes before and after they were connected. This is likely, however, not to give us a true picture of the effects generated by the programme.

Firstly, a growing economy yields income growth for many people irrespective of electricity. Hence, building lines to growing areas gives a correlation between growth and connections, but there need not be any causal relationship from electricity supply to economic growth. Secondly, it is only the richest segment that can afford to connect to the grid, and rich people usually have more resources for income generation than the poor. If this results in higher income growth for the rich, we also get a correlation between growth and electricity where there need not be any impact of electricity on incomes. Moreover, a household with high income growth may be more prone to take on the expenses of connecting to the grid in the expectation of higher income levels in the future. Vice versa, a stagnant or declining household economy would make the household more reluctant to invest in electricity in the expectation that incomes might fall below the level where consumption of such a luxury good can be maintained. Both patterns could give a relationship where income growth and electrification are strongly correlated, but where there are no impacts of electrification on household incomes.

In these situations the evaluator must try to establish the counterfactual: what would be the impact if the programme had not been implemented. Constructing a counterfactual for such programmes may not be too time-consuming or difficult since it is rather obvious that key characteristics determining the selection of beneficiaries of the programmes (entrepreneurship in the case of micro-finance and initial economic strength in the case of electrification) impact on the result of the programmes. Then the evaluator must compare the results of the participants with the same indicators for similar groups that did not have the opportunity to participate. Thus, as long as the characteristics that determine the selection of the participants in a programme is known and taken into account in the analysis, the evaluator will get a correct indication of the impact of the programme.

Another selection problem occurs when programme participants can choose whether or not to participate in the programme:

Are there any opportunities for the participants to select themselves into, or out of, the programme?
In cases where participants can determine whether they participate or not, one may expect that those who choose to participate are also those who will benefit more. Vice versa, those who see no benefit from their own participation will not participate. Hence, comparing impacts for participants with non-participants may exaggerate the impacts of the programme. The point is evident from the above example in that those with more entrepreneurial skills are likely to choose to participate in an income-generating programme, while those that do not possess

such skills are not. Hence, the non-participants do not constitute a comparison group that can act as a counterfactual for the entrepreneurs in the programme. The latter group would probably have quite different incomes in the absence of the programme than those with less entrepreneurial skill. Hence, the evaluator needs to find out how this self-selection bias affects the impacts for participants as well as non-participants. In the micro-credit example above the challenge would be to find a comparison group with the same entrepreneurial skills as the programme participants.

There is plenty of research that finds large biases in estimates of impacts from evaluation studies that do not take into account selection problems. Three examples illustrate the point. First, what is the effect of rural roads on poverty reduction? van de Walle (2002) demonstrates that comparing the incomes of villages that have a rural road with those that do not gives the result that building a rural road will generate large increases in incomes. However, van de Walle shows that if the analysis takes into account the fact that the level of economic activity mattered in selecting which villages got a road, it will show that the real effects on incomes from building the roads are much smaller.

The second example is evaluating training programmes for the unemployed where participants choose whether to join the programme. It is likely that those who are most committed and/or able to get out of unemployment will join the programme. Hence, comparing the rates of unemployment for former programme participants with non-participants will indicate that the effect of the programme is larger than is actually the case since many of those on the training programme would probably get a job irrespective of the training.

A third example is the evaluation of the impacts of flip charts on academic performance in Kenyan schools. The study found that parent-teacher associations became more active in schools that received flip charts, and that parents were more eager to get their children to study harder (World Bank 2006). Hence, the improvements in academic performance by schools that received flip charts may have been caused by the extra engagement from parents and teachers and not from having the flip chart.

The general point for impact evaluation is that some participants may be encouraged by being selected for the programme and may perform better, and some may be discouraged by not being included and may perform worse.

2.2.5 Spill-over effects

Some aid programmes carry with them spillover effects on people that were not among the intended beneficiaries. Hence, the evaluator should be aware of possible hidden impacts when attempting to create a comparison group.

Are the effects on non-participants included in the evaluation of the project?
Successful income-generating programmes may serve as an illustrative example. If a group of people included in such a programme starts to earn much money from the project, then their spending of this money will also have an effect on other people that share the same markets. It is possible to envisage a scenario where the extra incomes of people in a programme increase the income of people in the comparison group. The true effect will hence be higher than the measured effect, which in this case is the difference in incomes between participants and non-participants.

Does the programme affect the market?
More complicated spill-over effects can be found once the programme impacts on the market. Programmes that guarantee a minimum wage for participants illustrate the point. With the presence of such programmes, few will be prepared to accept other work for wages below this minimum wage. Hence, the evaluator would most probably observe a concentration of wage rates at the minimum wage, both for participants and non-participants. Comparing the programme group with a comparison group of non-participants would then indicate that the programme did not have any effect since the wages for both groups are the same. In reality, however, there are large impacts for both participants and non-participants since many of those who previously received wages below the threshold are now paid the minimum wage.

Hence, not taking into account the functioning of the market may distort the analysis of a programme's impacts.

2.2.6 Fungibility

If a donor decides to finance a road that the government would have built anyway, then the true effect of the aid is likely to be that some other projects are financed from the extra funds available to the government. This is called fungibility in the literature. In such cases the effect of the aid would be found from measuring the impacts of these other projects funded by the government. The evaluator needs to ask:

How does the financing decision of the donor relate to the government's own priorities? Would the selected intervention have been financed by the government independently of the aid?

A concrete example highlights the point. If there are ten roads with the same cost ranked on a priority list – e.g. according to their benefits to society – and the donor steps in to fund the first road on the list and the government funds the next five roads, then the impact assessment of the aid should calculate the benefits from the sixth project. The counterfactual situation with no aid would be that the government funded the five roads highest on the list. A study by van de Walle and Cratty (2005) analyses this issue, but dealing with fungibility in aid evaluations is difficult since there is often no official ranking of proposed projects and donors' real preferences may not be known to the evaluator. However, it is important for the credibility of the evaluation that the issue is assessed.

2.3 How can the analysis be tailored to practical constraints?

A challenge for aid impact evaluations is to balance scientific requirements for impact assessments with practical matters such as limitations of data, time and funding. Building on the literature in the field of programme evaluation and the review of a sample of evaluations (below), an aim of this study is to address some of the trade-offs between practical limitations and scientific requirements both in short-term consultancies, affording two weeks for touring a project area and writing an assessment, and more comprehensive evaluations where the design of the project has incorporated an impact evaluation component with detailed data collection before and after a project that also includes treatment and comparison groups.

The resources used for evaluating projects are often weighed according to the resources used for the particular aid project under scrutiny. Hence, small projects are often neglected when it comes to evaluating impacts, while medium and large projects are subjected to rapid assessments if they are evaluated. In some cases, however, it may be more useful to undertake thorough analysis of a few small and medium or large interventions respectively to see what similar projects can learn from those under study, than to perform shallow, quick assessments of many projects in each category.

It is important for the evaluator to acquire information about the details of the administrative and institutional aspects of the programme, typically available from the programme administration. Such data are key to the design of the survey for constructing the counterfactual and assessing the impacts of the programme. Moreover, the knowledge of the programme context and design can be important for the evaluator in handling any problems arising from unobservable characteristics of the participants that are likely to influence the impacts of the programme.

Evidently, the data requirements for a thorough impact evaluation can be very demanding, and there is a temptation to rely on less formal, unstructured interviews with participants. It has proven to be very difficult to use this approach to assess the counterfactual. Most respondents will have great difficulty in assessing the counterfactual for their own situation. Asking questions like "what would you be doing if the programme had not been initiated", and "how would that situation compare with the contemporary situation" will in most cases not be useful as anything more than a complement to quantitative survey data, and seldom be a credible impact evaluation on its own.

The inaccurate measurements of many living standard indicators also impose a limitation. Income figures are renowned for their notorious inaccuracies, especially when it comes to

underreporting actual earnings and to difficulties in measuring the incomes of poor rural households with their own large production. Take the poor farmer who is asked about his/her income. How is s/he supposed to know that the seeds bought for planting should be deducted from the farm revenues when computing income, while the seeds bought in the same purchase for own consumption should be reckoned as household expenditure? Consumption figures are found to be more reliable indicators of well-being than income figures, but collecting information on these indicators is time-consuming. Moreover, a selective non-response problem often also occurs when richer households are not interested in revealing their incomes and thus refuse to participate in the survey. There is much literature on measurement errors in income and consumption surveys and how to deal with them. Deaton (1997) gives a good introduction to the issue.

Finally, a note to evaluators using different data sources for comparing programme and comparison groups is warranted. Several research findings indicate that differences in the design of survey instruments can cause severe bias in such estimates (see Ravallion 2006 and the references therein for more information).

3. The Methodology of this Review

We have reviewed a limited sample of evaluations of Norwegian aid commissioned either by the Ministry of Foreign Affairs or Norad or in two cases by Norwegian People's Aid. The selection is influenced by two main factors. Firstly, we were looking for evaluations that explicitly addressed the question of impact. Hence, studies of whether or not operational procedures have been followed in specific projects, or whether the actual performance of a project has proceeded according to plan, have not been included in this study. The same goes for different types of reviews and descriptions of aid efforts. This factor, looking at evaluations commissioned since 1996, in fact drastically reduced the potential sample; the majority did not address impact, since the focus was on outputs and effectiveness of funding (aid modalities and channels) and output targets. Secondly, we looked for evaluations that served as illustrations of different and commonly encountered methodological challenges.

