Australasian Evaluation Society

Why do many evaluations have a positive bias? Should we worry?

Michael Bamberger*[1] *September 3, 2009

jmichaelbamberger@gmail.com

Many of the approaches used by international development agencies to assess the impacts and effects of their development interventions can potentially result in the findings having a positive bias so that the impacts and benefits of the interventions are over-estimated, while the proportion of the target population who do not benefit may be under-estimated and the potential negative consequences of the interventions may frequently be ignored. These biases are due to a combination of factors including: budget and time constraints, limited access to data – particularly baseline data; how evaluations are commissioned and managed; and political and organizational constraints and pressures. These biases have important consequences when evaluation findings are used to inform future management and policy decisions. While these biases are more obvious for under-resourced evaluations with unrealistically short deadlines, there are also a number of potential positive biases that can also affect well resourced "strong" quantitative evaluation designs. A number of recommendations are proposed to reduce positive bias and strengthen the validity of evaluation findings by strengthening how evaluations are managed and through strategies to strengthen the evaluation methodology, even when operating under budget and time constraints.

The two worlds of program evaluation[2]

Over the past decade that has been a steadily growing demand for a better assessment of the effectiveness and impacts of development assistance (Morra and Rist 2009. Chapter 2). "Does aid work?" is a question constantly asked by parliaments, ministries of finance and development, academia and civil society in both donor and aid receiving countries. International aid agencies have come under attack for the lack of rigor in assessing the effectiveness of their aid programs (CGD 2006) and for their willingness to accept process indicators (e.g. number of training courses organized) and output indicators (e.g. kilometers of roads built, number of children vaccinated) as convincing evidence that their investments are contributing to, for example, the reduction of poverty or progress towards the achievement of the Millennium Development Goals.

---

[1]. This paper draws on a guest lecture by the author at the International Program for Development Evaluation Training (IPDET) at Carleton University, Ottawa in June 2009

[2] Michael Bamberger is Senior Advisor for Evaluations, Social Impact (SI). SI is a global social enterprise dedicated to helping international agencies, civil society and governments become more effective agents of positive social and economic change. For more information on SI's work to improve development effectiveness visit: http://www.socialimpact.com

In response to these concerns there has been a movement towards the application of the principles of biomedical research to apply experimental designs to the assessment of the impacts of social and economic development programs (Bamberger, Rao and Woolcock *forthcoming*). The key features of these approaches are the use of random assignment of subjects (individuals, households, communities or organizations such as schools, health clinics or village banks) to treatment and control groups and the assessment of impacts on the basis of a limited number of quantitative indicators. Some authors go so far as to argue that randomized experimental designs are the only valid way [the "gold standard"] to assess program impacts (Banerjee 2007, Duflo and Kremer 2005), while others argue that when they can be applied experimental designs are the best approach with strong quasi-experimental designs as the next best option. Others (Deaton 2009; Ravallion 2008), while accepting the central importance of quantitative evaluations, question the exclusive reliance placed by the "randomistas" on randomization and make the case that careful theorizing and tests of hypothesis that derive from theory, along with other types of quantitative methodologies – propensity score matching, careful structural modeling, and instrumental variables should be not be so easily dismissed (Bamberger, Rao and Woolcock forthcoming).

While most development researchers recognize the theoretical benefits of randomization as a way to control for selection bias, the heated arguments for and against the use of RCTs and strong quasi-experimental designs has tended to obscure the fact that even on the most generous estimate it has probably not been possible to apply strong statistical designs in more than 25% of development evaluations (many would estimate that the proportion is closer to 10%), and that RCTs have probably not been applied in more than one or two percentage of impact evaluations[3]. So even if we accept all of the claims for the benefits of strong statistical designs, the question still remains how to assess the results (impacts) of the remaining 75-90% of projects and programs where these statistical designs cannot be used. While there are many training programs and publications on the application of strong statistical designs, there is surprisingly little guidance available on ways to produce methodologically acceptable estimates of the effects (impacts) for the majority of development programs that do not lend themselves to the application of these statistical designs[4].

So there really do seem to be two separate worlds of program evaluation. The minority of programs that lend themselves to quantitative evaluation and where the evaluations are often well funded and well documented in research journals; and the majority of programs where little guidance is available on how to produce methodologically sound estimates of results or impacts. Many statistically oriented researchers have pronounced that only evaluation designs with a counterfactual based on statistically matched comparison group can be considered as legitimate impact evaluations. Partly because quantitative researchers greater credibility with central government planning and finance ministries in developing countries, many of these agencies have accepted that it will not be possible for them so assess the impacts of many of their

[3] The author organizes a workshop every year at IPDET (International Program for Development Evaluation Training). The workshops are attended by 20-30 experienced evaluators and usually there will not be more than two or three who have ever had the opportunity to apply an experimental or strong quasi-experimental design throughout their whole professional career.

[4] For examples of publications that do address these questions of conducting evaluations under resource constraints see: Morra and Rist (2009); Leeuw and Vaessen (2009); Bamberger and White (2007); Bamberger, Rugh and Mabry (2006).

investments (when strong statistical designs cannot be used), while others make claims about achieving objectives but with little concern about the validity of the evaluation methodology. On the other hand many critics of the quantitative approach have claimed that there are alternative ways to estimate program impacts and refer, often somewhat vaguely to "alternatives to the conventional counterfactual".

So while many development agencies will ***occasionally*** support a more rigorous and high profile evaluation, often as a joint initiative with another agency, the standard approach to impact evaluation for many of these agencies will often be to commission most of their evaluations at the end of a project or program, often only allowing consultants to spend a very short time in each country or project location. As a result consultants are often only able to spend a short time in the field and do not have sufficient preparation time or the organizational support to efficiently plan their field work. Consequently consultants often only have time to meet with project beneficiaries or the agencies and stakeholders directly involved in the project and they may meet with very few (or sometimes no) people who are not participating in the project, and they find it even more difficult to identify and meet with any groups who may be worse off as a result of the project.

Does the widespread acceptance of rapid and under-funded program evaluations really matter?

