DISCUSSION PAPER / 2007.01

Methodological Challenges in Impact Evaluation:
The Case of the Global Environment Facility (GEF)

Jos **Vaessen**
David **Todd**

University of Antwerp

IOB
INSTITUTE OF DEVELOPMENT POLICY AND MANAGEMENT

Comments on this Discussion Paper are invited.
Please contact the authors at <jos.vaessen@ua.ac.be>

Instituut voor Ontwikkelingsbeleid en -Beheer
Institute of Development Policy and Management
Institut de Politique et de Gestion du Développement
Instituto de Política y Gestión del Desarrollo

Postal address:
Prinsstraat 13
B-2000 Antwerpen
Belgium

Visiting address:
Lange Sint-Annastraat 7
B-2000 Antwerpen
Belgium

Tel: +32 (0)3 275 57 70
Fax: +32 (0)3 275 57 71
e-mail: dev@ua.ac.be

http://www.ua.ac.be/dev

DISCUSSION PAPER / 2007.01

Methodological Challenges in Impact Evaluation:
The Case of the Global Environment Facility (GEF)

Jos **Vaessen***
David **Todd****

January 2007

* Jos Vaesen is researcher at the Institute of Development Policy and Management, University of Antwerp.
** David Todd is Senior Evaluation Officer at the GEF Evaluation Office, Global Environment Facility.

IOB
INSTITUTE OF DEVELOPMENT POLICY AND MANAGEMENT

University of Antwerp

Contents

	Abstract	**4**
	Résumé	**4**
1.	**Introduction**	**5**
2.	**The GEF Evaluation Office Impact Evaluation**	**6**
3.	**Methodological challenges**	**8**
3.1.	The problem of the independent variable	8
3.2.	The problem of the dependent variable	10
3.3.	Methodological responses to the attribution and aggregation challenge	15
4.	**A theory-based impact evaluation approach**	**18**
4.1.	Addressing the attribution challenge	18
4.2.	Addressing the aggregation challenge	23
5.	**Conclusions**	**26**
	References	**28**

Abstract

In this paper, we explore some of the methodological challenges that evaluators face in assessing the impacts of complex intervention strategies. We illustrate these challenges, using the specific example of an impact evaluation of one of the six focal areas of the Global Environment Facility; its biodiversity program. The paper discusses how theory-based evaluation can provide a basis for meeting some of the challenges presented.

Résumé

Défis méthodologiques dans l'évaluation d'impact: le cas du Fonds pour l'Environnement Mondial (FEM)

Dans cet article nous explorons quelques-uns des défis avec lesquels les évaluateurs sont confrontés quand ils s'interrogent sur l'impact des stratégies d'intervention complexe. Nous illustrons ces défis en utilisant l'exemple spécifique d'une évaluation de l'impact d'un des six domaines focaux du Fonds pour l'Environnement Mondial : son programme de biodiversité. On présente l'approche « theory-based evaluation » (évaluation par la théorie d'action) comme base des solutions aux défis soumis.

1. Introduction [1]

In recent years, as results-based management has become a central concept and practice in the management of development assistance, multilateral and bilateral donor organizations have increasingly demanded 'hard evidence' of the results of the policies, programs and projects they support. However, this clamor to measure results has not yet been matched by a proportionate increase in funding and achievement in measuring outcomes and impacts (e.g. Picciotto, 2003; CGD, 2006).

The Global Environment Facility (GEF) recently prepared its approach to an impact evaluation of its biodiversity program, one of six focal areas of intervention. In this paper, we explore some of the methodological challenges that the evaluators are facing. The GEF biodiversity portfolio serves the overarching objective of protecting globally important biodiversity. In practice however, there is still uncertainty regarding to what extent and how this objective is achieved by the projects to which the GEF contributes. More specifically, this uncertainty is fuelled by the fact that:

- projects often encompass a wide range of discrete activities while it is often not clear how these different activities contribute to project objectives and higher-level program goals (GEF, 2004a);
- data on the outcomes and impacts of biodiversity projects (including appropriate indicators for measuring these effects) are scarce (GEF, 2004a);
- environmental change processes are complex and changes may only become apparent years after a project has been completed (MEA, 2005).

These constraints as well as the inherent complexity of linking specific interventions to global biodiversity gains pose particular methodological challenges to impact evaluation. After discussing these challenges we will argue that theory-based evaluation can provide a basis for meeting some of the challenges presented.

Our discussion is inspired by White's (2003) triple-A assessment of development agency performance. The three A's are: attribution, aggregation and alignment. Although White applies the concepts in the assessment of the quality of agency performance reports, they also represent key concerns in evaluation and are particularly relevant to impact evaluation. Applied to the context of impact evaluation they can be defined as follows. Attribution refers to the problem of establishing a causal link between intervention outputs and observed changes in impact variables. In order to be able to isolate the effect of an intervention on a particular target from the influence of other variables (e.g. the policy environment, socio-economic trends), evaluators often rely on the principle of a

[1] This paper discusses an impact evaluation managed by the GEF Evaluation Office. The content of the paper is the sole responsibility of the authors and does not commit the GEF Evaluation Office or any other actors involved in the evaluation to the authors' views. An earlier version of this paper was presented at the UKES-EES Conference in London in October 2006. The authors would like to thank participants at the conference, Frans Leeuw, Robrecht Renard and Osvaldo Feinstein for their comments.

counterfactual scenario (what would have happened without the intervention). Aggregation concerns the question of how micro-level impact data can be meaningfully aggregated across interventions. This is crucial in impact assessment studies of clusters of interventions (as opposed to a single (site-specific) intervention), like the biodiversity portfolio. Related to this, alignment touches upon the issue of whether data collected at micro-level are relevant with respect to an agency's overall objectives. Our discussion is mainly constructed around two of the three concepts, the challenges surrounding the issues of attribution and aggregation.

2. The GEF Evaluation Office Impact Evaluation

The Global Environment Facility (GEF) is an international organization that provides grants to initiatives in developing countries directed at the protection of the (global) environment. The GEF is governed by a Council which consists of representatives of the different member states contributing to the GEF and of recipient countries. Operations are overseen by the GEF Secretariat and evaluated by the GEF Evaluation Office, which reports directly to the Council. Individual GEF interventions are mainly managed by its implementing agencies, the World Bank, The United Nations Development Program (UNDP), and the United Nations Environment Program (UNEP); and (to a lesser extent) by its executing agencies (e.g. the Asian Development Bank, the Inter-American Development Bank). Implementation in the countries and regions of intervention is often handled by governmental, non-governmental or private sector organizations in collaboration with the above-mentioned agencies. GEF representatives in different countries (focal points) and an international team of experts (Scientific and Technical Advisory Panel) are additional elements of support in the preparation and implementation of interventions.

GEF funding is directed at six principal focal areas: biodiversity, climate change, international waters, ozone depletion, land degradation, and persistent organic pollutants. In order to induce changes within these focal areas, the GEF employs a number of instruments and initiatives: full-sized projects (projects of more than US $ 1 million), medium-sized projects (projects of up to 1 million US $), enabling activities (aimed at fostering a policy environment conducive to environmental protection), the small grants program (a mechanism to support local small projects up to US $ 50,000), and a small and medium enterprise program (in collaboration with the International Finance Corporation).

The biodiversity program represents the largest portfolio of interventions financed by the GEF and has been selected as the evaluand of the impact evaluation. In the period 1991 to August 2006 approximately $ 2.22 billion of GEF funding with some $ 5.16 billion of co-financing was allocated to biodiversity projects (GEF PMIS data base, August 30, 2006). The biodiversity portfolio is the operational mechanism of the Convention on Biological Diversity (CBD). The CBD's Conference of Parties every two years provides guiding principles as a strategic foundation for the GEF's biodiversity portfolio. As a general mission statement, projects in the biodiversity focal area seek to reduce biodiversity loss attributable to human behavior. More specifically, biodiversity is to be protected at three levels: ecosystems, species and genes. By doing so, the GEF aims to promote three types of behavior vis-à-vis biodiversity: conservation, sustainable use and benefit sharing.