The purpose of this review is to assess how evaluators have solved the methodological challenges of establishing or refuting any causal relations between an intervention and stated impacts, i.e. the impacts that are referred to in the ToR of the evaluation. It is worth noting that the DAC Evaluation Quality Standards, for instance, are completely silent on the issue of methodology and on how to handle difficult trade-offs.[4] Our main approach in this review has been to look at the literature on programme evaluations (see Section 2) and assess to what extent applying this methodological framework would have improved the impact evaluations of the selected reports. We have also assessed the evaluations more generally to see to what extent conclusions on impact are based on a systematic and objective analysis of empirical information.

A number of basic conditions influence choice of methodology in an impact evaluation. This challenge differs according to the length of the perceived chain of causality between intervention and societal effect (outcome/impact). When effect is defined in terms of indicators that represent very complex processes (e.g. personal income or national GDP) the likelihood is great that no robust conclusions can be drawn. Where effect is defined at an intermediate, shorter-term level (outcome), the identification of causality becomes somewhat easier. The exact location of a study on this continuum of effects from the immediate to the long-term, and from near to higher-level impacts, has implications for the analytical design. We have assessed how this issue is treated both in the terms of reference and by the evaluators.

Borrowing terminology from natural science, the most basic challenge relates to the "without" question – to establish a counterfactual.[5] We also ask to what extent the type of intervention studied renders itself to testing for or even speculating on the counterfactual. To what extent has the design of the intervention itself been influenced by evaluation requirements? Would it have been possible to carry out randomised studies among beneficiaries and non-beneficiaries?

As stated above, it is critical to distinguish between the effects of aid and the effects of a particular intervention (project/programme, policy etc.). Aid is rarely the only financial source – even more so aid from a particular donor – and rarely the only trigger of an intervention. Aid is also not only money, but comes with additional attributes such as transfer of knowledge/technical assistance and also forms of conditionality. Have the evaluators studied the aid influence on a particular intervention/programme or the impacts of intervention as such? It is quite possible that aid can have negative effects on a programme achieving positive results, or vice versa. We have in particular studied the methodology and conclusion sections of the evaluations.

4 http://www.oecd.org/dataoecd/34/21/38686856.pdf
5 The situation or condition which hypothetically may prevail for individuals, organisations or groups were there no development intervention.

The sample for this review contains studies undertaken by some of the most relevant institutions carrying out impact evaluations of development aid in Norway: CMI, COWI, ECON, FAFO, NIBR, Nordic Consulting Group, NUPI, UMB Noragric, Scanteam and Norconsult. From this sample seven evaluations were selected based on the following main criteria: what can we learn from impact evaluations of Norwegian aid and what are the concrete lessons to be learned with respect to methodological challenges and deficiencies? We selected the evaluations so that they would differ both in terms of the types of intervention (programmes/projects/activities) that are the focus of analysis, and in what kinds of impact (or outcome) have been anticipated in planning documents and given as justification for the support. In order to streamline the review, we have chosen not to include studies that reveal lessons similar to the seven selected evaluations. In only one study (4.1) did we find that the evaluators also looked for unintended impacts. In the table below we have organised the selected studies according to the characteristics above.

Table 1. Categorisation of the selected studies

	Name of study	**Commis-sioned by**	**Institution responsible for evaluation**	**Type of intervention**	**Impacts identified in plans, ToR and/ or study**
1	Credible Credit. Impact study of the Dedebit Credit and Savings Institution (DECSI), Tigray, Ethiopia. Borchgrevink et al. (2003)	NPA	NUPI	Micro-credit	Income Consumption Assets Community development
2	Marginalised groups, credit and empowerment. Borchgrevink et al. (2005)	NPA	NUPI	Micro-credit	Empowerment
3	Evaluation of Norwegian Support to Psycho-Social Projects in Bosnia-Herzegovina and the Caucasus. Agger et al. (1999)	MFA	COWI	Support to groups – mainly women	Return to normal functions New initiatives
4	Study of the impact of the work of FORUT in Sri Lanka: Building Civil Society. Baklien et al. (2004)	Norad	NIBR	Micro credit Service provision	Capacity building Poverty reduction Good governance
5	Study of the impact of the work of Save the Children Norway in Ethiopia. Helland (2004)	Norad	CMI	Education HIV/AIDS Child rights	Democracy Human rights Poverty
6	Evaluation of the Tanzania-Norway Development Cooperation 1994-1997. ECON (1999)	MFA	ECON	Country programme	Reduced aid dependence
7	Review of Norwegian support to FIFAMANOR. Aune et al. (2005).	Norad	Noragric	Support to public utility company	Income diversification Nutrition

4. What can We Learn from Norwegian Impact Evaluations?

The review is conducted and reported along the following lines. First, it examines the ToR to see to what extent the commission is clear on what impacts are to be analysed. Second, it assesses whether the ToR specifies the methodology that is to be used in the evaluation and whether the concretisations of the tasks are in line with the impact evaluation aims of the commission. This is important because it may clarify what results are expected from the evaluation and the internal consistency of the ToR. Then it scrutinises the methodology used in the particular assessment. How does the methodological approach used match the analytical challenges with re spect to potential impact identified in the ToR or purpose of the study? Further, we discuss to what extent the results from the evaluation are obtained through a proper implementation of the methodology and whether there is coherence between the methodology in use and the conclusions drawn. This forms the basis for an assessment of how the study could have been designed differently (i.e. improved) given the time and financial resources made available.

For the purpose of this study to spell out what we can learn from impact evaluations of Norwegian aid, we have selected studies where there are concrete lessons to be learned with respect to methodological challenges and deficiencies. Hence, some of the evaluations we reviewed have not been incorporated in the presentation below. Perhaps needless to say, there are also examples of evaluations where the aims expressed in the ToR correspond well with the methodology used and the resources made available for conducting the assessment. We have observed, however, that this is the case primarily with evaluations that do not pretend to assess impact. The review below starts with the cases representing the most elaborate methodological approach.

4.1 Evaluation of the Debit Credit and Savings Institution (DECSI) in Tigray, Ethiopia

Borchgrevink et al. (2003) analyse the impacts of an Ethiopian micro-credit scheme, DECSI, which is co-financed by Norwegian Peoples Aid. Their main objective was to assess to what degree the programme contributed to poverty reduction in terms of income, assets and reduced vulnerability, and to development efforts in terms of agricultural production, marketing of agricultural products, income diversification and the spin-off effects of the financed activities. The evaluation input was 7 weeks of work for two Norwegian consultants, 6 weeks for a local researcher and 4-5 weeks for local enumerators with a budget of approximately NOK 650 000. This study is chosen because the study team believes it shows how far one can get in impact evaluation within modest evaluation budgets when the evaluators have the right competencies. To keep focus on the issues important to impact evaluation we confine the assessment to the poverty impacts.

How does the evaluation report discuss methodology for impact assessment?
The report has a separate methodology chapter that discusses the details of collecting primary quantitative household data and the related selection process, the collection of qualitative data, the approach towards triangulating the methods and a thorough discussion of the methodological problems and challenges, including selection bias.[6] The issue of constructing a counterfactual is explicitly discussed and applied in the design of the quantitative survey and the qualitative approach.

There were no restrictions on people entering the DECSI programme, which implies that it is necessary to control for self-selection (see Section 2.2.4 above). The authors grasp this point, and substantiate the hypothesis that the wealth level of the participants may influence the impact on those who participated, i.e. that poorer households are more likely to perform worse

6 The discussion of selection bias is elaborated in footnote 25, page 53 of the report.

in the productive activity financed by the loan. Hence, richer people may choose to participate in DECSI while poorer people do not. Taking this into account, the authors collected data from programme participants based on a random selection of those who had benefited from the project. These were ranked according to wealth measures and then contrasted with data from randomly chosen households that had not participated in the programme. Based on this, they selected a comparison group of non-participants based on the number of participants in each wealth-group in each district. This implies that when they interviewed 9 participants in a district where 2 were rich, 3 were medium wealthy, 2 were poor and 2 ultra-poor, then the 9 non-participants also contained 2 rich, 3 medium wealthy, 2 poor and 2 ultra-poor households.

What conclusions are made with respect to impact?
With respect to impacts at the household level, the following information taken from the conclusions in the report illustrates that the authors found significant improvements in the economic well-being of many of the DECSI participants, while service delivery seems to be unaffected.

Table 2. Percentage of clients and non-clients reporting improvements over the last 5 years

	Clients	Non-clients
Improved living standard of the household	59%	33%
Improved income of the household	47%	35%
Increased asset base of the household	51%	30%
Increased quality of food consumption	54%	38%
Increased quantity of food consumption	49%	32%
Improved health of household members	62%	62%
Improved education of household's children	76%	72%
No children dropped out of school in 5 years	75%	74%
Improved access to clean water	72%	69%
Female members' community participation	60%	63%

The authors discuss weaknesses of the estimates and draw a general conclusion: "The DECSI programme has had a positive impact on the lives of the clients…. Their situation has improved in terms of income, consumption and assets. They also seem more food secure and less vulnerable to shocks. Improvements are equally distributed among clients of different wealth categories."