A serious consequence of the rapid and under-funded program evaluation strategies of many development agencies is that a significant number of the evaluation reports have a *systematic positive bias* that tends to:

- Over-estimate the positive effects of the interventions they support
- Ignore or under-estimate the proportion of the population excluded or not benefiting from the interventions and
- Ignore or down-play many of the negative consequences - some of which can be very serious

The potential institutional and policy implications of this bias include:

- Agencies may continue to fund projects that might be producing less benefits than claimed
- Reduced resources and incentives to develop and test alternative programs
- Failure to take measures to reach-out to excluded or under-served groups – often the poorest and most vulnerable
- Failure to identify and address negative consequences, resulting in a failure to follow the policy of "Do no harm"

Box 1 gives two typical examples of how positive bias can occur and how this can lead agencies to draw misleading conclusions concerning program benefits.

In the first example the evaluators only meet with entrepreneurs who have received loans under the project. Many of the loan beneficiaries were able to use the funds to reduce the cost and time to access markets and consequently many were able to increase their sales and profits. As these findings were both positive and confirming the implementing agency's program theory

model there was no incentive (and perhaps no time or resources) to dig more deeply into the context in which the project was operating. In other similar instances the evaluation may be

Box 1 The danger of over-estimating project impact when the evaluation does not collect information on comparison groups who have not benefited from the project intervention

For reasons of budget and time constraints, a high proportion of evaluations commissioned to assess the effectiveness and impacts of ODA projects only interview beneficiaries and the agencies directly involved in the projects. When this happens there is a danger of a positive bias in which the favorable effects of the interventions are over-estimated, and the negative consequences are ignored or under-estimated. However, if these negative effects and impacts are taken into account, the net positive impact of the project on the total intended target population may be significantly reduced.

- An evaluation of a micro-credit program promoting the manufacture and marketing of traditional carpets in Bolivia, interviewed a sample of carpet makers who had received loans. It was found that on average their income from the sale of carpets *had increased significantly*. It was concluded that microcredit was an effective instrument for increasing the income of traditional artisans and reducing poverty. However, when carpet manufacturers who had not received loans were interviewed it was found that on average their income *had declined*. One of the reasons was that loan recipients were able to rent or purchase a vehicle to get their carpets to market more quickly and cheaply, which gave them a competitive advantage. On the basis of this additional information it was estimated that the total sales of carpets had probably not increased very much, but that a larger share of the market was now controlled by loan recipients. The inclusion of a control group (artisans who did not receive loans) radically changed the conclusions of the evaluation.
- Consultants commissioned to evaluate the impacts of food-for-work programs on women's economic and social empowerment were only able to spend an average of 3-4 days visiting project locations and meeting with affected communities in each country. Typically the consultants met with the local government agencies managing the projects, the local NGOs responsible for implementing the projects and residents of the communities where the programs operated. In each community consultants met with groups of women participating in the food-for-work programs and with many of their husbands. It was apparent that the project had produced significant increases in the income of the women, that their husbands were very supportive of the economic activities of their wives and that there was convincing evidence of the women's economic empowerment and increase in their self-confidence and social empowerment. In most cases the evaluation ended at this point and a very positive report was produced. However, in several cases the consultants also contacted key informants not involved in the project in order to obtain information on the experiences of other women who had not participated in the project. For example, a local nurse who regularly visited women in these same communities reported that many women who had attended the initial meetings had been forbidden by their husbands to participate in the project and in quite a few cases had been beaten for attending without his permission. Many men were unemployed and felt humiliated that they were not able to fulfill their traditional role of providing economically for their family. This again illustrates that a completely different image of the project was obtained when an effort was made to obtain information on the situation of non-participants and not to simply base the evaluation report on meetings with those who had benefited.

Source: Michael Bamberger (2009a) *Institutionalizing Impact Evaluation Systems in Developing Countries: Challenges and Opportunities for ODA Agencies*. Trends in Development Assistance Series 5 (Minato, N and Fujita, N Editors). Tokyo. Foundation for Advanced Studies on International Development.

mainly based on project records, which would tell a similar story, and the evaluator may not even have met with many of the entrepreneurs.

As the project was achieving its intended objectives and as the positive benefits could be easily seen there was no incentive to explore further. However, if the project did not appear to be achieving its objectives it is quite likely that the implementing agency would have felt the need to study in the situation in more depth.

In the second example, the consultants were only able to spend a limited time in each project location, and most of their information came either from a meeting with a group of (in this case very enthusiastic) women who had participated in the food for work program and who had experienced positive economic benefits and felt more empowered; or with the agencies implementing the project. Many of the husbands also attended the meeting and indicated their support for their wife's economic activities. In a country with a strong macho culture where many women were not allowed to work, the eloquent support of the men made a deep impression on the consultants, and many emotionally powerful quotes from the meetings were included in the consultant report. Having received convincing testimony directly from the beneficiaries in their own communities that "proved" the project was achieving its intended objectives, combined with the time constraints (wishing to return to town to avoid driving on treacherous rural roads in the dark) meant that consultants did not feel the need to explore the situation more deeply. Furthermore, the local NGO responsible for implementing the project was convinced of the efficacy of the project approach was able to supply many more examples of other communities where women had become similarly empowered.

Again the project was achieving its intended objectives, and this, combined with the logistical difficulties of finding more time to find additional respondents meant that the consultants felt no strong need to explore further.

Why are so many biased evaluations conducted and why are they accepted by the client?

Although no statistics are available on the proportion of program evaluations that contain a positive bias, it is the impression of the author that a significant proportion of development agencies employ evaluation strategies that contain a significant risk of *systematic bias*. Several pieces of circumstantial evidence support this conclusion. First, as discussed earlier, a high proportion of evaluations do not use a pretest/posttest comparison group design, in other words they do not include a statistical counterfactual that can control for selection bias and eliminate alternative explanations for the observed changes in the project population. Second, many development agencies have a policy of only allowing consultants to spend a very short time in the field so that even if they wished to, it would be difficult to interview a wider range of informants. Third, strategies for defining an alternative to the conventional statistical counterfactual are still at an early stage of development.