The biodiversity portfolio is divided into 5 operational programs representing different ecosystems targeted by interventions: Arid and Semi-Arid Zone Ecosystems; Coastal, Marine, and Freshwater Ecosystems; Forest Ecosystems; Mountain Ecosystems; Conservation and Sustainable Use of Biological Diversity Important to Agriculture. In order to improve the coherence of the portfolio and provide better guidance to individual interventions, four strategic priorities were recently introduced: Catalyzing Sustainability of Protected Areas; Mainstreaming[2] Biodiversity in Production Landscapes and Sectors; Capacity Building for the Implementation of the Cartagena Protocol on Biosafety; Generation and Dissemination of Best Practices for Addressing Current and Emerging Biodiversity Issues. Increasingly, guidance for project preparation and implementation, monitoring and evaluation, and (to some extent) allocation of funding is expected to be structured by these priorities.

[2] Mainstreaming biodiversity "involves integrating the values and goals of biodiversity conservation and sustainable use into economic sectors and development policies and programmes" (GEF, 2004b: 2).

The diversity in operational programs and strategic priorities is an indication of the broad scope of the biodiversity project portfolio. This diversity and corresponding complexity creates particular challenges for evaluation.

The GEF Council recently approved the Evaluation Office proposal to undertake impact evaluations. The stated "objective of this modality will be to evaluate the long-term results of GEF interventions, a few years after GEF support is concluded and to assess the sustainability and replication of the support as well as to extract lessons learned (GEF, 2006a: 6). Following the Council's decision, an approach paper (GEF, 2006b) was prepared by the Evaluation Office as an initial step in developing an impact evaluation of GEF biodiversity interventions.

The impact evaluation is basically concerned with two issues: what has happened (in a descriptive sense), and the causal relationships

between the intervention and the changes that have been observed. These constitute the heart of the attribution challenge, i.e. uncovering the effects of an intervention on particular phenomena (while taking into account the influences of other pertinent variables). While a comprehensive treatment of all intervention activities lies beyond what is practically feasible, at the same time it is expected that the exercise should transcend the limitations of isolated perspectives on the impacts of particular projects or activities within projects. For this reason, we have opted for a portfolio perspective to impact evaluation as opposed to a single intervention perspective. Inherent to this is the challenge of aggregating evidence from single intervention activities to the biodiversity portfolio level, as well as drawing linkages between field level results and global level information resources on biodiversity. In order to address these challenges, the evaluation "will be based on the theory of change underlying specific portfolios and projects, which will form the theoretical base of the impact evaluation" (GEF, 2006b: 13).

3. Methodological challenges

3.1. The problem of the independent variable

A first key question evaluators need to raise is the question of impact of what? Two key issues come to mind: the delimitation issue and the choice of level(s) of analysis. Regarding the former, Pawson and Tilley (1997), among others, illustrate that an exact delimitation of an intervention can be problematic. Rather than constituting a clearly delineated mechanism, an intervention resembles more an open system, a social system embedded in a larger social system in which it is often not easy to determine where an intervention ends and the 'external' world begins. In the case of the GEF evaluators are furthermore confronted with the issue of 'blending' of interventions. For example, in the case of the World Bank as Implementing Agency, GEF grants are often blended with World Bank loans where the GEF project de facto is part of a bigger intervention package. In principle the GEF grant is designed to account for the incremental costs associated with generating global benefits as opposed to the local benefits generated by the loan package. In practice however it is often very difficult to clearly distinguish between the two. Similar problems sometimes occur when GEF projects are part of broader intervention strategies of other Implementing or Executing Agencies.

The second problem, the choice of level(s) of analysis, is one of the key issues to be resolved in portfolio (impact) evaluations. What level(s) of analysis is/are appropriate for making (in the simplest

and most straightforward manner possible) plausible and coherent statements about attribution? For example, should we analyze the impact of projects, activities within projects, operational programs, etc.? Correspondingly, at what level(s) of analysis should evidence about impact be aggregated? For the GEF Council it would be desirable that impact evidence could be aggregated to the portfolio level and put into perspective with global trends, so that changes put into motion by GEF funding can be identified. Apart from that however, one can raise the question whether other types of aggregation of evidence on impact would be useful, especially from the point of view of knowledge management.

This is not easily resolved as different levels of analysis each present their own advantages and disadvantages. The project is the basic administrative unit of intervention and as such presents a natural choice as a focus for impact assessment exercises. In addition, data on performance, outputs and (to a lesser extent) biodiversity-related data are collected and reported at this level. A disadvantage is the fact that projects are not always clearly aligned to higher-level program objectives (GEF, 2004a). This makes it difficult for many projects in the portfolio to aggregate evidence at this level to higher levels of analysis. A second choice would be the level of the operational program, since projects within the portfolio are classified according to the operational program (e.g. Forest Ecosystems) they adhere to. While the ecosystem-related program categories are meaningful at one level, many projects adhere to multiple operational programs. Most importantly, the variety in terms of objectives, activities and institutional structures between projects within one operational program makes this unit of analysis difficult to use for evaluation purposes. The more recent strategic priorities could be useful for evaluative purposes, as the categories (e.g. Catalyzing Sustainability of Protected Areas) represent different groups of projects with (as a group) more coherent objectives and strategies. Nevertheless, the categories are quite general, still harboring a substantial variety of intervention activities. In addition, projects can serve multiple strategic priorities. Finally, the strategic priorities have only recently been introduced and as such are not useful as units of analysis in the impact evaluation, which focuses mainly on completed projects.[3]

[3] Since many of the processes of change induced by GEF interventions are likely to produce observable effects on biodiversity only after a certain period of time, a focus on relatively older projects seems justified.

Given the difficulties associated with these (what can be called) traditional levels of analysis, the evaluators have considered alternatives. An interesting level of analysis is the thematic area of intervention. To a large extent projects (and intervention activities within projects) can be categorized in a coherent manner on the basis of the main theme addressed. Examples of thematic areas of intervention are: protected area management, alternative livelihoods, research on innovative practices, and particular mainstreaming models (e.g. sector-specific legislation). A second interesting level of analysis is that of policy instruments. Policy in-

struments are the basis of public intervention everywhere. Examples of generic policy instruments are: economic incentives (e.g. tax reductions, subsidies), regulations (e.g. laws, restrictions), and information (e.g. education, technical assistance). As argued by several authors (e.g. Salaman, 1981; Vedung, 1998; Pawson, 2006), a classification of different policy instruments recurring throughout the portfolio in relation to specific purposes and contexts can constitute a useful tool for the assessment of effectiveness of interventions as well as institutional learning. "Rather than focusing on individual programs, as is now done, or even collections of programs grouped according to major 'purpose' as is frequently proposed, the suggestion here is that we should concentrate on the generic tools of government that come to be used, in varying combinations in particular public programs" (Salaman, 1981: 256). Acknowledging this central role of policy instruments enables evaluators to take into account lessons from the application of particular (combinations of) policy interventions elsewhere, in the first place relating to the field of environmental protection and development, but also beyond (see Bemelmans-Videc and Rist, 1998).

3.2. The problem of the dependent variable

A second key issue is the question of impact on what? An important consideration concerns the question on which point in the causal chain between intervention output and final (desired) impact one should focus? The primary objective of the GEF is to generate global environmental benefits. Ideally, the effects of all GEF interventions should therefore be traceable up to changes in global environmental benefits, e.g. in the case of biodiversity (positive) changes at the levels of ecosystems, species and gene pools. However, if the impact evaluation were to concentrate on impact at these levels, its utility would be severely hampered by the substantial challenges of attribution (establishing to what extent environmental changes can be shown to result from GEF interventions) and aggregation (the extent to which localized biodiversity changes resulting from interventions can be seen to contribute to higher-level (ideally) global changes). More specifically the following complicating factors play a role.