What is the analytical basis for the conclusions?
The analytical basis for the authors' conclusions is strong and relies on both qualitative and quantitative data analysis with the data having been collected specifically for the purpose of analysing impacts. The authors' approach of constructing a counterfactual would give the true impact of the credit programme on poverty if no other unobserved characteristics impacted on the outcomes except wealth. If, however, the DECSI participants were better entrepreneurs than non-participants, we would not get the correct effect of the programme since these people would probably get finance for some income-generating activities even if DECSI had not been implemented.

The qualitative interviews may give some feedback on this selection question, and it is interesting to note that all the case stories selected by the authors and referred to in the report indicate that the DECSI participant is a true entrepreneur. One had started a business before entering DECSI, one started several highly profitable businesses (building houses for rental, producing and selling wine and starting a shop), and one started seasonal trading (100% profit a year). However, the fact that four out of five cases seem to be entrepreneurs may not be representative of the whole sample if the stories are included only to illustrate specific points. Borchgrevink et al. state that it is difficult to judge whether and to what extent such selection effects have influenced their results, but more use of the qualitative aspects could have shed further light on the issue (see footnote 25 in the report).

It is crucial for an impact evaluation to do a thorough investigation of the selection problem, in particular for micro-credit schemes, and the Ethiopian study reveals that this can be

difficult even when one constructs the questionnaires purposely for the impact study. In order to reveal whether there was a larger share of entrepreneurs in the group of DECSI participants as compared to the comparison group, one could have asked the respondents in both groups to provide information on previous engagements in entrepreneurial activities. Moreover, other indicators that were actually collected in the Ethiopian survey may give important information on entrepreneurial skills, or abilities in general, for example literacy, education and occupation. These data are not used for such purposes in the report.

The report would also have benefited from a section with descriptive statistics so that the reader could ascertain whether the DECSI and comparison groups were similar along other dimensions important for income growth and poverty reduction. The age of the breadwinners, for example, is often found to matter for income and income growth, and thus for poverty reduction. Young household heads tend to have lower incomes, but the growth over time can be high. Old household heads also tend to have low incomes, but declining over time, while the group in between has the highest incomes. These patterns can have an impact on assessments similar to this one since the authors investigate change in income during the last five years. However, without the data from the questionnaires it is not possible for the reader to verify whether both groups actually have the same characteristics important for income growth.

Finally, it is evident that more of the information collected in the survey could have been used to cross-check the results. For instance, if the authors were right in their assumption that only the wealth of the household matters for the outcomes, then the DECSI group should have significantly higher levels of income and consumption as compared to the comparison group. Moreover, a proper econometric analysis would probably also enable the authors to indicate the magnitudes of the impacts. This could in turn be related to the magnitudes of the inputs, giving a more informative assessment that compares the value of the input with the value of the output.

Lessons

The main lesson from this study is that Borchgrevink et al. (2003) show that it is possible within a reasonable budget frame to design and implement an impact evaluation of high quality that

- collects data from participants and non-participants and constructs a counterfactual,
- makes use of both qualitative and quantitative information, and
- provides analysis that probably separates out the true impact of the programme on beneficiaries.

However, it is also important to recognise that

- matching the programme participants and non-participants along more dimensions reduces the possibility of having unobservable characteristics that influence the assessed impact of the programme (which would skew results),
- displaying the descriptive statistics of the two groups improves transparency of the approach,
- using the full range of the collected data to assess the impacts usually improves confidence in the conclusions drawn, and
- collecting data on quantitative measures yields the opportunity to estimate the quantitative effects of the programme.

As a general point of view, we would argue that it is a better use of money to commission one such evaluation with scope to draw interesting conclusions on impacts than, say, two smaller evaluations unable to reach robust conclusions on impact.

4.2 Marginalised groups and empowerment – a study of DECSI in Ethiopia

Borchgrevink et al. (2005) analyse the impacts on marginal groups of the Ethiopian micro-credit scheme DECSI, which is the same programme discussed in the previous section. Their main objective was to assess to what degree the programme contributed to the empowerment of women and youth in the programme area of DECSI. This was not possible to infer from the previous impact analysis, Borchgrevink et al. (2003), so another round of data collection was initiated for this project. This study is chosen because the study team believes it shows how unintended results of aid programmes can be accounted for, and to illustrate the importance of

taking into account the impacts of a programme on markets. The study was commissioned by the Association of Ethiopian Microfinance Institutions (AEMFI) and Norwegian People's Aid (NPA), but the ToR is not attached to the study. The budget of the study was approximately NOK 1.2 million with 11 weeks for the Norwegian expert and 11 weeks for each of the three Ethiopian team members.

How does the evaluation report discuss the methodology for impact evaluation?
The report contains detailed information on the methodology and the study is designed in a way that allows for the construction of a counterfactual. However, the report is not as thorough as the previous study in elaborating on the methodological issues. The first phase of the data collection consisted of a quantitative household survey containing 520 households in the area of operation of DECSI. The total sample contains 39 % clients of DECSI, 17 % incoming clients, 28 % non-clients and 17 % ex-clients. The data were then processed and analysed in order to develop hypotheses for the second phase of data collection, which consisted of qualitative interviews and focus group discussions in selected localities.

First note, in contrast to the study assessed above, that the report presents the descriptive statistics for the various parameters of interest. This gives a good overview of the characteristics of the different groups and may be used for developing hypotheses about how these would impact on the outcome of the credit programme. For example, would credit have different impacts on a youth-headed household as compared to an adult-headed household? If so, one has to compare the impacts of the credit programme for youth households participating in the DECSI programme with non-participating youth households. If the area of residence also matters to the outcome, for example urban/rural, then youth clients in the urban areas must be compared with youth non-clients in urban areas, and so on.

What conclusions are made with respect to impact?
The report states that 64 % of the clients report that their living conditions have improved during the last five years. Seen in isolation, i.e. without assessing how non-participants experienced the last five years, this could have been interpreted to be a great improvement resulting from the DECSI programme. However, the authors emphasise that 55 % of the non-clients also report that their life has been improving during the last years, and that the difference between them can be taken as the impact of the programme. In other words, 9 % of the DECSI participants have improved their life during the last five years due, most likely, to the impact of the DECSI programme.

The authors also find two unintended side effects of the DECSI programme – indebtedness and deterioration of living standard, and that debtors withdraw their children from school to engage them in the income-generating activity financed by the loan.

The study is designed for the purpose of also including ex-clients of the programme to assess possible negative outcomes for DECSI participants. Contrary to expectations, however, they find that none of the ex-clients report that indebtedness arising from DECSI credit was the reason for their decline in living standard during the last five years. Moreover, 14.3 % of the current clients report a deterioration of living standard due to indebtedness from DECSI credit. Hence, the debtors are still in DECSI, which means that they are not excluded due to the large debt. This could imply two things. First, the debtors' decrease in living standard is due to the low yield on the investment financed by the loan. Hence, the DECSI client may be servicing the debt with other income sources. Second, the client is on the verge of being excluded from the programme. One lesson from this work is nevertheless that a substantial share of individuals participating in a programme may experience a decrease in living standard due to their inclusion in the programme, and this must be addressed in the evaluation.

The second side-effect of the DECSI programme is a labour market response from new labour-intensive activities being financed by the credit. The evaluation finds that there is a markedly higher primary school drop-out rate for DECSI clients as compared to non-clients.

What is the analytical basis for these conclusions?
The analytical basis for the assessments is twofold. First, quantitative data are summarised, descriptively compared between the different groups of interest and in some instances also

econometrically analysed. The background for the econometric analysis is not presented, and hence it is not possible to discuss the basis for this analytical approach. Second, the qualitative information collected is used to dig deeper into the results, in particular into the unintended effects, and case stories are used to illustrate the particular patterns that may lie behind the aggregated figures.

The particular reason for rather high co-variance between participants and non-participants when it comes to improvements in living conditions are not discussed in the report. One source could be natural variations in agriculture, which impacts on most people in such areas. If agriculture has improved in recent years (there was a drought in 2002) many farmers will report that their situation has improved. This must then be separated from the effect of DECSI, and hence it is crucial to compare the trajectory of the DECSI clients with that of the non-clients. We find the difference between the groups to be statistically significant,[7] so the DECSI programme probably had a positive effect. However, the main lesson is that without comparison groups the evaluator may get huge errors in making inferences about programme impact. Also, it is important to note that the samples are sufficiently large to allow for testing of statistical significance.

The report cites literature on the widely studied issue that debtors can, due to unforeseen negative shocks, be caught in debt traps, i.e. a situation where they are never able to service their debt. This would imply that the debtor is excluded from the programme, which in addition to credit often provides other important inputs for the person's economic activity, and could result in a sharp deterioration in living conditions. Patterns of failure in programme participation leading to a worse outcome for the client than not participating are not unique to credit programmes. The authors do refer to other studies in discussing the patterns found in the DECSI areas. It remains a puzzle, however, why none of the ex-clients reported that the DECSI loan was responsible for the deterioration in their living standard while many current clients reported that it was.