If we accept the prima facie case for the possibility of a significant proportion of program evaluations with at least a potential positive bias, the question arises as to why this is the case. Given the widespread recognition of the need for greater accountability for the use of aid funds

and the need to assess whether the intended objectives are being achieved, why do we find at the same time a significant number of evaluations being produced and accepted which do not achieve even a minimum standard of methodological rigor in addressing these questions? A number of factors seem to be at play.

A first set of factors relate to the fact that many evaluations are conducted with budget and/or time constraints [Bamberger, Rugh and Mabry 2006]. These constraints frequently limit the time that the evaluators (either internal or external) can spend in the field and, equally important the amount of time they can spend on the planning and preparation of the field work. The latter factor is often critical because much of the limited time in the field is wasted while setting up initial meetings and working on travel logistics, all of which could have been handled in advance. Even when sufficient funds are available the evaluators are often working under time pressure because the evaluation is not commissioned until late in the project or because there is a deadline for completing the final report. Often the available time is significantly shorter than it need have been because the evaluation manager did not start planning the evaluation until the last moment, often because she or he is managing a number of different evaluations at the same time.

Many evaluations also face data constraints. When evaluations are not commissioned until late in the project cycle estimates of the condition of the project and comparison groups prior to the start of the project will either have to rely on secondary data or on recall. While secondary data can provide very valuable (and often the only) data on baseline conditions, it must be treated with care as its value is often limited by the fact that it does not provide exactly the required information, may not cover all sectors of the target and comparison populations, may not have been collected at the time when the project was being launched, may not have been collected from the right people (only the "household head" but not the spouse or other relevant household members), or the quality of the data may not be very good. These data constraints will often introduce a positive bias because information is only collected from the mainstream, better off and more secure sectors. For example, the voices of wives or other family members have less control over household and community decisions are not heard, and squatters and those engaged in illegal or marginal economic activities are the sectors most likely to be excluded from the samples.

A frequent challenge occurs in using monitoring data and administrative records from the project being evaluated. While these sources are potentially valuable, the data has normally been collected for purposes other than the evaluation and may be incomplete, unreliable or may include a positive bias. With respect to this latter point, project staff often tend to put a positive spin on the data they record, particularly when this may be used for the purposes of staff evaluation; and they often tend to downplay problems. Another example are the frequently cited sources of bias in school attendance records. When school directors or teachers are evaluated on the basis of the number of students attending, there is an incentive to under-report non-attendance; in other cases where children are legally required to attend school up to a certain age, parents who do not wish their daughters to continue attending school after puberty may bribe the school to record their daughters as being present when in fact they are at home, and in other cases where there are no secondary schools for girls, families may send their daughters to

a boys secondary school where the school – perhaps for fear of sanctions from the education authorities, will record their sex as male.

Recall is also another valuable source of data but it is subject to a number of potential biases that cannot easily be controlled for. Other data constraints, particularly when data is collected exclusively through structured questionnaires may involve difficulties of collecting information on sensitive topics (such as domestic violence, illegal drug use, corruption), or problems of identifying and interviewing difficult-to-reach groups (such as sex workers, illegal immigrants, victims of domestic violence or people who are HIV positive). Of course in some cases the constraint on access to data may be mainly due to budget and time constraints which limit the possibility of collecting more difficult and time-consuming types of data.

All of these budget, time and data constraints can potentially introduce a positive bias. When working on a tight budget and with time constraints it is easier to meet with beneficiaries and it is more difficult, expensive and time-consuming to contact people excluded from the projects or who may have been affected negatively. In cases where the project is operating reasonably well, it is likely that a significant proportion of beneficiaries will report positively on the project, although this is of course not always the case.

There are also a number of institutional and political factors that can introduce a positive bias. Within agencies there are often subtle or sometimes not-so-subtle pressures to not "rock the boat" and to not ask sensitive questions or present critical findings on some of these issues. Examples include downplaying the importance of corruption or not challenging perceived cultural norms, many of which affect the status of women. In many organizations the evaluation function is relatively low in the organization hierarchy and the concerns of program staff and policymakers can often override the requirements of a rigorous or timely evaluation. In other cases the problem may be institutional inertia. It might be administratively difficult to change the procurement procedures for consultants, to contract local resource persons to conduct preparatory work to increase the productivity of the limited time that foreign consultants spend in-country, or to allow international consultants to spend more time in the field.

Another set of factors concern how consultants are contracted and managed. Often the request for proposals (RFP) or the terms of reference (TOR) provide very little detail on the required methodology. For example, there is frequently no requirement to identify and interview non-beneficiaries or groups who may have been negatively affected by the intervention, and consultants are also not required to select a sufficiently broad group of key informants to ensure that different sources of information are provided on who has benefited and who has not. Consequently there are often no requirements or incentives for consultants to go beyond collecting information from the easily accessible beneficiary groups. Similarly many agencies do not use quality assurance procedures to assess the evaluation methodology (either at the proposal submission stage or when the final report is presented) so that many findings that are based on limited evidence will often not be challenged. Many evaluation department staff are also responsible for managing a number of different evaluations at the same time, so that their time constraints oblige them to rely heavily on the consultants to determine the methodology and evaluation managers may even accept the consultant's findings and recommendations without having the time to really check them carefully

There may also be political pressures within the host country to avoid sensitive questions or to not interview groups that may be critical of the government. For example, there are cases where the evaluators are discouraged or forbidden from interviewing control groups that have not benefited from the project. In some cases this is to avoid interviewing groups critical of the project while in others in may be to avoid creating the expectation that these groups will receive compensation for having been relocated or having suffered other losses[5]. These political pressures may discourage, or in some cases forbid the evaluator from interviewing these groups.

Often the political pressures are more subtle as when the implementing agency arranges meetings for the evaluation consultants but reports that it was not possible to locate certain groups[6] or these "problem" groups simply come to meetings for reasons that are difficult to determine. In other cases the local agency will handpick who is invited (and not invited) to participate in focus groups. Other forms of political bias may include the partner agency deciding who receives copies or is asked to comment on the draft final evaluation report, or in extreme cases government may prepare, and put a certain "spin" on the summary of findings that is circulated.