The nature of environmental change. The complexity of processes of environmental change continues to be a challenging and elusive area of scientific inquiry. Large-scale scientific efforts such as the recent Millennium Ecosystem Assessment (MEA) have contributed to strengthening the scientific consensus on a number of issues regarding environmental processes, more particularly the key role of ecosystems in sustaining life on earth. A promising framework linking ecosystem services[4], their underlying drivers of change and different aspects of human well-being has been developed by this project. On

[4] The MEA distinguishes four main groups of ecosystem services: provisioning (e.g. food, water, fiber, fuel), regulating (climate regulation, water, disease), cultural (spiritual, aesthetic, recreation, education), and supporting (primary production, soil formation) (MEA, 2005).

the other hand, once again many of the limitations in our understanding of these processes have been pointed out. In particular the non-linear nature of environmental change, the time scales over which changes occur and the interaction effects between different drivers of change (e.g. the interplay of climate change and economic activity and the effects on various ecosystem services) are often not well understood. A specific complicating factor is the irreversibility of many environmental change processes (e.g. habitat loss, species extinction); once a certain point of change (a threshold) has been passed a process of restoration towards the former state of the environment is no longer possible (Rao, 2000; MEA, 2005).

The concept and measurement of biodiversity. The GEF has adopted the CBD's definition of biodiversity, which encompasses the diversity in species, gene pools and ecosystems. Biodiversity as a whole as well as the three subcomponents cannot be easily captured by simple indicators and requires multiple indicators representing the different aspects of (genetic, ecosystems, species) biodiversity (Duelli and Obrist, 2003). Comprehensive indicators are often contested as they are clearly value-laden (i.e. including specific dimensions of biodiversity with certain relative weights) and involve adding up different types of biodiversity which (in some cases) might be negatively correlated (ibid., 2003). Regarding the latter, for example there is sometimes a negative correlation between species diversity (e.g. the number of different fish in a lake) and species abundance (e.g. the amount of fish of one species). This trade-off can become problematic when minimal thresholds for species survival are threatened. Despite all this, comprehensive indicators can be very useful in determining priorities for resource allocation. Under the new country-based resource allocation framework the GEF uses a comprehensive indicator of biodiversity encompassing species and ecosystem biodiversity. The GEF biodiversity indicator can be broken down into two components: representation of ecosystems and species diversity, and threats to ecosystem quality and species. The impact evaluation will primarily focus on the second component as GEF interventions are mainly focused on reducing biodiversity threats (GEF, 2006b). Regarding the effect of GEF interventions on levels of biodiversity (representation), data on biodiversity aspects are often not readily available (see below). Measurement of the different aspects is often not straightforward nor easy and therefore can be very costly. Consequently, it can be worthwhile to choose proxy indicators which are highly correlated with multiple aspects of biodiversity (Duelli and Obrist, 2003). Further analysis is needed to reveal what proxy indicators might be useful in such a role. In the case of the GEF, the intensity of land use or the number of hectares of protected area could be useful proxies for biodiversity (GEF, 2004a).

Current data availability. A major input to impact evaluation studies is the existing information base. As a result, evaluators first of all

inquire whether there is useful existing evaluative evidence (at project level) to inform impact evaluation studies. Second, the question arises how existing evaluative evidence (at project level) can be usefully aggregated to inform impact assessment at portfolio level. Recent studies have pointed out the lack of information on impact in existing end-of-project evaluations. (GEF, 2004a; GEF, 2005). In addition, the same studies have reported that in general projects do not have adequate (standardized) reporting systems on biodiversity impact data. The recently introduced strategic priorities and corresponding tracking tools to monitor performance and impact-related indicators represent a positive development towards such a reporting system and increasingly, projects are systematically collecting data on biodiversity conservation and sustainable use.

The type of intervention supported by the GEF. In recent years, in the GEF biodiversity portfolio there has been a relative shift from site-specific interventions to interventions that support a broader agenda of advancing biodiversity concerns not directly related to a specific site. Many of the latter group encompass intervention activities at national or regional (group of countries) level while also including localized intervention activities as pilot and demonstration sites of some of the principles promoted at higher levels of administration. The causal chain connecting this type of intervention activities to biodiversity variables is often more indirect and diffuse than in site-specific intervention activities, making it more difficult to resolve the attribution problem. For example, it is hard to establish clear causal links between enhanced political will of a national government to put biodiversity on the political agenda (e.g. as a result of a GEF-funded national policy dialogue process) and actual changes in biodiversity indicators, given the large number of intermediate steps (from political will to resource allocation to policy design to policy implementation etc.), the influence of other variables (e.g. other policy priorities, resource constraints, institutional alliances, institutional capacities, etc.) and the uncertain time path of these processes. At the same time, there is an explicit interest from within the GEF to measure impact of its interventions at intermediate levels of the causal chain towards biodiversity conservation and sustainable use. Given the increasing importance of interventions focusing on issues like awareness and political will, policy design and implementation capacity, institutional collaboration and coordination, there is a growing demand for knowledge about in what ways and to what extent the GEF has achieved positive results in these fields.

The problems of attribution and aggregation discussed above have led some previous studies to conclude that the impact of GEF activities is best measured at the level of behavioral changes of

actors (e.g. GEF, 2003). First, as discussed in the previous paragraph, this refers to the behavior of individuals and institutions that influence policies and markets, which in turn (in)directly affect biodiversity variables. Second, it refers to the behavioral changes among end users of natural resources (e.g. farmers, fishermen, the public, etc.), more specifically, behavioral changes in the conservation, sustainable use and benefit sharing of biodiversity. Analogous to the institutional level, impact at the level of behavior of end users of natural resources represents an important intermediate level of impact relevant to actors within (and outside) the GEF network. In what ways and to what extent have GEF interventions changed the behavior of these actors? What types of interventions in what settings are the most successful? From that point onward one can venture further down the causal chain towards changes in biodiversity. In some cases the links between certain patterns of behavior and biodiversity are straightforward and attribution issues can be resolved relatively easily; for example, the link between an increase in intercropping systems and on-farm biodiversity (e.g. in terms of plants, insects and birds). In other cases, one can only assume that there is a positive causal link between behavior and biodiversity on the basis of existing (scientific) evidence. In the worst case, the causality is highly contested as the current state of the art of knowledge about the interplay of different variables and their effect on biodiversity is insufficient to draw conclusions about causality and attribution (from human behavior to environmental change).

Apart from the specific challenges in attribution and aggregation faced by evaluators when assessing the impact of GEF interventions on biodiversity changes there are other reasons for shifting time and resources to more intermediate impacts. One of the ten GEF operational principles emphasizes the catalytic role of the GEF in its mission of seeking to maximize global environmental benefits. More particularly the GEF seeks to induce catalytic effects in at least three different ways (GEF, 2005):

- to maximize co-financing contributions and leveraging resources from other institutional actors;
- to maximize the replication of successful GEF intervention approaches at different levels;
- to promote mainstreaming of environmental concerns in (sector) policies and legislation.