The authors suggest two explanations for the finding that DECSI client households have a higher primary school drop-out rate than non-client households. The first is provided by comparing the descriptive statistics between the two groups. The number of extremely poor is higher in the client group and this could in part explain the pattern since the poor are more vulnerable to adverse shocks, which in turn oblige them to withdraw their children from school. The second involves the labour market. The purpose of the credit is to finance productive enterprise. If successful, one would expect this to increase labour demand. Using both quantitative and qualitative techniques, the authors find that a larger share of the non-clients as compared to clients had children dropping out of school due to the need for child labour. However, as the qualitative investigations reveal that extremely poor people are more inclined to take their children out of school, the authors should have compared drop-out rates between extremely poor clients and extremely poor non-clients to see if there were marked differences in responses to whether child labour was important for dropping out. Nevertheless, the approach is a good example of how triangulation may give a more thorough foundation for reaching conclusions on impacts.

Lessons

Several specific lessons emerge from this impact evaluation:

- probable unintended effects should be discussed in the ToR,
- the design of the evaluation should take into account unintended effects both for participants and for non-participants and former participants in the aid programme,
- impacts that work through the market may be important to the outcome of the programme, but can also be difficult to trace, and
- triangulation of methodologies may give a more thorough foundation for conclusions about impacts.

4.3 Evaluation of Norwegian support to psycho-social projects in Bosnia-Herzegovina and the Caucasus

This is an evaluation of projects run by Norwegian NGOs to help war-affected people heal traumas. The main target groups of these programmes have been women and children, and the

7 Using a standard two-sample test of proportion, we find that the difference between the responses is statistically significant at the 9 % level.

psycho-social rehabilitation work aimed at individuals as well as social relations in the war-torn societies. The evaluators are asked "to find out to what extent these programmes have had an effect and whether such programmes should be supported in the future" with a total budget of approximately NOK 1.1 million of which NOK 820.000 were to cover salary for the consultants (1338 hours).

The ToR specifies a long list of criteria defining "qualitative impact", several of which concern aspects of implementation and not impacts as such. As regards impacts, the evaluation was to investigate the extent to which the programmes have:
- "helped ignite other initiatives among the users"
- "helped the users return to normal functions"
- "helped repair relationships"
- "prevented children from being recruited into the armed forces".

Furthermore, the ToR indicates elements of a methodology for assessing impact by stating that:
- "the aims and objectives for each programme should be used as a baseline against which the effect can be measured";
- "in order to asses the effect on individual participants in the programmes, their psycho-social state as they entered the programme should be used as a baseline (if at all available)".

How does the evaluation report discuss methodology for impact assessment?
There is a small section on methodology that merely gives a summary record of interviewees and secondary sources consulted. Thus, the report does not discuss how to conduct the analysis of impacts identified in the ToR, which is surprising given the team's composition (four psychologists/psychiatrists and one political scientist) and the obvious need for some form of counterfactual research design in this case. How can the effects of the programmes on the change processes indicated above be isolated?

What conclusions are made with respect to impact?
The report has no explicit terminology with regard to concepts such as 'result', 'effect' and 'impact', and the presentation of findings tends to mix conclusions with respect to management and approaches used, outputs and effects in a rather unstructured manner. In fact, the concluding chapter contains few statements regarding impact. The report states conclusions about the "overall positive results of the six psycho-social projects reviewed", but there is no reference to the specific impacts mentioned in the ToR.

There is some discussion about the selection of beneficiaries in the programmes. Who were attracted to and able to enlist themselves in these projects? Were they among the most traumatised? In some project surveys none of the women reported having been raped, but it is likely that many would have been reluctant to disclose such information, being afraid of the social stigma involved. Rape trauma was a major focus of international attention, but there seems to be some uncertainty to what extent the projects reached out to those affected.

The project's roles in facilitating peace processes and promoting post-conflict stability are discussed at a theoretical level with statements such as:
- "In the projects evaluated, all of them could be said to make contributions which have the potential to create a peaceful society…"
- "The evaluated projects could be said to contribute to social stability through creating opportunities in safe environments for the re-establishment of social networks, building skills such as literacy and trust among people."

What is the analytical basis for these conclusions?
The conclusion about overall positive results is based solely on statements from project staff and beneficiaries. Participants in the projects "emphatically endorsed the significant value" of the attention, support and care they were offered. We have no reason to question this as an overall impression, but a more structured and randomly sampled user survey might have given a more nuanced picture.

With respect to impacts on individual participants, the report refers to attempts by some projects to track changes in people's traumatic reactions before and after inclusion in the

programmes, but is critical of the value of this methodology for evaluating psycho-social improvement. Nevertheless, the report seems to rely solely on such project data for its own general assessment.

As to wider societal impacts, the analysis is merely theoretical and reiterates the basic assumptions about impact that justified these projects in the first place. The only critical assessment deriving from this theoretical analysis is the low level of genuine beneficiary participation in the projects – they "remain passive beneficiaries rather than active participants" (p.10).

Overall, the evaluation fails to answer several of the questions in the ToR with respect to impact. There is no explanation of this failure, nor any attempt to describe the methodological approach. We find there is a weak analytical basis and flimsy presentation of data supporting the conclusions made. Surprisingly, there has been no attempt to introduce comparison groups, by interviewing non-beneficiaries or looking at similar programmes supported by other agencies. Apparently, there is no methodological approach to the selection of interviewees – either randomised or structured sampling.

Lessons
The object of evaluation in this case represents programmes delivering certain services aimed at specific target groups. Norway contributed to the financing of the programmes.

- We would have expected the evaluators to have discussed more explicitly how to assess the aid element versus broad programme effectiveness. The impacts of the aid on local institution building, for instance, are not analysed.
- This is a typical case where principles of programme evaluation, as presented in section 2, would be highly relevant.
- We find that the evaluators have not been able to get a handle on the issue of impact, and by and large shy away from concrete assessments. Furthermore, there is no attempt to assess critically the assumptions about impact made in project documents.
- The ToR are fairly clear about the need to analyse impact, but apparently the client and the consultant seem not to have had a thorough discussion of methodological implications and what could be achieved in practice, given data availability and resource constraints.

4.4 Evaluation of Norwegian support to FORUT, Sri Lanka

Baklien et al. (2004a) assess the impacts of the work of FORUT in Sri Lanka, a project carried out in parallel with an assessment of Save the Children Norway's (SCN) work in Ethiopia (see below). A separate report gives the methodological background for the study (Baklien et al. 2004b). The total budget for the study, including the contributions to the methodology report and the local partners, is approximately NOK 1,180 000.

FORUT's programme in Sri Lanka has implemented projects aiming to strengthen civil society, contribute to capacity building, reduce poverty and promote democratisation and good governance. The study was commissioned by Norad and the ToR specifies three broad areas selected for an impact evaluation: democracy, human rights and poverty reduction. The ToR further states that the objective of the study is to carry out an impact evaluation with respect to these three outcomes: "In order to address the impacts of the organisations' work, it is necessary to conduct a more thorough and long-lasting study than usually undertaken in evaluations. This study will be carried out over a period over two years." Even though the ToR is very precise as to the purposes of assessing impacts, they are ambiguous and vague in the section that specifies the concrete approach to be taken under each of the three categories of impact. This is particularly so when it comes to the impacts on poverty reduction; in addition, most of the statements also direct the assessment to focus on outputs instead of impacts.

How does the evaluation report discuss methodology for impact assessment?
Despite the clear aim of the ToR to conduct an impact analysis, and its specification of the outputs to be described, the authors define their task in a third way, i.e. not to assess outputs or impacts:

> "…this is not an evaluation of FORUT and SCN as such, but rather a study, or an assessment that aims to describe the type of impact that NGOs may have in terms of achieving the

> objectives set out in the Norwegian guidelines for support to NGOs, and suggest ways in which impact can be reported, the interventions that we study are "best practice" example of interventions. In other words, the interventions have been chosen because they are likely to illustrate *intended* impacts of FORUT's interventions." Baklien et al. (2004b) p. 15.

The authors' aims of "describing the type of impact that NGOs may have" and "illustrat[ing] *intended* impacts" may be interpreted as corresponding to a paragraph in the methods section of the ToR: "The study will use a case study approach and select illustrative activities of the Norwegian organisations and partner organisations." However, selecting interventions that illustrate the impacts (as the authors do), and not the activities (as the ToR specifies), is not informative for the overall purpose of the study. Moreover, we have not found any information on why the aim was altered in this manner, and why this was accepted by Norad.

What conclusions are made with respect to impact?
Note first that, as for the DECSI study above, we confine our assessment of the study to the impacts on poverty. The poverty component in FORUT's programme comprises three elements: savings and credit programmes, agricultural loans and provision of services (health, education and infrastructure). In the initial phase of the authors' assessment, four workshops were carried out – one with partner organisations, one with FORUT staff, one with beneficiaries and one with government officials/political authorities. One aim of these workshops was to discuss possible impacts of FORUT's work.

The workshop participants reported that a range of improvements in socio-economic status had taken place. The study further asserts that FORUT's savings and credit programme are "more of a coping strategy to alleviate symptoms of poverty than a development strategy to eradicate the root causes of poverty. The main reason is that loans … are relatively small, averaging Rs. 5000." However, the *average* loan is, according to the report, sufficient to buy half a cow, which in turn indicates that a poor household could buy many productive assets for this amount. The study also states that the investments are in activities with marginal economic returns.

Other conclusions, which the authors claim serve the purpose of illustrating the intended impacts, are confined to insignificant general statements, for example: "Agricultural loans schemes have helped farmers to increase incomes, ensure food security and enhance the stability of household incomes"; and "financial support to CBOs for constructing drinking water wells and toilets has improved the water and sanitation conditions in the villages."