Most of these institutional and political pressures tend to produce potential positive bias for similar reasons to those discussed earlier. As many of these factors, including the weak evaluation methodology, tend to show the program in a positive light, implementing or funding agencies may feel less incentive to challenge the methodology and findings than they would if the evaluation report showed the program in a poor light[7].

Potential sources of positive bias in strong statistical impact evaluations

Up to this point we have been discussing sources of positive bias in evaluations that are conducted under budget, time or data constraints and where many of the sources of bias result from having to conduct an evaluation with too small a budget to be able to interview enough respondents, too little time to properly prepare and implement the evaluation design, or where reliance has to be placed on inadequate secondary data. However there are also a number of potential sources of positive bias in many supposedly strong statistical designs where there is sufficient budget and time to implement an evaluation design that provides unbiased estimates of project impacts. While it could be argued that many of the following examples result from poor evaluation design, several are inherent in the nature of many statistical designs.

---

[5] Many projects have funds to compensate certain groups, such as people with land titles who are forced to relocate, but they may not have funds (or the desire) to compensate other groups, such as people without land titles or illegal squatters. If these groups that will not receive compensation are interviewed as part of a control group, it is then difficult for the government to later claim that they did not know these groups existed.

[6] In one case the government agency coordinating a program review meeting reported that they had not been able to locate any NGOs working in the areas of poverty reduction and food security and only government agencies attended the meeting.

[7] An example where government did challenge a methodology that showed their programs in a poor light was an evaluation of rural health service delivery in a North African country. The NGO conducting the evaluation had conducted a number of village case studies which reported frequent sexual harassment of female patients by doctors and a lack of respect for less educated patients who did not speak French. The Ministry of Health criticized the case study methodology as not being "professional research" and consequently dismissed the findings as being biased. The Ministry's criticisms may have been valid, but the point is that similarly weak methodologies tend not to be challenged when the findings are positive.

Many quantitative evaluations use as their sampling frame an existing list that is intended to cover the treatment and or control groups. Often this will be a government register that was prepared to identify individuals, households or communities eligible to receive financial support or other public benefits and that covers, for example, the school-age population, the internally displaced population, the low income population or families living in certain types of informal or substandard housing. However, many of these registers are not complete, either because they have not been updated or because they do not adequately capture all of the target population, and often the groups most likely to be left out are the poor and vulnerable. For example, in an evaluation of the impacts of government services on the internally displaced population [IDP] in a Latin American country it was estimated that the official IDP register used by the evaluators to select the sample did not include perhaps as much as 40% of the IDPs. In some cases people did not register because they wished to remain incognito, for example because of fear of reprisals for problems in their region of origin or because a man or woman had started a new family and did not wish their former spouse to be able to locate them. In other cases IDPs did not trust the government, or they may not have known about the register or how to enroll. Although it was not possible to assess the reasons for not registering, it is quite likely that the people left out would on average have been more unstable and vulnerable, in which case the sample used for the evaluation would have a positive bias. Often the target population is required to take the initiative to register and many fail to do so for many different reasons, while in other cases the government agency may use inadequate methods for locating and registering the eligible population[8]. So again it seems likely that the people most difficult to locate might be poorer or more vulnerable.

Another potential source of positive bias is that quantitative evaluations frequently collect data through the use of a structured questionnaire. It is widely acknowledged that these questionnaires are not well suited to capture sensitive information or to locate difficult to reach groups, so when used alone these surveys may under-estimate domestic violence, drug use, the prevalence of sex workers or people who are HIV positive. The limitations of the data collection instrument are often exacerbated because information is often only collected from the "head of household" in many communities and cultures the majority will be males, and he may not be fully informed about the situation or problems of other household members or may tend to downplay some of the problems, such as the time that women have to spend collecting water.

Many quantitative evaluations also have potential issues of construct validity as they tend to use a limited number of quantitative indicators to measure outcomes and impacts, and these are often inadequate to fully capture multidimensional constructs such as poverty, welfare, security or empowerment. Often these simple quantitative indicators may not capture some of the important qualitative dimensions of project impacts as when welfare is defined in terms of income and capital expenditures but ignores issues such as the level of domestic or community violence. Quantitative designs are also often not good at capturing unexpected outcomes, many, but not all, of which may be negative.

[8] For example, social service agencies often try to identify people eligible for their services by inviting people to come to meetings (and many do not attend), by making a single visit to each house and enrolling people who are at home (missing out those not at home), or by asking people to identify neighbors who would be eligible (so that people who do not have good relations with neighbors are less likely to be identified).

Most statistical evaluation designs also do not examine local contextual factors (e.g. whether the local economy is growing or declining, whether the local political environment is supportive or opposed to the project, whether local institutions provide the intended support) that may affect project outcomes. As the locations for pilot projects are often selected to maximize the likelihood of a successful outcome, for example because of strong support from local politicians, it is often the case that outcomes in the pilot locations may be more successful than they would be in more typical locations (where a larger project would be implemented) without this exceptional level of support or other favorable conditions. Consequently the lack of attention to contextual factors may introduce a positive bias.

A final potential source of positive bias is the treatment of *unobservables* in econometric analysis (Bamberger and White 2007; Bamberger, Rao and Woolcock *forthcoming*). Many statistical designs are based on the analysis of secondary data and have to rely on the information included in household income and expenditure surveys, demographic and health surveys or other available censuses or surveys. A frequent problem is that statistically significant differences between the project and control groups in the change in the dependent variables (impact indicators) may be due to the effect of the project, but they may also be due to pre-existing differences between the project and comparison groups that are not measured in the secondary data set. For example, a number of surveys in countries such as Bangladesh have found that women who borrow from village banks tend to increase their expenditures on children's education, household necessities and housing improvements more that do women with comparable socio-economic conditions who did not get a loan from the Bank. The problem in the interpretation of the findings is that the differences in these outcome indicators may be due, at least in part, to pre-existing differences between the project and control groups with respect to factors such as previous business experience, self-confidence or level of control that women exercise over household resources and decisions. So it might be the differences in these characteristics that make these women more likely to borrow money from the village banks or it could be that they would have increased their income (and hence household expenditures) more than the average women *even if they had not obtained a loan*. Econometricians often make the (questionable) assumption that these unobservables are *time-invariant* so that their effect on the outcome variables can be ignored. However, in many cases the analyst does not even know what these unmeasured factors might be so the assumption that they will be time-invariant is often highly questionable. If in fact some of the observables were not time invariant, their presence might explain part of the assumed project impact – in which case ignoring them could introduce a positive bias because part of the assumed project impact is in fact due to these unmeasured factors[9].