Some of the key mechanisms through which these effects are expected to occur are: fostering awareness and political will to act on biodiversity concerns; fostering institutional alliances and partnerships among public and private actors; building capacities; supporting research projects; and demonstrating and disseminating good practices on innovation. In principle catalytic effects can occur at different levels. Many types of catalytic effects (especially replication effects) occur more or less spontaneously. Recently, GEF interventions are increasingly developing

specific strategies to maximize the catalytic role, for example by focusing on innovation, demonstration and dissemination, or by designing explicit replication strategies. The methodological implication of the foregoing is that from the point of view of the GEF network, there is a strong demand for evidence on the achievements of GEF interventions in terms of the three types of catalytic effects. This is to be addressed by an evaluation on catalytic effects which started in late 2006. The subsequent links to biodiversity changes are far more difficult to assess. Causality is diffuse, complex and subject to many external (context-specific) influences. As a result, evaluators' work would be first and foremost centered around the question of how these relationships work in particular contexts. The question of attribution of changes to GEF interventions let alone the determination of the magnitude of intervention effects would be largely out of reach. Not only the causal relationships with biodiversity changes are important, evaluators need to be aware of other types of (unintended) effects. For example, catalytic effects induced by GEF interventions might result in trade-offs between biodiversity issues and other public spending.

A final note regarding the 'dependent variable' concerns the dimension of sustainability. Previous studies, in line with their conclusions on impact assessment, have signaled the potential difficulties in sustainability assessment (GEF, 2004a, GEF 2005). Questions about the sustainability of impacts are often even more shrouded in fog than questions of attribution of changes to an intervention. Sustainability is a highly contested concept that is difficult to pin down in terms of indicators or fixed goals (Mog, 2004). Nevertheless, evaluators can make headway by looking into the factors that make it more or less likely for particular changes to be sustainable (ibid.). In doing so, they should distinguish between different relevant units of assessment (e.g. institutions, ecosystems) and different dimensions of sustainability (e.g. financial sustainability, ecological sustainability)[5]. Examples of questions evaluators could ask are the following. Are particular institutional structures (e.g. management structures of protected areas) likely to be financially sustainable? Are technological innovations (e.g. intercropping systems) likely to be appropriated and integrated into existing practices? Are particular enabling environments (e.g. political dialogue, institutional collaboration, legislation on biodiversity) likely to be politically sustainable? Are particular practices (e.g. selective harvesting of non-timber forest products) likely to have a lasting positive influence on biodiversity variables (e.g. ecosystem quality)? What are the main contextual variables obstructing/enabling these processes? To some extent, these questions can be translated into measurable indicators. However, the scope of such questions is almost endless and, as a result, a challenge for evaluators lies in the 'economical' incorporation of sustainability concerns in the overall impact exercise.

[5] The biodiversity program study (GEF, 2004a) briefly discusses different dimensions of sustainability relevant to biodiversity interventions.

This is a particular challenge for the GEF, where the concept of a global environmental benefit implicitly incorporates the concept of sustainability, since it attempts to counter current unsustainable patterns of natural resource use.

3.3. Methodological responses to the attribution and aggregation challenge

Before we introduce the basics of the methodological approach applied in the GEF impact evaluation it is worthwhile to reflect briefly on the current methodological debate on impact evaluation.

In several policy fields such as health, education and criminal justice, and to a lesser extent development interventions, 'rigorous impact evaluation' is mostly equated with randomized controlled trials (RCT) or close derivatives (quasi-experiments). The core idea is that observed changes can only be interpreted if they are objectively compared to a counterfactual situation (i.e. that which would have happened without the intervention). In the case of RCTs this works by randomly separating an 'intervention' group from a control group for the duration of an intervention. As a result, differences in target variables between the two groups can be attributed to the intervention as for all other variables conditions are the same (due to the random participation in the intervention). If random assignment is not possible, control groups are constructed to reflect intervention groups as closely as possible in order to be able to attribute differences in target variables to the intervention. Several variations of this principle are applied[6], depending, among other things, on data availability (before and after) and budgetary constraints (Rossi et al., 2004; IEG, 2006).

[6] Basically, one can discern two groups of approaches. The first group of approaches employs experimental design as a basis for isolating intervention effects (preferably before an intervention has started, enabling ex ante – ex post comparisons). The second group relies primarily on advanced statistical analysis to isolate intervention effects from other influences. The latter group of approaches is mostly applied in cases where there are insufficient or no experimental design controls.

Nowadays, in the fields of evaluation and applied policy analysis one can notice a strong current in favor of more applications of (quasi-)experimental impact evaluation (CGD, 2006), proponents perceiving this type of methodology (either as a stand-alone procedure or embedded in a mixed method design) as the most rigorous and trustworthy way to resolve the attribution issue. Moreover, by capturing impact in terms of effect sizes they generate indications of the magnitude of impact. Very importantly, results of single impact evaluation procedures can be relatively easily aggregated by using quantitative meta-analysis.

Why then are there so few applications of this type of rigorous impact evaluation in development intervention (including the particular field of environment and development in which the GEF operates)? A few reasons can be stated that apply to impact evaluation in general. First of all, the results stemming from rigorous impact evaluation studies are

usually freely available and to a large extent very useful to different organizations working on similar intervention activities. Consequently, individual organizations are facing a disincentive to engage in rigorous in-depth impact evaluation[7], as useful results might be produced by others and therefore available at little or no cost. In addition, there might be other disincentives such as the fear of finding negative impacts or insufficient positive evidence which might put in jeopardy future support for funding (Pritchett, 2002).

[7] Instead opting for cheaper less in-depth studies.

Other reasons why there are relatively few (quasi-)experimental evaluations are the following. First of all, they can be very costly and time-consuming. For example, the World Bank, since 1980 has conducted only 23 of this type of evaluations with costs ranging between US$ 200,000 and US$ 900,000 while taking sometimes more than two years to complete (OED, 2005). Second, there are a number of technical and practical considerations which raise the threshold of doing this type of evaluation. These include the high demands in terms of statistical analysis skills, and planning and organization of experimental designs. Regarding the latter, studies are mostly of a quasi-experimental nature as randomization in social policy is often simply not possible (people cannot be excluded at random) or unethical to implement (withholding benefits from particular people while providing them to others). Another constraint concerns the fact that impact evaluations are often not part and parcel of the regular policy cycle and are often commissioned ad hoc. Rigorous (quasi-)experimental evaluation on the other hand (ideally) requires careful planning from the start of an intervention, enabling an adequate set-up of the design as a basis for reliable baseline and ex post data (Rossi et al., 2004). Finally, there are also critical signals stemming from academic debate which raise doubts about the 'superiority' of quasi-experimental evaluation from a conceptual-methodological point of view, and as a result dampen enthusiasm for application. An important critique comes from the field of 'realist evaluation'. This critique is mainly centered around the reductionist nature of quasi-experiments and meta-analysis. It highlights elements such as the incorrect equation of apparently similar intervention mechanisms (e.g. several projects on health education) which in reality might work in different ways, the oversimplification of outcomes, and the concealment of intervention contexts (Pawson, 2002; see also Pawson, 2006).

Each of the above-mentioned points is relevant for the GEF impact evaluation and raises justifiable concerns about the potential utility of quasi-experiments in the impact evaluation design. The limited budget for the GEF impact evaluation would not permit conducting a rigorous quasi-experimental evaluation unless one would be willing to accept a substantial loss in scope, reducing the range of lessons which can be generated to help improve future performance.

Yet the most compelling argument for choosing not to apply a (quasi-)experimental methodology concerns the nature of many GEF interventions. As discussed earlier, the growing importance of GEF interventions directed at awareness building, natural resource management systems, legislation design, capacity building, political support at national or regional level, as well as other catalytic effects central to the GEF's role in supporting the global environment cannot be adequately assessed on the basis of quasi-experimental designs. These interventions and their intended effects are completely different from the relatively well-delineated site-specific interventions with clearly identifiable target groups which usually are the subject of quasi-experimental impact evaluation In global environment interventions it is much more difficult to isolate the intervention from the wider institutional and policy environment, while effects are complex, diffuse and uncertain, making it impossible to determine counterfactuals.