What is the analytical basis for the conclusions?
Carrying out broad workshops with the focus on discussing possible impacts of the programme under scrutiny is a sound starting point for an impact analysis. However, much of the empirical background used for the conclusions seems to be taken from the workshops. Taking into account that FORUT had an annual budget of NOK 24.6 million in 2003 and that the interventions were chosen because they were likely to illustrate the intended effect of FORUT's work, it is perhaps not a surprise that some workshop participants reported socio-economic improvements. However, the question of interest is how many of the FORUT clients actually did benefit from the programmes, how much did they benefit and at what cost. No information on these basic questions is provided in what was labelled by Norad a long-term impact evaluation. The outputs from the programmes are listed in Appendix 1 in the report (Baklien et al. 2004a) and in Appendix 5 in the methodology report (Baklien et al. 2004b), but no use of this information can be traced in the main report when it comes to impacts on poverty.

The assessment of the poverty-reducing impacts of FORUT's work does not point to any sources when it draws its conclusions. Despite the firm conclusions on the impacts of loans on economic returns, food security and income, no overview of the profits of the different investments is given, no consumption or nutrition figures are discussed, no information on income-smoothing strategies in use by FORUT clients is presented, and discussion of the chain from FORUT's inputs to increased incomes is absent. Moreover, not even the outputs from the programme are assessed.

Similar grave shortcomings are found in the brief discussion of whether the credit programme actually benefits the poor. It is a rather easy task to gather information to get an indication of

how the programme excludes or includes the poor. However, no such efforts are undertaken by the authors despite the explicit aim of the ToR to investigate such patterns.

Lessons

- The main lesson is that the vagueness of the ToR yields the opportunity for the evaluators to define the assessment into something that does not correspond to the aims of the ToR. This issue is further exemplified in the next section.
- Moreover, when it comes to the impact assessment component in the study, it is our opinion that the authors of the FORUT study did not have the competence for carrying out this part of the assignment. This judgement is based upon the lack of analytical basis for the conclusions that were drawn on impacts of the poverty component of FORUT's programme.

4.5 Evaluation of Save the Children Norway (SCN) in Ethiopia

The Save the Children Norway (SCN) programme under evaluation is the Ethiopia programme for the period 2002-2005, of which 64 % of the resources went to education, 16 % to child rights, 11 % to HIV/AIDS prevention, 5 % to poverty reduction and 3 % to disabled children. The budget for 2003 amounted to USD 3.5 million, of which 85 % went to project support and 15 % to SCN administrative costs. A quarter of the funds of the operational budget were targeted towards emergency relief interventions.

The study of the impact of SCN (Helland 2004) was commissioned jointly with the FORUT study assessed above and shared the same ToR. Hence, it was specified as an impact evaluation to be carried out focusing on democracy, human rights and poverty reduction. Recall also that the ToR then proceeded to list outputs as the issues that the evaluation was to focus on. These were whether partner organisations included the poor as members, board members, beneficiaries or employees in the projects, how partner organisations linked up with other local institutions, and whether SCN helped poor people to access resources. The only mention in the ToR of any indicator that comes close to serving as an indicator of well-being was whether the projects enabled the poor to "access the necessary resources". As indicated in the previous section, the vagueness in the ToR may explain the different interpretations of what was to be analysed, and in this evaluation it is not clear what "necessary resources" really means in this setting. Finally, the ToR specifies that in Ethiopia the focus was to be on the situation among the poorest children in both rural and urban contexts.

The methodology section of the ToR specifies that the study is intended to be a learning process for SCN and partner organisations, that the study's major focus is on groups, organisations and institutions, but also on indicating impacts on individuals. However, no methodology is specified other than the statement that the study should use a case study approach, which in itself is not very informative. The section describes what is to be assessed through the approach, but nothing about how this assessment is to be conducted. The evaluation of SCN, including the contribution to the methodology report (Baklien et al. (2004b), had a budget of NOK 1,020 000.

How does the evaluation report discuss methodology for impact evaluation?

Helland (2004) underlines in the beginning of the methodology section of the evaluation that the report is not intended to be an impact evaluation, but rather to present a brief account of the activities of SCN during 2002-2005 in Ethiopia. Nevertheless, the main approach of the author is to place the intentions behind the SCN's efforts, and the activities that derive from them, into the Ethiopian context and to "trace outcomes of project implementation in terms of benefits for the intended target groups." This is, however, carried out in such a general way and seemingly without primary data collection that it is more precise to describe the review using the first formulation, i.e. as an account of the SCN's activities.

What conclusions are made with respect to impact?

Turning to chapter 8 in the report, "Impact on democratization, human rights and poverty reduction", we are not offered any tracing of "outcomes of project implementation in terms of benefits for the intended target groups." A very brief comment on the success of the educational efforts is provided, but is not documented empirically in any way.[8] A brief

8 The assertion that the educational efforts of SCN have been successful is also stated in other chapters, but substantiation is lacking.

comment on poverty reduction is also included, but instead of discussing expected impacts on poverty by the SCN's efforts targeted towards poor households (5% of the SCN budget – USD 175 000 for 2003), the author chooses instead to contrast SCN's efforts with Ethiopia's poverty challenge at the national level. This deviates again from the author's specified approach of tracking impacts to the intended target groups, and the purpose of relating a minor poverty-reducing effort to national poverty levels in a country with 65 million people seems unclear. One could argue that the ToR invited the evaluator to include possible national impacts, but this implies that the inputs studied are of a magnitude that might possibly have national implications.

What is the analytical basis for the conclusions?
While the account of SCN activities is rich in historical detail and thorough in the description of the projects and the context they are embedded in, the analysis is usually confined to short, general descriptions of the outputs and efforts. No substantial empirically based analysis of impacts on beneficiaries is carried out, and no data are used for the assessments made. Hence, the report corresponds to the author's statement that it is a brief account of recent SCN activities.

Lessons
It is our the opinion that the conflicting statements in the ToR when it comes to whether the commission was an impact evaluation or an account of NGO activities, and the vagueness in stating the methodology, led to the approach of the consultant. The lack of focus on concrete impacts of the SCN's activities is nevertheless also in part a conscious choice from the consultant's side, and we pose the question whether this was tacitly approved by the commissioning body or whether it was a deliberate change that occurred during the evolution of the consultancy work. Hence, the specific lessons are to

- avoid conflicting statements at the different levels of the ToR (overall aim, specific issues and methodology), and to
- ensure that the methodology section concretely specifies methodologies (i.e. qualitative versus quantitative approaches, what type of quantitative and qualitative methodology is acceptable, primary or secondary data, descriptive versus analytic, reliance on literature in the assessment etc.).

4.6 Evaluation of the Tanzania-Norway Development Cooperation 1994-1997

The ToR states that (our emphasis):

> "The main purposes … is a *systematic assessment of the results* and experiences derived from the total development cooperation during the period 1994-97"
>
> "… whether the *degree to which goals were achieved*, is in reasonable proportion to the use of resources"
>
> "… the evaluation should as far as possible assess the degree to which Norwegian development assistance may have *contributed to sustainable development* during this period."

The ToR makes no use of the term 'impact', whereas 'results' is frequently used, and the evaluators are asked to assess results achieved with reference both to the overarching goal of the cooperation with Tanzania – "the promotion of a sustainable economy that is not dependent on development assistance" – and to a number of subsidiary goals, such as expanding infrastructure, strengthening political reform and improving environmental management. The evaluation was to be carried out mainly on the basis of secondary data, augmented by interviews of key stakeholders and a short visit to the country. In terms of resources, the study had a total budget of approximately NOK 1.1 million where research costs accounted for NOK 0.9 million (1873 hours).

How does the evaluation report discuss methodology for impact assessment?
There is no section explicitly presenting the methodology used, but the study devotes a whole chapter to 'results'. The report broadly distinguishes between two levels – programme level and activity level. The former refers to sectors and cross-cutting issues in Tanzania's development, while the latter refers to Norwegian-funded projects.

What conclusions are made with respect to impact?
The impact of Norwegian aid at the *programme* level is discussed in rather general terms, noting that "impact is … difficult to evaluate due to a lack of a more focused outcome objective. Spreading resources along a number of different activities the way Norway's programme has been structured necessarily means that impact assessments will always be difficult if not impossible to carry out" (p. 106). The study basically says that impacts at programme level cannot be discerned because the relative role/size of Norwegian aid is too small and too dispersed. But it also argues that sluggish development in Tanzania is evidence of a low aid impact, as illustrated by the following quote: "Norway had implicitly an extremely ambitious agenda: to contribute to economic and political transformations – that is, structural change. As has been noted in the previous chapters, a number of these change processes have moved very slowly and thus no discernible impact has taken place" (p. 104).

The two statements above seem to represent a contradiction. Is it possible both to conclude that the impacts of Norwegian aid at the programme level cannot be identified, while at the same time argue that slow development is evidence of low aid impact? The answer might be in the affirmative if the latter refers to aid generally, but requires a different analytical perspective. What is important in this context is that the ToR ought to have specified the need to assess the degree to which results were achieved, in reasonable proportion to the use of resources, and outlined methodological implications.