[9] See Roodman and Morduch 2009 for an econometric study that reanalyzes several well-known economic assessments of the impacts of women's access to microcredit in Bangladesh and claims, through the use of more sophisticated analysis techniques to have eliminated potential sources of biases in the earlier estimates of the impacts on women and their households. However, the study makes no reference to the potential interpretation problems posed by unobservables such as those we have discussed, so that a major potential source of bias was not considered and in fact could not have been addressed in the analysis as the required information on factors such as women's role in control of household resources or previous business experience was not included in the data base. Addressing these questions would probably have required additional mixed method field work to reconstruct the situation of women with respect to variables such as household decision-making and prior business experience in project and control group families prior to the start of the village bank programs.

So one or more of these factors could introduce the possibility of positive bias in the estimates of project impact obtained from conventional quantitative evaluation designs. Incomplete sampling frames may leave out some of the most vulnerable groups while structured surveys may fail to capture some of the most sensitive issues, may not interview the most vulnerable groups or may fail to identify and collect data from groups not wishing to be surveyed – again often the most vulnerable groups. Also the use of a limited number of quantitative impact indicators may fail to capture some of the important qualitative dimensions of the constructs used to define and measure impacts. Finally, the lack of contextual analysis may over-estimate project impacts by ignoring some of the important local factors that enhance outcomes but that would not be present if the project were replicated.

Ways to reduce positive bias in the evaluation findings

Even when operating under real-world constraints, positive bias can be reduced by combining better management of evaluation with a number of measures to strengthen the evaluation design.

Strengthening the management of program evaluations

An important first step is for development agencies to assess the procedures they use for commissioning, managing, disseminating and using program evaluations to determine whether the procedures are vulnerable to any of the threats to validity discussed in the previous sections. If any systemic problems are identified it is then important for the agency to assess the seriousness of the problems and how they affect the quality of the evaluation information used for program quality control and future budget and policy decisions. Depending on the nature, pervasiveness and seriousness of the problems one or more of the following steps should be considered:

a. Strengthen the Request for Proposals [RFP] or the Evaluation Terms of Reference [TOR]:
    i. Provide a clearer and more explicit statement of the objectives of the evaluation, how the findings will be used, the required level of precision and the kinds of management and/or policy decisions to which the findings will contribute.
    ii. Define clearly the required minimum acceptable methodological standards for the evaluation. These may be provided in fairly general terms or they may be very detailed and specific. Some of the minimum requirements for evaluations operating on a limited budget and tight timeline might include: (a) the definition of a counterfactual together with an explanation of the sources of data that will be used to test it; (b) a specified minimum number of meetings with non-beneficiaries; (c) identifying and obtaining information on groups who may be negatively affected by the project; (d) selection of a sample of key informants that are both familiar with the project and with the broader political, economic and socio-cultural context within which it operates; (e) methodological procedures for selecting participants for focus groups (if these are to be used) and for conducting the discussions and reporting the findings.

b. Build-in quality assurance procedures for assessing the methodology and the presentation of findings and conclusions. This may include:

i. *Preliminary evaluation feasibility analysis* to ensure the proposed TOR could be achieved with the available resources, within the required time frame and at this point in the project cycle. This should be conducted before the TOR are issued as a reality test to avoid embarking on an evaluation that clearly could not achieve its stated objectives. Two common reasons why the stated objectives could not be achieved include: the budget is insufficient to implement the proposed design and collect the required data; and insufficient time is allowed for the design, implementation and analysis of the evaluation. Timing must also take into account climatic factors such as the onset of the rainy season, or factors such as public holidays, the start of election campaigns, periods when national agencies will not be able to actively participate due to periods of financial and organizational planning; (c) the likelihood of strong political opposition [which does not mean the evaluation should not go ahead but which may cause delays or difficulties in access to information]; (d) it is too early in the project cycle to be able to measure the required outcomes or impacts; and (e) unrealistic estimates of the time required for administrative arrangements such as approval by the host government, procurement of local consultants or other local services, time for review and approval of inception reports.

ii. *Evaluability analysis*. Once the evaluation proposal is received (from agency staff if it is to be conducted internally or from consultants) an evaluability assessment should be conducted to determine (a) whether the proposal fully responds to the terms of reference ; (b) whether the proposed methodology is technically sound; (c) whether it will be possible to implement the methodology within the budget and time constraints; and (d) whether there a reasonable likelihood that the required data can be collected (either from secondary sources or through the proposed primary data collection procedures).

iii. *Application of a threats to validity checklist*. A threats to validity checklist (Bamberger 2009c) can be applied at various points in the evaluation to identify potential threats to the validity of the evaluation findings and recommendations[10]. The checklist can be used at various points in the evaluation cycle: the start of the evaluation as part of the evaluability analysis, mid-way through the evaluation, to assess and propose revisions to the draft final report, or once the final report has been submitted. The checklist can be used by funding agencies, the evaluation team, the project implementation agency or national planning and policymaking agencies.

iv. Require the sources to be given for each finding and recommendation in the evaluation report and check on the accuracy and adequacy of the sources. The executive summaries of many evaluation reports include statements like "many