Given the complexity and uncertainty surrounding GEF impacts on biodiversity, the evaluation should be at least as much about generating insights about processes of change instigated and/or influenced by GEF interventions as about the actual demonstration of change attributable to the GEF. In practice, the latter cannot be established in a reliable manner without the first. In general, one can question, at least for the type of intervention activities sketched above, whether the determination of attribution of changes to GEF interventions is at all realistically possible. Accordingly, some authors talk about contribution instead of attribution, which basically entails a more comprehensive perspective on causality without a claim on determining the precise (magnitude of) causal effect from the intervention to change the dependent variable (Van den Berg, 2005).

Based on the foregoing, the impact evaluation must begin by mapping different processes of change related to different intervention activities within the portfolio. Then, at different levels of analysis, more precise data will be gathered to establish more precise claims of attribution. Theory-based evaluation constitutes a suitable framework for this type of approach. In the next section we will discuss some elements of a theory-based approach and how they help to resolve some of the methodological challenges sketched earlier.

4. A THEORY-BASED IMPACT EVALUATION APPROACH

Over the past two decades theory-based evaluation has developed into an important methodological current in evaluation theory and practice (see for example Weiss, 1997; Rogers et al. 2000; Donaldson, 2003). Although particular theory-based approaches[8] differ in terms of the way theory is perceived and handled in evaluation, all approaches share the basic idea of theory as a set of assumptions underlying the way an intervention is supposed to work (i.e. the intervention theory). Consequently, the task of evaluators lies in reconstructing the main assumptions that underlie an intervention and subsequently, testing whether these assumptions are valid.

[8] Theory-based evaluation (introduced by Weiss) is probably the most commonly used term. Other terms are used in the literature to refer to broadly similar approaches.

4.1. Addressing the attribution challenge

Let us briefly outline the two methodological steps that are at the heart of theory-based evaluation: theory reconstruction and theory testing. A common interpretation of an intervention theory is that it starts out from a systematic representation of the expectations and assumptions held by intervention staff and decision makers. These intentions and assumptions are only in part made explicit in formal documents (such as formal logical frameworks) and thus require further reconstruction. In other words, it is not a priori altogether evident what an intervention actually 'is', what it is meant to achieve and how it is meant to achieve it. There are several methodologies available for reconstructing intervention theories. A particularly useful methodology is called the policy-scientific approach (Leeuw, 2003). It is based on a five-step procedure to uncover the main assumptions that make up an intervention, mainly based on documents (official and working documents produced by decision makers and staff) and interviews. The reconstruction process is finished as soon as evaluators feel that the theory constitutes a balanced and realistic representation of the major intentions and assumptions held by decision makers and staff in a particular organization or network of organizations related to the intervention. Klein Haarhuis and Leeuw (2004) for example apply this approach in an evaluation on the World Bank anti-corruption program. A variation on this procedure is illustrated by Carvalho and White (2004) who analyze the impact of the World Bank social fund program. They show how for different assumptions, so-called 'anti-theories' can be defined, for example embodying the assumptions of opponents of particular types of interventions. The subsequent assessment process serves the purpose of adjudicating between the rival theories and ultimately arriving at better explanations of processes of change leading to impact.

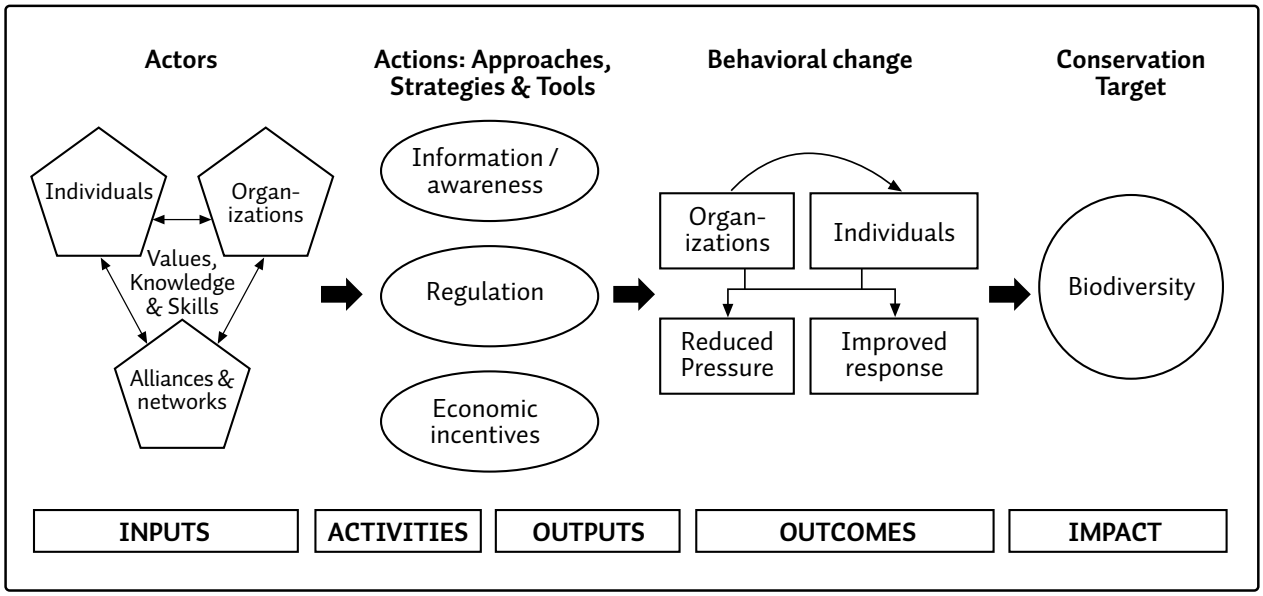
This brings us to the second issue of how to test an intervention theory. To what extent are the assumptions underlying the theory valid? The specific methods and approaches that evaluators can use to test the intervention theory are almost endless. In this sense, theory-based evaluation is not method-specific. For example, an intervention theory can constitute the basis for a quasi-experimental evaluation. Yet in many theory-based evaluation exercises the principle of analyzing causality and attribution works quite differently from quasi-experimental evaluations. Rather than trying to control for all the possible exogenous influences on impact variables in order to determine an intervention's effect on these variables, the evaluator relies primarily on logical argumentation, by carefully tracing all the assumptions underlying the theory (from inputs to outputs to impacts). Depending on the type of assumption, different sources of evidence come into play. For many types of assumptions (e.g. the influence of local norms and beliefs on institutional performance) 'hard' evidence is difficult to come by, nor can evidence always be guaranteed to be collected in a 'scientifically rigorous' manner. As a result, theory-based evaluations rely on the principle of triangulation of methods and sources of information, bringing as much evidence as possible into play (from different perspectives) in the assessment of hypotheses and assumptions. In many cases, there is no clear distinction between reconstruction and assessment as evaluators start out from a simple intervention theory and gradually work towards a more refined, empirically tested intervention theory that will help draw conclusions on attribution as well as serving as a basis for institutional learning.

In the GEF impact evaluation intervention theories are reconstructed at two levels: project level and portfolio level intervention theories. Here we largely focus on the latter. From the perspective of the portfolio, the most coherent evaluand at the highest level of aggregation is the 'thematic area' of intervention (e.g. protected areas, ecotourism, biosafety), representing major intervention strategies. A first step in the impact evaluation is to identify the main thematic areas and their respective weights in the biodiversity portfolio. For example, in a desk study in preparation of the impact evaluation, covering a sample of 30 projects from the biodiversity portfolio, it was found that, 20 out of 30 projects comprise intervention activities on land, water or species management. Furthermore 15 out of these 20 projects include activities on protected area management and 11 out of 20 develop activities on compatible resource use.

Figure 1 shows the generic results chain underlying the logic of many conservation projects (see GEF, 2006c). This model provides a useful starting point to elaborate more specific intervention theories of strategies at portfolio level. The basic idea is one of a network of actors collaborating in the implementation of a number of actions designed to

induce certain (behavioral) changes in organizations and/or individuals which ultimately affect biodiversity.

Figure 1. Basic intervention theory conservation projects



Source: Adapted from GEF (2006c).