At the *activity* level the report makes an interesting attempt at getting an overview based on a systematic assessment of close to all major activities financed by Norway in the period. The evaluators, based on a review of reports, interviews and their own assessments, rated the projects according to 19 evaluation criteria, one of which was 'impact':
- In terms of impact, 72 projects (of the 90 reviewed) were rated 0 to 3 – being equivalent to "unsatisfactory", "poor", "good", "excellent". 6 of these projects were rated "excellent" and only 1 project was rated "unsatisfactory", while 41 had made "good" impact and 24 "poor" impact. In other words, 65% of activities were considered as at least having had "good" impact.
- In terms of overall rating, combining the different criteria, this was done for 67 of the activities, and as many as 81% received the rating "good" or better.

What is the analytical basis for these conclusions?
There is an underlying argument in the text that overall developmental trends, whether sluggish or positive, are taken as corresponding evidence of aid effectiveness. However, the study makes no attempt to explain what the links are between aid inputs and trends in macro impacts. Nor is this assumed relationship between development trends and aid corroborated by the findings on the activity level.

The methodological problems of validity and subjectivity of individual rating and in creating overall ratings are carefully discussed. Any attempt to condense complex social and economic processes into simple numeric indicators is fraught with difficulties, but the approach introduces some uniformity in the way information is interpreted and the results matrix yielded some interesting findings. Three interesting observations can be made from the findings in the study:
- Comparing programme and activity level, we find that the assessment of individual activities is far more positive than what is said about macro trends. This may well be the true picture – individual activities may well succeed in an adverse environment, but may partly be explained by a bias in the aid industry to be optimistic and that the evaluators base their assessment mostly on the self-reporting of stakeholders directly involved.
- This just serves to underscore that we cannot judge aid impact purely as a factor of changes in macro indicators. Aid interventions need their own set of evaluation criteria.
- Infrastructure projects generally score quite well. Projects with well-defined output targets are likely to be easier to manage, and since results are more tangible the assessment of impact tends to be more positive.

Lessons
This is an example of an evaluation typical of development aid cooperation. With limited resources, consultants are asked to assess the effect of Norwegian aid to a particular recipient, in this case a partner country, and there are several other cases of similar evaluations looking

at aid to particular NGOs and UN organisations. Typically, the recipient uses the assistance for a wide range of activities and it becomes a formidable challenge to assess impact. The tendency, therefore, in such evaluations is to emphasise primarily on how the aid partnership functions.

There is no attempt in this study to analyse causality. There is no attempt to assess the counterfactual, or to establish relevant comparisons. The report provides strategic advice on a broad range of issues based largely on hegemonic ideas at the time about the need for less fragmentation, more coordination and stronger recipient ownership. The report provides no insight, however, on the extent to which aid effectiveness underpins the advice provided. Three concrete lessons can be drawn:

- Evaluations designed in this way can be useful inputs to strategic discussions, but cannot yield reliable analysis of impacts.
- This requires a more selective approach, through which certain assumptions about impact are tested by comparative methods and the study of changes over time.
- The ToR needs to define more clearly terms like 'impact' and 'results'. As the study above shows, some critical questions with respect to 'results' emerges. Firstly, where goals are more physically perceived achievement tends to be higher, and secondly, aid is often perceived as successful despite overall negative trends. Both questions warrant further scrutiny of how we define and assess results.

4.7 Evaluation of Norwegian support to FIFAMANOR, Madagascar

The Norwegian support to FIFAMANOR, the national centre for research and extension services in agriculture and husbandry in Madagascar, amounted to 56.5 million NOK during the programme period that is evaluated. This amounted to 50 % of the total running costs of the institution. The Norges Vel has had the overall financial and reporting responsibility to Norad. The overall aim of the Norwegian contribution is to improve the living standards of farmers in the Vakinankaratra region of the High Plateau of Madagascar. The specific programme supported by Norad includes 8 components of different farm inputs which were to be promoted or developed and have the main focus during the implementation.[9]

The ToR specifies that five broad issues are to be covered, which are identical to the DAC criteria – efficiency, effectiveness, relevance, impact and sustainability – all with a range of specific problems to be assessed. However, the ToR confines the methodology for the review to studying programme documents and interviewing key stakeholders. The review is expected to cover an assessment of the programme's impact and the effects of the long-term support to FIFAMANOR. The ToR further elaborates that the discussion of impacts should include effects at the household, community, institutional and policy levels. Finally, the ToR states that the assessment should cover possible effects on gender, health, nutrition and staff capacity/qualifications, including extension workers. The budget for the study is limited to NOK 530 000 with a time allocation of approximately 3 weeks for fieldwork and 2 weeks for preparation and report writing, all for the Norwegian consultants.

How does the report discuss methodology for impact assessment?
There is no discussion about methodology for impact assessment in the report – there is in fact only one sentence about approach: "The methodology used in this evaluation is based on reviewing existing documentation and interviews."

What conclusions are made with respect to impact?
Several impacts or indications of impacts on the people's living standards from the FIFAMANOR projects are noted in the report:

- "FIFAMANOR has made a valuable contribution to promote milk production in Vakinankaratra. Milk production has also contributed to income diversification and to the development of a private diary industry…. Many of the diary farmers have increased their standard of living".
- "… application of new higher yielding varieties developed by FIFAMANOR has contributed to substantially higher yields".

9 These components were 1) grain production; 2) potatoes, sweet potatoes and other tubers; 3) milk production, fodder production and genetic improvement of the livestock; 4) a social programme for the integration of women into productive activities; 5) a competence centre for research related to the above; 6) support to farmers' organisations; (7) effective use of scientific and technical knowledge; and 8) a training centre for farmers and extension workers.

- "… research and development activities of FIFAMANOR have contributed to improved human nutrition in the region".
- "The activity with the biggest development impact may be FIFAMANOR's sale of seed potatoes".
- "FIFAMANOR's creation of women and men's associations have greatly promoted development in the region".

The report also states that: "Considering the large efforts of FIFAMANOR to assist these groups, there is little doubt that these farmers have strengthened their incomes, capacities, knowledge and practices." Finally, on the extension component the report states that there were 5400 direct beneficiaries of FIFAMANOR's efforts, and an additional 500 farmers that were followed up more closely.

What is the analytical basis for these conclusions?
There is no analytical basis in the evaluation report for the conclusions stated above. No attempts to substantiate the assertions are made and the reader may believe that the conclusions are not derived from empirical facts. We may only assume that the claims made above are either from the documents scrutinised or from the stakeholders interviewed.

No attempt is made to assess the magnitude of the input (56.5 million NOK) against any impacts and whether these impacts are in a reasonable proportion to what could have been expected from the perspective of efficiency. It is more disturbing, however, that there is no attempt to give a description of outputs that are relevant for the living standards of the households. The consultants state that "the annual report gives a very good overview over the activities of FIFAMANOR and the same indicators are measured over several years. This gives a good overview of the development of the projects and the results obtained." However, the consultants' own report does not contain any such information despite the clear request in the ToR to provide such information, and no substantial information is provided that could have given suggestions of probable impacts on people's living standard. The report also states that the monitoring and evaluation system developed for FIFAMANOR does not provide sufficient information to assess the impact of the project on the living standards of the project beneficiaries. This was perhaps why the ToR specified that it was the consultant's task to discuss the impacts.

The importance of taking the response of the market into account is illustrated in the FIFAMANOR study. Much of the support has been used to stimulate the production of milk in the district, but the milk producers are now complaining about the low prices of milk and cows. Hence, if the support to milk production has driven prices down, it is evident that those milk producers that are not receiving support would be worse off from the intervention. From the figures in the study, there are 10 000 households with milk cows in the district, but the consultants do not come up with figures on how many of these have received support from FIFAMANOR, what type of support and how it has impacted on the relevant markets (and in turn on living conditions for beneficiaries and non-beneficiaries).

Under the time frame of the study, it is not possible to cover all five issues raised in the ToR in a thorough manner. Hence, our discussion of one of the issues of this study, i.e. the "relevance and impact" component, must be seen in light of the limited time at the evaluators' disposal. Moreover, based on the methodology section of the ToR it is not surprising that there is no analysis of any impact of the support to FIFAMANOR on living conditions in the district. This implies that our findings are also relevant for the commissioning body.

Lessons
Summing up, it is the study team's opinion that the review would have been substantially improved within the resource frame specified in the ToR by
- providing data on the outputs of the programme components;
- relating these to the resources that had been used, and
- discussing the likely impact of these outputs on the living standards of both beneficiaries and non-beneficiaries of the project. For example, what was the economic value of the support to the average milk producer and did this support come as an investment so that it also contributed to an improved living standard in the longer run?

However, if the aims of the ToR with respect to "assessment of programme impact" (p. 31) were to be fulfilled, then a substantially different approach should have been taken. The methodology section should then have included primary data collection with the aim of assembling household data on beneficiaries and non-beneficiaries to construct a counterfactual. This would have required a substantially larger budget for the assessment. Hence,

- the aims of the ToR with respect to impact evaluation must be coupled with adequate resources for conducting the study, and
- the methodology must be specified according to this aim.

5. Practical Recommendations for Impact Evaluations

Building on the presentation of methodological challenges (Section 2) and the seven cases reviewed (Section 4) what follows below is our attempt to distil practical recommendations for both commissioning agencies and evaluators.