[10] Bamberger (2009c) identifies 7 dimensions of validity: (A) Objectivity: are the conclusions drawn from the available evidence, and is the research relatively free from researcher bias? (B)Reliability: is the process of the study consistent, coherent and reasonably stable over time? (C) Internal validity: Are there any reasons why the assumed causal relationship between two variables (project interventions and outcomes) may not be valid? (D) Statistical conclusion validity: Are there any reasons why conclusions about statistical association (i. e. Differences between the project and comparison groups) may not be valid? (E) Construct validity: Do the indicators used to define processes, outcomes and impacts accurately reflect the essence and scope of the key constructs being assessed? (F) External validity: reasons why inferences about how study results would hold over variations in persons, settings, treatments and outcomes may not be correct (G) Utilization: How useful were the findings to clients, researchers and the communities studied?

respondents stated that …." or "most women had encountered problems with respect to …..". Quite often "many" or "most" in fact only refer to one or two people attending focus groups or who were included in case studies. In other cases some of the findings are not consistent with the evidence presented in the main report or sometimes there is no evidence at all to support the statements. It is not unusual for the Executive Summary to present a more positive assessment of project effects than the evaluation findings actually justify. As many readers only look at the Executive Summary, and if this presents their program in a favorable light they may not feel the need to check the validity of the findings[11] or the evidence on which they are based.

At the time of contract negotiation the evaluators should be advised that all of their findings will be checked for accuracy and all must be documented. It is of course essential that the agency commissioning the evaluation does in fact systematically follow-up and actually check the sources. Where the findings are not supported by the evidence the consultants must be asked to revise the report, or even some cases to return to the field.

c. Assess more realistic minimum time and budgets for different kinds of evaluation. The goal should be to propose the minimum increases that could make a significant difference but that would be realistic within the agency budget and operating procedures. These more realistic minimums should be applied on a pilot basis to selected evaluations and an assessment made of the value-added of the additional resources and time.

Consider including in the revised budgets time and funding to hire local research support in planning and preparing for the visit of the external consultants. A local consultant could help with, among other things: preparing field visits to ensure groups and people met include beneficiaries and non-beneficiaries; preparing focus groups or community meetings to ensure all key groups are represented; ensure reliable information is obtained on negative consequences of project; identify and arrange meetings with key informants; prepare briefing on some of the key issues to be addressed; and arrange one or more meetings with a representative group of civil society organizations. If possible at least one video or phone conference should be arranged with the local consultant to discuss these preparatory activities.

[11] An evaluation was commissioned in a South American country to assess the impact of rural roads on access to health, education and other services. The executive summary reported that the construction of rural roads significantly increased women's utilization of rural health centers. This finding was widely quoted by the Ministry of Transport. In fact the main report indicated that the impact of rural roads on women's use of health centers was quite limited, partly because the husband controlled the household budget and would often not give his wife the money for the bus fare if she "didn't look sick" or if he would have to mind the children while she was travelling, and partly because many rural communities did not believe in the utility of modern medicine. Unfortunately, not many people actually read the main report so the positive impacts of rural roads continued to be cited.

d. Develop practical guidelines for consultants and agency staff on how to strengthen evaluations while working under real-world constraints.

Ways to strengthen the evaluation methodology to reduce bias – even when working under budget, time and data constraints

a. There are a number of strategies for reducing the costs and time required to collect and analyze data while ensuring an acceptable minimum level of methodological rigor and reducing bias [Bamberger, Rugh and Mabry 2006; Bamberger and White 2007]. Some of the approaches include:

i. Simplifying the evaluation design. For statistical designs this will usually involve eliminating either the baseline data collection for the project and/or comparison group, or the post-test comparison group. There is always a trade-off between saving cost and time and ensuring an acceptable minimum standard of methodological rigor. When exploring these trade-offs it is helpful to use the threats to validity checklist to identify potential validity issues.

ii. Clarifying what information the client really needs for decision-making and what information is only "interesting". Often the data collection instruments can be significantly shortened (resulting in cost and time-saving) by cutting out non-essential information. Resources can also be saved by reducing the amount of disaggregated analysis that is required. For example, it is often much faster and cheaper to assess the average impact of the project on the total beneficiary population than to compare the impact on different sectors of the population, or to compare impacts in different project locations.

iii. Creative use of secondary data. There are often potentially valuable sources of secondary data that could be used to create a comparison group or to reconstruct baseline data (Bamberger 2009b). Bamberger, Rao and Woolcock (forthcoming) illustrates the wide range of secondary data sources that were used to assess the impacts of a rural road construction project in Eritrea on access to schools and health centers, the cost and time to transport agricultural products to market, access to government agencies and social networks. Sources included: school attendance and enrolment records, the records of local health centers, records of purchases and sale in the local farmers cooperative markets, regional and national data on agricultural prices, and vehicle registration records.

iv. Determining the required sample size. The two main determinants of the required sample size for estimating the statistical significance of project impacts are the estimated effect size and the required power of the test. Sample size estimation is a subject on which many evaluators are quite weak, and it will often be found that the sample size used in the evaluation is either too large, so that the cost of data collection is higher than necessary, or too small so that there is a danger of not being able to detect small but operationally important impacts. It is sometimes, but not always possible to reduce the required sample size either by accepting a lower power of the test, by finding ways to increase the effect size or by increasing the efficiency of the sample design (Bamberger, Rugh and Mabry 2006 Chapter 14).

v. Reducing the costs of data collection, processing and analysis. There are a number of ways that the costs of data collection can be reduced. Examples include: direct observation instead of surveys (for example, observing and documenting how people travel to work or the time taken to collect water); self-administered questionnaires; collecting information through group interviews (e.g. focus groups) rather than individual interviews.

b. There are also a number of evaluation design strategies that can strengthen the methodological rigor and reduce bias while working under real-world constraints. Space does not permit a full discussion of these techniques but examples include:

i. Using program theory models to strengthen the analysis of causality and to help identify the key hypotheses that should be tested [Bamberger, Rugh and Mabry 2006 chapter 10].
ii. Using mixed method designs to strengthen the validity of analysis based on relatively small samples through the use of triangulation (Rao and Woolcock, 2003; Woolcock 2009).
iii. Using strategies to "reconstruct" baseline data for evaluations that do not begin until late in the project cycle (Bamberger 2009b). These strategies combine the creative use of secondary data, recall, key informants and the use of PRA and other participatory group techniques such as the construction of timelines, historical transects, cause-effect diagrams and impact diagrams (Kumar 2002).
iv. Defining alternatives to the conventional counterfactual for attribution analysis when it is not possible to use a statistically matched comparison group. This is a new field but some of the promising approaches include concept mapping (Kane and Trochim 2007), Program Theory of Change (Morra and Rist 2009), general elimination strategy (Scriven 2008), participatory group techniques and the creative use of secondary data.