For the purpose of impact evaluation it is useful to distinguish process theory and impact theory (Rossi et al., 2004). The former refers to the assumptions and expectations underlying the processes of inputs leading to outputs, while the latter concerns the assumptions regarding particular outputs inducing processes of change resulting in final impacts. This differentiation is important in order to determine whether an observed lack of change is (mainly) due to problems of implementation, often referred to as implementation failure, or whether the concept of intervention (the idea that particular outputs lead to desired impacts) is fallacious, which is called theory failure (Suchman, 1967). As a result, in impact evaluation evaluators should have at their disposal substantial data about intervention outputs and the implementation process producing these outputs as a basis for further analysis. Only then can the more complex question of attribution in impact theory, i.e. the interaction between intervention outputs and external variables, the (potentially) complex and diffuse causal chain linking outputs to impacts, be addressed.

In this paper we focus on impact theory, particularly the question of how common combinations of policy instruments under certain circumstances contribute to processes of change in institutions and the behavior of end users of natural resources and ultimately affect biodiversity variables. In practice, there is often a strong association between the thematic area of intervention and certain combina-

tions of policy instruments that occur throughout projects. For example, protected area management relies on policy instruments like: capacity building intended to strengthen the institutional and financial sustainability of the protected area framework; imposing restrictions on land use and natural resource exploitation with the intention of generating changes in land use and natural resource exploitation; and awareness raising on natural resources in order to support changes in land use and natural resource exploitation. This connection between the thematic area of intervention and policy instruments is crucial as theories on the effectiveness of (combinations of) policy instruments constitute essential building blocks of intervention theories at thematic area level.

Some work on establishing this type of connection has already been done. For example, in another desk study covering medium-sized and full-sized projects from three different portfolios[9] an inventory of site-specific interventions related to agriculture was made (GEF, 2006d). First, major thematic areas were identified. It was found for example that approximately 59 % of all the projects included alternative livelihoods activities, 40 % ecotourism, 49 % sustainable land use techniques, 14 % reforestation, etc. In addition, an inventory was made of different policy instruments. For example, 14 % of the projects included microgrants (economic incentives), 24 % microcredit (economic incentives), 42 % education and awareness building (information/awareness), 26 % technical assistance (information/awareness), 32 % community-based natural resource management (regulation), and 8 % land user agreements (regulation). Further analysis would make it possible to identify the major patterns of thematic areas linked to particular combinations of policy instruments.

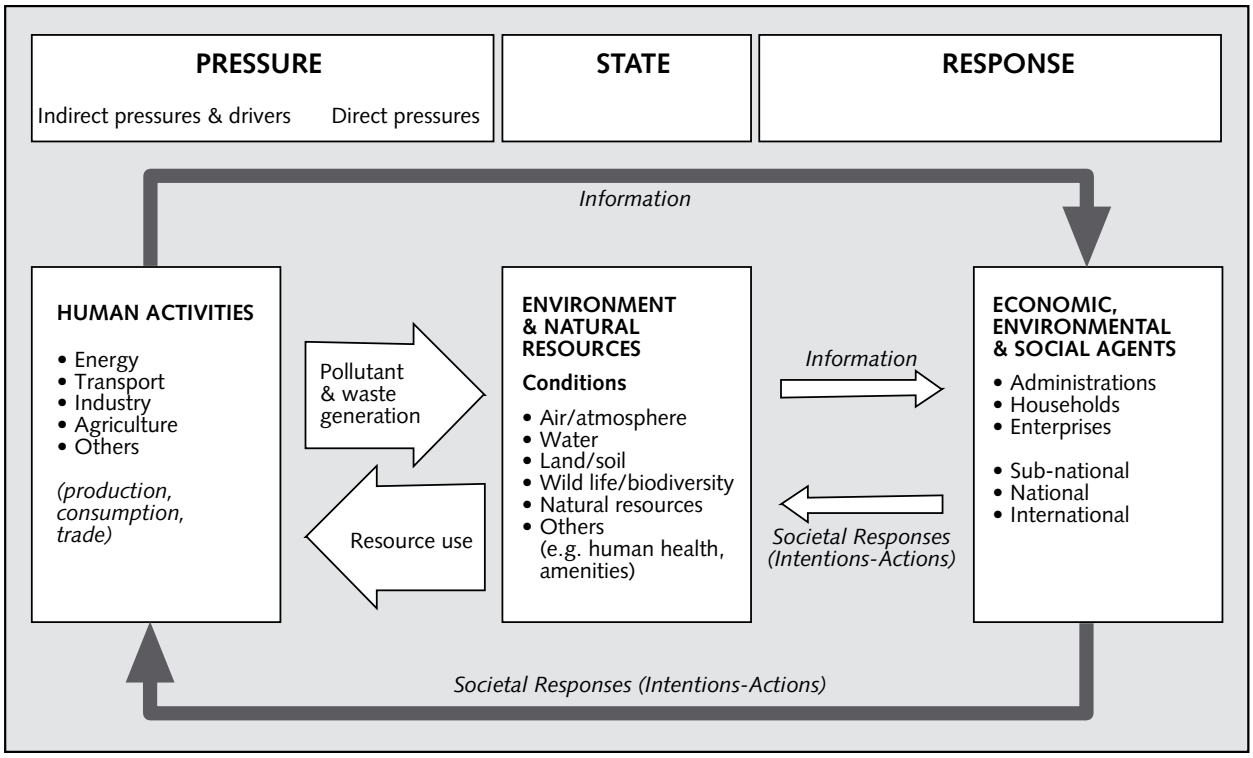
[9] Biodiversity, Land Degradation and Multi-Focal Areas. All medium-sized and full-sized projects approved between January 2000 and June 2005 (n = 332).

The causal chain linking GEF intervention outputs to changes in the behavior of end users of natural resources can be relatively short and straightforward, for example in site-specific GEF interventions that directly target end users of natural resources (e.g. technical assistance resulting in crop diversification). In other cases, such as in GEF intervention activities directed (at least in the first instance) at changes at institutional level (e.g. capacity building resulting in improved legislation), the subsequent causal effects on end users of natural resources are more diffuse and difficult to capture. In the latter case evaluators require a workable model in order to better conceptualize these potentially complex processes.

The same goes for the human behavior-environment interface which is the most complex part of the impact theory. In a few cases the linkages between human behavior and environmental change are straightforward and the assumptions connecting changes in institutional and individual human behavior to changes in environmental benefits can be reconstructed and tested in a relatively simple manner. In other cases,

the evaluator is facing the frontier of the state of the art of research in the natural sciences (e.g. biology, ecology) and can only make very rough assumptions about these complex causal relationships. Several models available in the literature can assist evaluators in the reconstruction of useful intervention theories. An example is the so-called Pressure-State-Response model (PSR) developed by the OECD in the 1990s (OECD, 2003). The basic logic of the model is represented in figure 1 and more elaborately illustrated in figure 2. The model helps to classify the effects of GEF interventions in terms of reducing pressures on the environment, improving the state of the environment or improving responses by institutional actors. In addition, for each category of effects the GEF's contribution can be assessed in relation to other influencing variables. In short, evaluators can conduct PSR analyses for selected projects and on this basis will be able to articulate an intervention's influence on wider processes of environmental change.

Figure 2. The Pressure-State-Response Model



Source: OECD (2003).