The ToR must be internally consistent
The ToR must be internally consistent to avoid differing interpretations of the assignment. The study team finds that this was not the case for several of the evaluations reviewed, and we see the need for a more rigorous use of terminology with a precise definition of the key concepts employed. This implies doing more than referring to the DAC Guidelines, and also requires a critical discussion, while preparing the ToR, as to whether all dimensions of an evaluation can realistically be included. This goes in particular for 'impact'.

It is of particular importance that the methodology section is clear and concise as for many evaluations this will be decisive for what kind of evaluation is conducted and hence what type of analysis is undertaken. The methodology section should be concrete on the particular methodologies to be used and specify whether qualitative and quantitative approaches should be applied, what type of quantitative and qualitative methodology is acceptable – primary or secondary data, descriptive versus analytic, reliance on literature in the assessment etc.

If the ToR is internally inconsistent, the evaluator should pay close attention to the methodology section in light of the resources provided for the evaluation. Where the ToR asks for an impact evaluation but confines the methodology to stakeholder interviews within a narrow budget frame, it must be noted that this will seldom be sufficient for an evidence-based assessment. Where there is a mismatch between the ToR's overall goals, scope, concrete issues to be covered, methodology and budget, the exact intentions of the study should be clarified before the evaluation takes place. We acknowledge that where procurement guidelines require public tendering this might pose a problem, and hence commissioning agencies should seek advice on methodology prior to finalising the ToR.

In cases where the ToR concretises the intended impacts which the study is to assess, it should also ask for an assessment of unintended effects.

It is notable that few robust aid impact evaluations have been carried out by Norwegian research institutions and consultancy firms during the review period 1996-2007. Furthermore, where impact is included there is a tendency to cover a wide range of aspects. The necessary resources for investigating impacts according to the aims of the ToR are mostly not provided, which in turn yields a superficial treatment of the impact evaluation components of the study. We recommend that commissioning bodies should limit the number of impact evaluations, and rather provide an adequate budget for studies which are to assess impact. One lesson from this review, however, is that it is possible within a reasonable budget to design and implement an impact evaluation of acceptable quality.

Provide a strong analytical basis for conclusions on impacts
On theory
Developing a theory, or a logical chain, on how the programme is expected to impact on the participants and non-participants can give important pointers to where the evaluator should focus resources for studying impacts. A good starting point for this work would be to conduct a workshop with key participants (clients, non-clients, programme staff members, donors, local authorities etc.) in order to spell out the expected sequences from inputs to impacts (see Figure 1 above) and the distribution of these impacts across different groups of beneficiaries (youth, women, poor, disabled, entrepreneurs, educated etc.). The same requirements for

building theoretically formulated hypotheses regarding impacts apply also to institutional impacts, i.e. where the aid is intended to increase skill levels, efficiency, innovation etc.

On methodology
For evaluations where the purpose is to assess the impacts of a specific programme assigned exclusively to certain observable units such as individuals, households, villages or regions, it is necessary for the evaluator to construct a counterfactual. This involves the collection of data from both beneficiaries and non-beneficiaries, or the use of secondary data if these are appropriate for the particular research question, or ideally, to use both data sources to cross-check results.

The ideal situation is where there is reliable baseline information – i.e. pre-intervention information on key characteristics of beneficiary and comparison groups. Unfortunately, this is usually not the case.

Even without baselines, data collection for impact evaluation can be greatly improved if:
- The programme beneficiaries interviewed are randomly selected from the total population of programme beneficiaries (drawn randomly from lists prepared by the administrators of an intervention).
- A comparison group for interviews is randomly selected from non-beneficiaries, excluding ex-beneficiaries. However, if programme beneficiaries were selected in a manner so that the characteristics of this group also influenced programme impact, then the comparison group must be selected so that these particular characteristics are similar for the two groups.
- It is important to identify and include ex-beneficiaries in analysing unintended effects of the programme.
- Contextual factors are controlled for through participatory approaches.
- The sample size of the data collected is sufficient for statistical analysis to be performed. Determining sample size and performing the statistical analysis requires specific knowledge of the relevant methodologies.
- The evaluators should look for alternative and simpler indicators where more complicated ones cannot be made available – e.g. use assets and quality of housing as indicators of well-being rather than income and consumption data.

This approach is strengthened by making use of qualitative information, for example to discuss whether the comparison group is similar to the beneficiary group. It is important to validate the aggregated results of a quantitative analysis through participatory methods. Hence, the best result is probably obtained by triangulation of methodologies.

Furthermore, potential selection biases may have to be addressed. One approach is to match beneficiaries and non-beneficiaries along a range of different characteristics. This approach deals with the selection problem in cases where unobserved beneficiary characteristics are likely to influence programme impact and it is difficult to find indicators for these characteristics. The issue is to find a comparison group that is as similar as possible to the beneficiary group.

We have found that impact evaluations generally provide rudimentary documentation of the data being used. There is evidently a trade-off between decision-makers' and bureaucrats' appeal for short and crisp reports and principles for scientific documentation, but we want to emphasise that displaying descriptive statistics improves the transparency of the methodological approach.

Keep an eye on unintended effects
The design of the evaluation should take into account unintended effects both for beneficiaries and for non-beneficiaries and former beneficiaries. Impacts that work through the market may encompass both intended and unintended impacts and may be important to the assessment of the programme's overall impacts.

Ensure the right evaluation competency
We would like to emphasise that impact evaluation is a discipline that requires specific knowledge about programme evaluation methodology in addition to advanced analytical skills.

This implies that all impact evaluations should be subjected to a peer review in order to bring an independent perspective to bear on assessing to what extent the commission has been fulfilled and to review the methodological approach used.

Timing
When can one expect the full impacts to emerge? The evaluator and the commissioner should draw on experiences of other similar programmes, both from evaluations and project documentation, to find out when one can expect specific impacts to emerge and the trajectory of the impacts. This may be important as the timing of the evaluation may not be in accordance with the emergence of the full effects.

The best impact evaluations are designed and implemented side-by-side with the programme itself.

References

Agger, I., E. Jareg, J. Mimica and C. C. Reiben (1999): "Evaluation of Norwegian Support to Psycho-Social Projects in Bosnia-Herzegovina and the Caucasus", Norwegian Ministry of Foreign Affairs, Evaluation Report 3.99.

Aune, J. B., M. Skortnes, A. W. Randriamamonjy (2005): "Review of Norwegian support to FIFAMANOR", Norwegian University of Life Sciences, Noragric Report No. 30.

Baker, J. (2000): *Evaluating the impacts of development projects on poverty: A handbook for practioners.* Washington D.C.: World Bank.

Baklien (2004a): "Study of the impact of the work of FORUT in Sri Lanka: Building Civil Society", Norad Evaluation Study 5/2004.

Baklien, B., D. Getachew, M. Haug, J. Helland, C. Weerackody (2004b): Study of the impact of Norwegian NGOs on civil society: FORUT (Sri Lanka) and Save the Children Norway (Ethiopia). Some methodological issues. A report based on the initial phase of the study.

Bamberger, M., J. Rugh and L. Mabry (2006): *Real World Evaluation: Working Under Budget, Time, Data and Political Constraints.* Thousand Oaks, CA: Sage.

Borchgrevink, A., Jo Helle-Valle og Tassew Woldehan (2003): "Credible credit. Impact study of the Dedebit credit and savings institution (DECSI), Tigray, Ethiopia", *NUPI report*, 12.03.2003.

Borchgrevink, A., T. Woldehana, G. Ageba, and W. Teshome (2005): "Marginalized Groups, Credit and Empowerment: A Study of Dedebit Credit & Saving Institution (DECSI) of Tigray", Report Commissioned by Norwegian People's Aid (NPA) and the Association of Ethiopian Microfinance Institutions (AEMFI).

Deaton, A. (1997): *The analysis of household surveys: A microeconometric approach to development policy.* Johns Hopkins/World Bank.

Hatlebakk, M. (2007): "Grameen Bank", Økonomisk Forum (in Norwegian), vol. 61 (4), pp. 14-22.

ECON (1999): "Evaluation of the Tanzania-Norway Development Cooperation 1994-1997", ECON Report no. 10/99.

Helland (2004): "Study of the impact of the work of Save the Children Norway in Ethiopia", Norad Evaluation Study 2/2004.

Ravallion, M. (2001). "The Mystery of the Vanishing Benefits". *World Bank Economic Revie*w. 15(1): 115-140.

Ravallion, M. (2003). "Assessing the Poverty Impact of an Assigned Program". Chapter 5 in Bourguignon, F. and Pereira da Silva, L.A. (eds). *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools.* Washington D.C.: World Bank.

Ravallion, M. (2006). "Evaluating Anti-Poverty Programs". In R.E. Evenson and T. P. Schultz. *Handbook of Development Economics Volume 4*, North-Holland, Amsterdam.

Smith, J. (2000). "A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies." Swiss Journal for Economics and Statistics. 136(3). Available at: www-personal.umich.edu/~econjeff/Papers/smith-neu.pdf

van de Walle, D. (2002): "Choosing rural road investments to help reduce poverty", *World Development*, vol. 30 (4).

van de Walle, D. and D-J. Cratty (2005): "Do aid donors get what they want? Microevidence on fungibility", Policy Research Working Paper 3542, World Bank. Available online at http://econ.worldbank.org/external/default/main?pagePK=64165259&piPK=64165421&theSitePK=469372&menuPK=64216926&entityID=000012009_20050321153125

World Bank (2006): "Conducting quality impact evaluations under budget, time and data constraints", World Bank Independent Evaluation Group, Washington D.C.