Conclusion

There are two distinct worlds of program evaluation: a few rigorous and highly publicized experimental and strong quasi-experiment designs, and a much larger number of evaluations that are conducted under budget and time constraints and that tend to use quite weak methodologies. While the evaluation literature provides very detailed guidance on, and critiques of the strong statistical designs, very little guidance is available on how to strengthen the rigor of the majority of evaluations that either use weak quasi-experimental designs or non-experimental designs. As a consequence of the time and resource constraints under which the latter category of evaluations are conducted and the very limited attention they receive in the literature , many of these evaluations tend to have a positive bias, over-estimating the impacts of the programs, paying little attention to the sectors of the target group that do not have access to the program services and largely ignoring the negative consequences, often quite serious that some programs can produce.

This widespread positive bias has important operational and policy consequences. Programs may continue to be funded even though they are producing smaller positive impacts than agencies assume, there are reduced resources and incentives to develop and test alternative programs, there may be a failure to take measures to reach out to under-served groups, and agencies may fail to live up to the principle of "Do no harm".

There are a number of reasons why, despite the increased demand for more rigorous assessments of aid effectiveness, such a high proportion of program evaluations continue to use weak methodologies and produce biased results. A first set of factors concern the severe budget and time constraints under which many evaluations are conducted. A second set of factors relate to the fact that probably the majority of evaluations are not commissioned until towards the end of the project cycle and do not have access to baseline data generated specifically for the evaluation. Consequently the information on which the estimates of the pre-project situation are based are often not well-suited to the purpose of the evaluation and may be inadequate, of low quality or in some cases not available at all.

Other causes of the systematic positive bias relate to the way that evaluations are commissioned and managed. Often the evaluation terms of reference [TOR] do not provide a clear definition of purpose of the evaluation, the questions to be addressed and the kinds of decisions to which the findings will contribute. Also many TOR make unrealistic demands on what the evaluation is intended to achieve, for example requesting an assessment of impacts when it is often too early in the project to even measure outcomes; presenting a long list of questions to be addressed without any prioritization; or requesting a high level of rigor that it will be impossible to achieve within the budget and time constraints. Many TOR do not provide any guidance on the preferred methodology or minimum standards of methodological rigor (such as the need for a comparison group or at least conducting interviews with groups who did not benefit from the project). While it is a legitimate strategy to permit consultants, especially very experienced consultants, to have flexibility in proposing the methodology; very often consultants may propose unrealistic methodologies and these are often not challenged by the client. For example, it is very common for consultants to commit themselves to assessing impacts, even when they should know that they will be impossible to assess impacts at this early stage of the project, with the approved budgets or with the available data sources. In the final report the consultants will often state that it was not possible to assess impacts as it was too early or they did not have sufficient resources, and frequently they are not called to account for having committed themselves to assessing impacts in their proposal – even when it should have been clear that this could not be done. The lack of supervision of consultants is often partly due to the fact that evaluation department staff are often managing several evaluations at the same time, so they have little time to carefully monitor the consultants they have contracted.

Other organizational factors creating a positive bias include pressures from within the agency (subtle or more direct) not to "rock the boat" by asking difficult or sensitive questions that might affect ongoing negotiations with the partner government. There may also be political pressures from within the country that may discourage, (or in some cases forbid, interviewing certain groups (e.g. families or communities that did not have access to the project or who are in the process of being forcibly resettled or losing their livelihood), or asking certain questions. These pressures may also determine who is involved (and not involved) in the consultation meetings on the evaluation planning, or who gets to receive a copy of the draft evaluation report or is invited to comment on the report. In some cases the host government may even prepare the summary of the evaluation that is distributed with the full report. All of these factors tend to produce a positive bias in the evaluation by screening out the groups who have not benefited or by screening out questions that focus on some of the problems or negative outcomes of the project.

While the sources of positive bias are often quite evident for under-funded evaluations conducted under time constraints, there are also a number of potential sources of positive bias in well-funded evaluations using a strong statistical design. A first source of potential bias is that many statistical evaluations use an existing register or list, usually prepared by a government agency to identify vulnerable groups eligible to receive government services. Often these registers do not capture the total target population and frequently the groups left out are the poorest or most vulnerable – potentially introducing a positive bias as the most needy groups may not be included in the study. A second potential positive bias is that quantitative surveys normally used a structured data collection instrument which is not well suited for obtaining sensitive information on topics such as domestic violence. Furthermore, many household surveys for reasons of cost only interview the "household head" (in most cases a male) so that groups such as women, the elderly or the handicapped may not be interviewed directly so that their problems or concerns may be under-represented. Structured survey techniques are also not well-suited for identifying and interviewing marginal groups such as illegal settlers, sex trade workers or drug users. Consequently the prevalence of these groups may be under-estimated and their particular problems and concerns not documented.

Quantitative techniques also tend to ignore the process of project implementation or the local contextual factors that affect outcomes in each location. As a result positive outcomes that may be partly due to special local circumstances, such as strong support from the mayor or the provision of additional resources (not recorded in the project budget) by a government agency wishing the project to go well so that the donor will provide additional funding, may be attributed to the project intervention – thus creating a positive bias.

A final factor is the way that econometric analysis, by its treatment of unobservables (assuming they are time-invariant and thus can be ignored), has a tendency to under-estimate the effect of factors such motivation, prior experience or special personal or family characteristics that may partly explain which the project group has performed better than the comparison group. This again can lead to over-estimating the project impact.