In the complex impact evaluation under discussion it is important to arrive at usable abstractions of the GEF's strategies and their effects on the global environment. However, one has to keep in mind the reductionist nature of such abstractions. So, the evaluator team may also need to develop some detailed case study analyses to complement the process of reconstructing and refining intervention theories. A narrative historical approach can be very useful to generate additional understanding about the complex linkages between GEF

interventions, the context in which they operate and possible outcomes and impacts. Such an approach would focus on the evolution of a particular GEF intervention or a series of GEF interventions in a particular region or country (e.g. in a biodiversity 'hot spot'). The in-depth illustration of the embeddedness of current GEF interventions in past interventions and strategies of other institutional actors will be particularly useful. In addition, of particular interest will be qualitative analyses of the sustainability of effects and the different patterns of replication, since these concepts are multi-dimensional and relatively difficult to analyze empirically. The evaluators will thereby gain a more detailed understanding of the complex interactions between interventions and social-institutional and environmental dynamics and will be able to develop a more detailed picture of the induced processes of change and new impacts on biodiversity.

4.2. Addressing the aggregation challenge

We can make a distinction between the aggregation of quantitative data from project level to higher levels of intervention (thematic area, portfolio, global trends) and the process of generalization or theory-building from project level intervention to impact theories at the level of thematic areas of intervention. Regarding the latter, intervention theories can potentially constitute a powerful basis for institutional learning and knowledge management on impacts of different types of GEF interventions in different settings. To develop a feeling of how this might work, let us briefly highlight a few points of attention on intervention theory-building relevant to the impact assessment of the biodiversity program.

Starting out from a generic results chain as shown in figure 1 and using information from different project documents of projects pertaining to a particular thematic area as well as other sources of information (e.g. existing literature, interviews with project staff), crude intervention theories can be reconstructed representing the basic causal linkages between inputs, activities, outputs and impacts. Subsequently, evaluators further refine and test these intervention theories on the basis of multiple sources of information, the main sources being: information available at project level (progress reports, end-of-project evaluations, field visits, staff interviews), existing evaluative evidence within the GEF (e.g. program studies, thematic cross-cutting studies, overall performance studies), studies on similar interventions elsewhere, expert interviews, and academic literature.

The process of intervention theory refinement basically works as follows. After the initial reconstructions, evaluators identify key assumptions to be tested and correspondingly define more focused study questions and indicators. These assumptions, questions and indicators

will provide a structure for more systematic empirical data collection at project level. In this process the thematic area intervention theory is more and more refined taking into account contextual factors of different projects pertaining to a thematic area. Pawson and Tilley (1997) introduced the principle of context-mechanism-outcome as the basic ingredients for theory-building about what works (for whom) under what circumstances. In our case, the team will look for outcome patterns at thematic area level, the major mechanisms contributing to these patterns (institutional structures, combinations of policy instruments) and the particular contextual settings (at project level) that condition these mechanisms. Project level information is subsequently fed back into the theory in order to further refine the assumptions. This process can be repeated in an iterative manner until the best possible explanatory model is achieved.

Intervention theories can be developed in relation to different levels of impact. As discussed earlier, depending on the type of intervention, and state of knowledge and available data on processes of environmental change, impacts are captured at four levels: catalytic effects[10], institutional changes (e.g., capacities, awareness, political will), behavioral changes of end users of natural resources (e.g. land use changes, reductions in harvesting of natural resources), and changes in biodiversity variables (e.g. species diversity, ecosystem quality). The evaluators will concentrate on developing causal theories that link particular thematic areas of intervention to the levels of impact which (in first instance) are deemed most relevant for these areas. For example: under what circumstances have the main thematic strategies of the GEF subscribing to the (broad) purpose of institutional change (e.g. enabling activities, biosafety projects) induced positive and sustainable changes at institutional level? Under what circumstances have the main thematic strategies of the GEF subscribing to the (broad) purpose of directly influencing the behavior of end users of natural resources (e.g. projects on mainstreaming) induced positive and sustainable changes at this level? Subsequently, the causal analysis is extended to biodiversity targets by using among other things the PSR model introduced earlier.

[10] Given the fact that the GEF Evaluation Office will organize a separate evaluation on the catalytic role of the GEF in the near future, the impact evaluation will not provide an in-depth coverage of these effects

In practice, the majority of interventions comprise objectives and activities which explicitly aim at inducing institutional change, behavioral changes of (groups of) individual users of natural resources as well as indirect (catalytic) processes bearing on biodiversity variables elsewhere. The above-mentioned evaluation strategy of linking thematic areas to levels of impact therefore should be triangulated through case study analyses (and other sources), of the type discussed earlier, in order to analyze the interplay between different interven-

tion mechanisms (policy instruments, institutional structures), specific contextual variables and effect patterns at all levels of change.

The second dimension of aggregation, the definition and measurement of impact indicators is related to the foregoing. The intervention theory is the principal basis for indicator development. In addition, it constitutes the basis for other types of empirical assessment regarding the contribution of GEF interventions to processes affecting biodiversity variables and intermediate levels of change.

Existing problems of alignment and aggregation of project level information cannot be fully compensated by the evaluation's data collection activities. The number and diversity of GEF interventions makes it too costly to be able to define relevant indicators and collect data for all types of interventions. As mentioned earlier the knowledge deficit on particular environmental processes poses an additional barrier to indicator development. This is the main variable which determines the type and precision of indicator to be measured.

Roughly, we can distinguish between three levels of precision. The scope for collecting and aggregating precise quantitative data on different types of biodiversity is limited. This may be feasible with regard to specific endangered species, such as the giant panda, which has been the target of a number of GEF projects, including the Qinling Forest Reserve project in China (GEF, 1995). For projects with broader aims, such as the protection of mangrove forests, it is likely to be far more difficult to define a specific set of impacts with corresponding measurable indicators. In projects in which data on biodiversity are not available but where it is relatively easy to identify the nature of the causal linkages (i.e. the intervention theory) between intervention outputs and processes of (environmental) change, the PSR model illustrated in figure 2 can constitute the basis for defining questionnaires and indicators in order to collect ordinal data concerning intervention effects on environmental variables. This type of data can also be relatively easily aggregated across interventions and compared to national or international trends. Finally, in projects that comprise intervention activities about which little is known regarding possible causal patterns towards environmental change, more attention should be paid to intervention theory reconstruction (preferably complemented by field assessments, expert interviews, and consultation of academic literature). This will provide the basis for collecting data on relevant proxy indicators of biodiversity (see Duelli and Obrist, 2003) such as the number of hectares of protected area, the intensity of land use or agricultural diversification. In addition, it is advisable in these cases to focus more on intermediate levels of impact.

5. Conclusions

The impact evaluation team faces a formidable challenge in assessing the impact of a portfolio which is complex in terms of both the diversity of its underlying interventions, and the nature of the interventions and the potential processes of change they aim to bring about.

In the first part of the paper we addressed the methodological challenges associated with the impact evaluation. Two principal lessons have emerged. First, a focus on thematic area of intervention and combinations of policy instruments underlying intervention strategies is recommended as it offers advantages in terms of addressing the issue of attribution while at the same time offering a structure for aggregating evidence and lessons on impact across interventions. Second, due to practical limitations and knowledge constraints as well as the nature of GEF interventions, evaluators should analyze causality at various points in the causal chain between outputs and impacts. In practice, this implies that evaluators address causal links between intervention outputs and changes at institutional level, behavioral changes at the level of end users of natural resources, or changes in biodiversity variables.

In this paper we introduced the notion of theory-based evaluation as a basic framework for the impact evaluation. In the field of policy evaluation different contexts and purposes of evaluation have resulted in a variety of theory-based approaches with divergent methodological and procedural features. We have sketched the foundations of a particular approach of intervention theory-building which first of all should help to systematize and deepen the understanding of processes of change instigated by GEF interventions. Subsequently the intervention theories should facilitate processes of gathering evidence on impact at different levels of intervention while at the same time providing a structure for logical argumentation on impact. Intervention theories should be defined, tested and refined at the level of thematic areas of intervention. In this process, evaluators focus on the trinity of patterns of outcomes linked to specific (combinations of) policy instruments embedded in particular contexts. Initially, intervention theories are reconstructed by focusing mainly on the expectations and assumptions harbored by staff and stakeholders at different levels in the GEF network (through document review and interviews), complemented by insights from documentation external to the GEF. Subsequently, these initial theories constitute the basis for indicator development, data collection at project level and further desk studies.