EVALUATION REPORTS

3.92 De Private Organisasjonene som Kanal for Norsk Bistand, Fase I

1.93 Internal Learning from Evaluations and Reviews
2.93 Macroeconomic Impacts of Import Support to Tanzania
3.93 Garantiordning for Investeringer i og Eksport til Utviklingsland
4.93 Capacity-Building in Development Cooperation TowardsIntegration and Recipient Responsibility

1.94 Evaluation of World Food Programme
2.94 Evaluation of the Norwegian Junior Expert Programme withUN Organisations

1.95 Technical Cooperation in Transition
2.95 Evaluering av FN-sambandet i Norge
3.95 NGOs as a Channel in Development aid
3A.95 Rapport fra Presentasjonsmøte av «Evalueringen av de Frivillige Organisasjoner»
4.95 Rural Development and Local Govemment in Tanzania
5.95 Integration of Environmental Concerns into Norwegian Bilateral Development Assistance: Policies and Performance

1.96 NORAD's Support of the Remote Area Development Programme (RADP) in Botswana
2.96 Norwegian Development Aid Experiences. A Review of Evaluation Studies 1986–92
3.96 The Norwegian People's Aid Mine Clearance Project in Cambodia
4.96 Democratic Global Civil Governance Report of the 1995 Benchmark Survey of NGOs
5.96 Evaluation of the Yearbook "Human Rights in Developing Countries"

1.97 Evaluation of Norwegian Assistance to Prevent and Control HIV/AIDS
2.97 «Kultursjokk og Korrektiv» – Evaluering av UD/NORADs Studiereiser for Lærere
3.97 Evaluation of Decentralisation and Development
4.97 Evaluation of Norwegian Assistance to Peace, Reconciliation and Rehabilitation in Mozambique
5.97 Aid to Basic Education in Africa – Opportunities and Constraints
6.97 Norwegian Church Aid's Humanitarian and Peace-Making Work in Mali
7.97 Aid as a Tool for Promotion of Human Rights and Democracy: What can Norway do?
8.97 Evaluation of the Nordic Africa Institute, Uppsala
9.97 Evaluation of Norwegian Assistance to Worldview InternationalFoundation
10.97 Review of Norwegian Assistance to IPS
11.97 Evaluation of Norwegian Humanitarian Assistance to the Sudan
12.97 Cooperation for Health DevelopmentWHO's Support to Programmes at Country Level

1.98 "Twinning for Development". Institutional Cooperation between Public Institutions in Norway and the South
2.98 Institutional Cooperation between Sokoine and Norwegian Agricultural Universities
3.98 Development through Institutions? Institutional Development Promoted by Norwegian Private Companies and Consulting Firms
4.98 Development through Institutions? Institutional Development Promoted by Norwegian Non-Governmental Organisations
5.98 Development through Institutions? Institutional Developmentin Norwegian Bilateral Assistance. Synthesis Report
6.98 Managing Good Fortune – Macroeconomic Management and the Role of Aid in Botswana
7.98 The World Bank and Poverty in Africa
8.98 Evaluation of the Norwegian Program for Indigenous Peoples
9.98 Evaluering av Informasjons støtten til RORGene
10.98 Strategy for Assistance to Children in Norwegian Development Cooperation
11.98 Norwegian Assistance to Countries in Conflict
12.98 Evaluation of the Development Cooperation between Norway and Nicaragua
13.98 UNICEF-komiteen i Norge
14.98 Relief Work in Complex Emergencies

1.99 WID/Gender Units and the Experience of Gender Mainstreaming in Multilateral Organisations
2.99 International Planned Parenthood Federation – Policy and Effectiveness at Country and Regional Levels
3.99 Evaluation of Norwegian Support to Psycho-Social Projects in Bosnia-Herzegovina and the Caucasus
4.99 Evaluation of the Tanzania-Norway Development Cooperation1994–1997
5.99 Building African Consulting Capacity
6.99 Aid and Conditionality
7.99 Policies and Strategies for Poverty Reduction in Norwegian Development Aid
8.99 Aid Coordination and Aid Effectiveness
9.99 Evaluation of the United Nations Capital Development Fund (UNCDF)
10.99 Evaluation of AWEPA, The Association of European Parliamentarians for Africa, and AEI, The African European Institute
1.00 Review of Norwegian Health-related Development Cooperation1988–1997
2.00 Norwegian Support to the Education Sector. Overview of Policies and Trends 1988–1998
3.00 The Project "Training for Peace in Southern Africa"
4.00 En kartlegging av erfaringer med norsk bistand gjennomfrivillige organisasjoner 1987–1999
5.00 Evaluation of the NUFU programme
6.00 Making Government Smaller and More Efficient.The Botswana Case
7.00 Evaluation of the Norwegian Plan of Action for Nuclear Safety Priorities, Organisation, Implementation
8.00 Evaluation of the Norwegian Mixed Credits Programme
9.00 "Norwegians? Who needs Norwegians?" Explaining the Oslo Back Channel: Norway's Political Past in the Middle East
10.00 Taken for Granted? An Evaluation of Norway's Special Grant for the Environment

1.01 Evaluation of the Norwegian Human Rights Fund
2.01 Economic Impacts on the Least Developed Countries of the Elimination of Import Tariffs on their Products
3.01 Evaluation of the Public Support to the Norwegian NGOs Working in Nicaragua 1994–1999
3A.01 Evaluación del Apoyo Público a las ONGs Noruegas que Trabajan en Nicaragua 1994–1999
4.01 The International Monetary Fund and the World Bank Cooperation on Poverty Reduction
5.01 Evaluation of Development Co-operation between Bangladesh and Norway, 1995–2000
6.01 Can democratisation prevent conflicts? Lessons from sub-Saharan Africa
7.01 Reconciliation Among Young People in the Balkans An Evaluation of the Post Pessimist Network

1.02 Evaluation of the Norwegian Resource Bank for Democracyand Human Rights (NORDEM)
2.02 Evaluation of the International Humanitarian Assistance of theNorwegian Red Cross
3.02 Evaluation of ACOPAMAn ILO program for "Cooperative and Organizational Support to Grassroots Initiatives" in Western Africa 1978 – 1999
3A.02 Évaluation du programme ACOPAMUn programme du BIT sur l'« Appui associatif et coopératif auxInitiatives de Développement à la Base » en Afrique del'Ouest de 1978 à 1999
4.02 Legal Aid Against the Odds Evaluation of the Civil Rights Project (CRP) of the Norwegian Refugee Council in former Yugoslavia

1.03 Evaluation of the Norwegian Investment Fund for Developing Countries (Norfund)
2.03 Evaluation of the Norwegian Education Trust Fund for Africain the World Bank
3.03 Evaluering av Bistandstorgets Evalueringsnettverk

1.04 Towards Strategic Framework for Peacebuilding: Getting Their Act Togheter.Overview Report of the Joint Utstein Study of the Peacebuilding.
2.04 Norwegian peacebuilding policies: Lessons Learnt and Challenges Ahead
3.04 Evaluation of CESAR´s activities in the Middle East Funded by Norway
4.04 Evaluering av ordningen med støtte gjennom paraplyorganiasajoner. Eksemplifisert ved støtte til Norsk Misjons Bistandsnemda og Atlas-alliansen
5.04 Study of the impact of the work of FORUT in Sri Lanka: Building CivilSociety
6.04 Study of the impact of the work of Save the Children Norway in Ethiopia: Building Civil Society

1.05 –Study: Study of the impact of the work of FORUT in Sri Lanka and Save the Children Norway in Ethiopia: Building Civil Society
1.05 –Evaluation: Evaluation of the Norad Fellowship Programme
2.05 –Evaluation: Women Can Do It – an evaluation of the WCDI programme in the Western Balkans
3.05 Gender and Development – a review of evaluation report 1997–2004
4.05 Evaluation of the Framework Agreement between the Government of Norway and the United Nations Environment Programme (UNEP)
5.05 Evaluation of the "Strategy for Women and Gender Equality inDevelopment Cooperation (1997–2005)"
1.06 Inter-Ministerial Cooperation. An Effective Model for Capacity Development?
2.06 Evaluation of Fredskorpset
1.06 – Synthesis Report: Lessons from Evaluations of Women and Gender Equality in Development Cooperation

1.07 Evaluation of the Norwegian Petroleum-Related Assistance
1.07 – Synteserapport: Humanitær innsats ved naturkatastrofer:En syntese av evalueringsfunn
1.07 – Study: The Norwegian International Effort against Female Genital Mutilation
2.07 Evaluation of Norwegian Power-related Assistance
2.07 – Study Development Cooperation through Norwegian NGOs in South America
3.07 Evaluation of the Effects of the using M-621 Cargo Trucks in Humanitarian Transport Operations
4.07 Evaluation of Norwegian Development Support to Zambia (1991-2005)
5.07 Evaluation of Development Cooperation through Norwegian NGOs in Guatemala

Norad
Norwegian Agency for
Development Cooperation

P.O.Box 8034 Dep, NO-0030 Oslo
Visiting adress:
Ruseløkkveien 26, Oslo, Norway

Telephone: +47 22 24 20 30
Fax: +47 22 24 20 31
Postmottak@norad.no
www.norad.no

Number of Copys: 200
February 2008
ISBN 978-82-754-272-1