The paper concludes with recommendations on ways to reduce positive bias and produce more precise estimates of project impact – even when conducting evaluations under budget, time and data constraints. The first set of recommendations concern ways to strengthen how program evaluations are managed. A crucial first step is to acknowledge the potential sources of positive bias in the evaluation findings and to recognize the consequences of these biases when evaluations contribute to the identification of future investment priorities and to policy formulation. A second step is then to improve the terms of reference for the evaluation. The statement of objectives must be made more precise and also more realistic, and consultants should be provided with more guidance on the preferred methodology and the minimum acceptable methodological standards. A second step is to apply quality assurance procedures. These may include a preliminary assessment of the feasibility of achieving the stated evaluation objectives before the TOR are issued, conducting an evaluability analysis of the evaluation methodology proposed by the consultants, and the use of a threats to validity checklist to identify and correct potential threats to validity at different stages of the evaluation. An additional step is to require all conclusions and recommendations in the evaluation reports to cite the evidence on

which they are based. It is then the responsibility of the evaluation manager to check whether the evidence does in fact support the findings (which it often does not do).

A practical problem is that many evaluations are commissioned with an inadequate budget and an unrealistically short time-line for the planning, implementation and analysis of the evaluation. An essential first step must be to assess experience from previous evaluations and define what would be a realistic minimum budget and time-line for different kinds of evaluation. The agency should then provide these additional resources to a sample of evaluations and assess whether there is evidence of value-added in terms of the quality, level of detail and practical utility of the better resourced evaluations.

A second set of recommendations concern ways to strengthen the evaluation methodology to reduce bias. Five strategies were proposed for addressing budget, time and data constraints. A number of techniques were also proposed for maximizing the evaluation rigor when working under these constraints. The techniques include: using program theory to strengthen the analysis of causality; using mixed method designs to strengthen the validity of the analysis, particularly when working with small sample sizes and limited access to data; using strategies to "reconstruct" baseline data when the evaluation does not begin until late in the project cycle; and defining alternatives to the conventional counterfactual to assess causality when it is not possible to use statistically matched comparison groups.

If the widespread nature of positive bias in evaluation findings and recommendations is recognized, the proposed steps may help reduce these biases and thus enhance the value of evaluation as an essential management and policymaking tool.

References

Michael Bamberger (2009a) *Institutionalizing Impact Evaluation Systems in Developing Countries: Challenges and Opportunities for ODA Agencies*. Trends in Development Assistance Series 5 (Minato, N and Fujita, N Editors). Tokyo. Foundation for Advanced Studies on International Development.

Bamberger, Michael (2009b) "Strengthening Impact Evaluation Designs through the Reconstruction of Baseline Data" *Journal of Development Effectiveness* 1(1): pages 37-59

Bamberger, Michael (2009c). "Checklist for Assessing Threats to the Validity of Findings and Recommendations of Impact Evaluations". Workshop on Conducting Impact Evaluations under Constraints. Ottawa. International Program for Development Evaluation Training. www.realworldevaluation.org.

Bamberger, Michael and Howard White (2007) Using Strong Evaluation Designs in Developing Countries: Experience and Challenges. Journal of Multidisciplinary Evaluation. Volume 4, Number 8. October 2007.

Bamberger, Michael; Jim Rugh and Linda Mabry (2006) *RealWorld Evaluation: Working under Budget, Time, Data and Political constraints*. Thousand Oaks. Sage Publications.

Bamberger, Michael; Vijayendra Rao and Michael Woolcock (forthcoming). "Using Mixed Methods in Monitoring and Evaluation Experiences from International Development" in Charles Teddlie and Abbas Tashakkori (Editors) *Handbook of Mixed Methods Research.* Second Edition. Thousand Oaks. CA. Sage Publications.

Banerjee, Abhijit (2007) *Making Aid Work*, Cambridge: MIT Press.

Deaton, Angus, 2009, "Instruments of Development: Randomization in the tropics, and the search for the elusive keys to economic development," The Keynes Lecture, British Academy

Duflo**,** Esther and Michael Kremer (2005) "Use of Randomization in the Evaluation of Development Effectiveness," from George Pitman, Osvaldo Feinstein and Gregory Ingram, ed., *Evaluating Development Effectiveness*, New Brunwick: Transaction Publishers.

Kane, Mary and William Trochim, W (2007). *Concept Mapping for Planning and Evaluation*. Thousand Oaks, CA. Sage Publications.

Kumar, Somesh (2002). Methods for Community Participation: a complete guide for Practitioners. Rugby, England. ITDG Publishing.

Leeuw, Frans and Jos Vaessen (2009). *Impact Evaluations and Development. NONIE Guidance on Impact Evaluation.* Draft prepared for the Cairo International Evaluation Conference. April 2009. Network of Networks on Impact Evaluation (NONIE)

Morra, Linda and Ray Rist (2009) *The Road to Results: Designing and Conducting effective Development Evaluations*. Washington, D.C. The World Bank

Rao, Vijayendra and Michael Woolcock (2003) "Integrating Qualitative and Quantitative Approaches in Program Evaluation", in Francois J. Bourguignon and Luiz Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools* New York: Oxford University Press, pp. 165-90

Ravallion, Martin, (2008) "Evaluating Anti-Poverty Programs," edited by Paul Schultz and John Strauss*, Handbook of Development Economics* Volume 4, Amsterdam: North-Holland.

Scriven, Michael (2009) "Demytholizing Causation and Evidence" pp. 134-152 in Stewart Donalson, Christina Christie and Melvin Mark (Editors) *What Counts as Credible Evidence in Applied Research and Evaluation Practice*? Thousand Oaks. CA. Sage Publications

David Roodman and Jonathan Morduch (2009). The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence - Working Paper 174. Center for Global Development 2009

Woolcock, Michael (2009) 'Toward a Plurality of Methods in Project Evaluation: A Contextualized Approach to Understanding Impact Trajectories and Efficacy' *Journal of Development Effectiveness* 1(1): 1-14