Theory-based evaluation provides a useful framework of analysis to confront the challenges of attribution and aggregation in a balanced way. Nevertheless, there are limitations to the extent that useful abstractions of processes of change leading to impacts can be constructed. The complexity surrounding each individual GEF project context can only be captured in a very limited way by the higher-level intervention theories. Consequently, the exercise of theory-building across interventions in itself is insufficient to grasp the nature of processes of change induced by GEF interventions and therefore will be complemented by case studies, i.e. project-specific or region-specific inquiries on impact, using the most appropriate mix of qualitative and quantitative methods.

We believe that the impact evaluation should not be viewed as an isolated exercise. Though limited in terms of the number of project-specific studies to be undertaken and the number of intervention theories to be reconstructed and tested in a detailed manner, the approach could form an important model for other GEF evaluation studies and a foundation for institutional learning and knowledge management on strategies to maximize the impact of GEF interventions. Without being exhaustive we mention a few options. First, conventional meta-evaluations of end-of-project evaluations focus on extracting lessons on both the content (e.g. performance of projects) as well as the quality of project level evaluations. In the future, as more and more GEF projects start reporting systematically on outcome and impact-related data, such exercises could be expanded by synthesizing impact-related evidence into portfolio-level statistics as well as more qualitative intervention theories covering major thematic areas of intervention. Improved guidelines on the type of information that should be reported at project level would benefit this type of analysis. Second, there are a number of learning projects on biodiversity issues currently funded by the GEF and managed by the Implementing Agencies. The insights produced by these projects would constitute important ingredients of knowledge management activities on GEF impact at portfolio level, complementary to the insights generated by existing thematic and meta-evaluations.

References

Bemelmans-Videc, M.L. and R.C. Rist (eds.) (1998) **Carots,Sticks and Sermons - Policy Instruments and Their Evaluation**, Transaction Publishers, New Brunswick.

CGD (2006) *"When Will We Ever Learn? – Improving Lives through Impact Evaluation"*, **Report of the Evaluation Gap Working Group**, Center for Global Development, Washington D.C.

Carvalho, S. and H. White (2004) *"Theory-based Evaluation: The Case of Social Funds"*, **American Journal of Evaluation** 25(2): 141-160.

Donaldson, S.I. (2003) *"Theory-Driven Program Evaluation in the New Millennium"*, in: S.I. Donaldson and M. Scriven (eds.) **Evaluating Social Programs and Problems – Visions for the New Millennium**, Lawrence Erlbaum Associates, London.

Duelli, P. and M.K. Obrist (2003) *"Biodiversity Indicators: The Choice of Values and Measures"*, **Agriculture, Ecosystems and Environment** 98: 87-98.

GEF (1995) **"People's Republic of China - Nature Reserves Management Project"**, Project Appraisal Document, Global Environment Facility, Washington D.C.

GEF (2003) **"Measuring Results of the GEF Biodiversity Program"**, Monitoring and Evaluation Working Paper 12, Global Environment Facility, Washington D.C.

GEF (2004a) **"Biodiversity Program Study"**, Global Environment Facility, Washington D.C.

GEF (2004b) **"Mainstreaming Biodiversity in Production Landscapes and Sectors"**, Discussion Paper, Global Environment Facility, Washington D.C.

GEF (2005) **"OPS3: Progressing Toward Environmental Results"**, Third Overall Performance Study, Global Environment Facility, Washington D.C.

GEF (2006a) **"Four Year Work Program and Budget of the Office of Monitoring and Evaluation – FY06-09 and Results in FY05"**, Global Environment Facility, Washington D.C.

GEF (2006b) **"GEF Impact Evaluations – Initiation and Pilot Phase – FY06"**, Approach Paper, Global Environment Facility, Washington D.C.

GEF (2006c) **"GEF Impact Evaluation – Final Report on a Proposed Approach"**, Working Document Prepared for the GEF Evaluation Office, Foundations of Success, Washington D.C.

GEF (2006d) **"Sustaining Global Environmental Benefits through Changes in Farmers' Behavior: A Review of GEF-funded Activities"**, Progress Report, unpublished, Global Environment Facility, Washington D.C.

IEG (2006) **"Conducting Quality Impact Evaluations Under Budget, Time and Data Constraints"**, Independent Evaluation Group, World Bank, Washington D.C.

Klein Haarhuis, C.M. and F.L. Leeuw (2004) *"Fighting Governmental corruption: The New World Bank Programme Evaluated"*, **Journal of International Development 16**: 547-561.

Leeuw, F.L. (2003) *"Reconstructing Program Theories: Methods Available and Problems to be Solved"*, **American Journal of Evaluation 24(1)**: 5-20.

MEA (2005) **Ecosystems and Human Well-Being: Synthesis, Millennium Ecosystem Assessment**, Island Press, Washington D.C.

Mog, J.M. (2004) *"Struggling with Sustainability – A Comparative Framework for Evaluating Sustainable Development Programs"*, **World Development 32(12)**: 2139-2160.

OECD (2003) **OECD Environmental Indicators: Development Measurement and Use**, OECD, Paris.

OED (2005) *"OED and Impact Evaluation – A Discussion Note"*, **Operations Evaluation Department**, World Bank, Washington D.C.

Pawson, R. (2002) *"Evidence-based Policy: In Search of a Method"*, **Evaluation 8(2)**: 157-181.

Pawson, R. (2006) **Evidence-Based Policy – A Realist Perspective**, Sage Publications, London.

Pawson, R. and N. Tilley (1997) **Realistic Evaluation**, Sage Publications, London.

Pritchett, L. (2002) *"It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation"*, **The Journal of Policy Reform 5(4)**: 251-269.

Picciotto, R. (2003) *"International Trends and Development Evaluation: The Need for Ideas"*, **American Journal of Evaluation 224(2)**: 227-234.

Rao, P.K. (2000) **Sustainable Development – Economics and policy**, Blackwell Publishers, Oxford.

Rogers P.J., T.A. Hacsi, A. Petrosino and T.A. Huebner (eds.) (2000) *Program Theory in Evaluation: Challenges and Opportunities*, **New Directions for Evaluation**, 87, Jossey-Bass, San Francisco.

Rossi, P.H., M.W. Lipsey and H.E. Freeman (2004) **Evaluation – A Systematic Approach, Seventh Edition**, Sage Publications, Thousand Oaks.

Salaman, L. (1981) *"Rethinking Public Management: Third Party Government and the Changing Forms of Government Action"*, **Public Policy 29(3)**: 255-275.

Suchman, E.A. (1967) **Evaluative Research: Principles and Practice in Public Service and Social Action Programs**, Russell Sage Foundation, New York.

Van den Berg, R.D. (2005) *"Results Evaluation and impact Assessment in Development Co-operation"*, **Evaluation 11(1)**: 27-36.

Vedung, E. (1998) *"Policy Instruments: Typologies and Theories"*, in: M.L. Bemelmans-Videc and R.C. Rist (eds.) **Carrots, Sticks and Sermons – Policy Instruments and Their Evaluation**, Transaction Publishers, New Brunswick.

Weiss, C.H. (1997) *"Theory-Based Evaluation: Past, Present and Future"*, in: D.J. Rog and D. Fournier (eds.) **Progress and Future Directions in Evaluation: Perspectives on Theory, Practice and Methods, New Directions for Evaluation**, 76, Jossey-Bass, San Francisco.

White, H. (2003) *"Using the MDGs to Measuring Donor Agency Performance"*, in: R. Black and H. White (eds.), **Targeting Development: Critical Perspectives on the Millennium Development Goal**, Routledge, London.

University of Antwerp

IOB

INSTITUTE OF DEVELOPMENT POLICY AND MANAGEMENT