GEF/C.41/Inf. 18
October 24, 2011

GEF Council Meeting
November 8-10, 2011
Washington, D.C.

Experimental Project Design in the Global Environment Facility: Designing Projects to Create Evidence and Catalyze Investments to Secure Global Environmental Benefits

(Prepared by STAP)

Report of the Chairperson of the Scientific and Technical Advisory Panel (STAP) to the GEF Council

Experimental Project Design in the Global Environment Facility

Designing projects to create evidence and catalyze investments to secure global environmental benefits

EXECUTIVE SUMMARY

Background

Budgets to supply local, regional and global environmental public goods are limited. Thus judging the effectiveness of environmental investments in different contexts is essential to ensuring that scarce funds go as far as possible in achieving desired outcomes. To maximize its return on investment and to catalyze investments from other sources towards effective environmental programs, the Global Environment Facility (GEF) needs to generate credible evidence about what works to achieve its environmental goals and under what conditions. The GEF needs answers to questions such as, "How best can we encourage commercial and residential energy customers to adopt energy-efficient technologies and, when they do adopt them, how much does energy consumption, and thus carbon emissions, decline?", "How do we encourage compliance with chemical regulations and, when compliance increases, how much do chemical emissions decline?", "Are marine and terrestrial ecosystems better managed when communities participate in management decisions?", and "How do we best encourage farmers to adopt soil conservation measures and, when they adopt them, how much additional carbon is sequestered?" There are indeed hundreds of similar questions, small and large, to which credible answers could substantially increase the return on GEF investments, either directly through the GEF portfolio or indirectly by catalyzing investments from other actors.

Objective

To help the GEF answer these questions, this advisory document describes one important, but heretofore neglected, approach: designing projects with features of experimental designs to test important questions related to effective project implementation. Unlike other advisory documents by the GEF-STAP, this advisory document does not describe what we know about what works. Rather, it describes how the GEF can contribute to *improving* what we know about what works. The advisory document is intended to induce curiosity, and reduce skepticism, about the applicability of such designs in the GEF portfolio, and to help GEF stakeholders identify projects that may be amenable to experimental design.

Main Messages

The document makes five key assertions:

1. In non-experimental project designs, attributing changes in environmental and social outcomes to project actions, rather than to other factors, is difficult.

2. Experimental and quasi-experimental project designs, when appropriately implemented, can make attributing changes in environmental and social outcomes to project actions easier.

3. Experimental and quasi-experimental project designs are feasible within the GEF portfolio and when feasible, there is little to gain, and much to lose, by giving up experimental control.

4. Experimental and quasi-experimental project designs can be used to test the effects of the main intervention funded by a GEF project, but they can also be used as modules within projects to test important design hypotheses related to project components (e.g., is information more likely to be used by polluting firms when transmitted via free workshops, targeted training for trend-setting firms, or through written material distributed by mail?).

5. Randomization of the project intervention across potential sites or participants is not the only way in which a GEF project can create experimental control. Other sources of experimental control include crafting participant eligibility requirements, randomizing the order of sites or participants when scaling-up a project, and randomizing the marketing of a project among the intended population.

In contrast, the advisory document does **NOT** do the following:

1. Serve as a detailed technical guide to designing experiments and analyzing experimental data. References are provided for readers who require greater depth. GEF-STAP can also connect project designers to relevant experts.

2. Serve as a guide to project monitoring or selecting the appropriate indicators for project outcomes. The decision of what to measure, and how to measure it, is independent of the decision to use an experimental project design.

3. Assert that all, or even most, GEF projects should use experimental designs. One can debate the optimal percentage of GEF projects using experimental designs, but the author believes the current percentage of close to 0% is difficult to defend.

4. Assert that experimental project designs are flawless or that other evaluation designs are uniformly less credible or informative. Context is important and the most important issue for the GEF continues to be asking the right questions.

In summary, this guideline document acknowledges that the GEF takes seriously its responsibility to generate knowledge about program effectiveness. It encourages the GEF's efforts in this domain and emphasizes that experimental project designs are an important component in these efforts. With the goal of transferring knowledge to other settings, some of the GEF project portfolio should aim to answer important questions about generating global environmental benefits in developing nations. Reliability in answering these questions can be improved, when feasible, with experimental project designs and with attention to the underlying theory of program effects and tests of what conditions enhance or detract from project performance.

1. Why does the GEF need experimental project designs?

Does the adoption of energy efficient technology lead to large reductions in energy consumption and emissions in developing nations and, if so, under what conditions do these reductions take place? Do protected areas alleviate or exacerbate poverty in neighboring communities? Do incentive payments promote alternative fuel use, soil carbon sequestration or marine ecosystem protection beyond what would have occurred in the absence of the incentives? Answers to such questions are crucial to maximizing the returns of GEF investments, but they are hard to uncover.

Consider a certification program designed to incentivize production practices believed to enhance global environmental benefits. These production practices could be related to any of the GEF's focal areas or cross-cutting programs; e.g., shade-grown coffee, sustainably harvested fish or wood, certified renewable energy, certified energy-efficient or dioxin-free products, or sustainably-cropped agricultural products. To make matters concrete, consider a shade-grown coffee project that is intended to induce farmers to retain native forest canopy above their coffee trees, thereby providing biodiversity habitat. The project hypothesizes that farmers are cutting down canopy to plant sun-loving coffee varieties because economic returns are higher than with shade-grown coffee. However, with access to a market and, perhaps, a premium for certified shade-grown coffee, some farmers would instead leave their forest canopy intact.

Assume that we have a baseline measure of forest canopy on farms that agree to participate in the program. Five years after the program starts, we go to these farms and observe that 98% of the original canopy is intact. What can we infer? Observer X infers that the certification created an incentive to protect the canopy and thus we observe most of the canopy intact (implicit assumption: less of the canopy would have been intact had there not been the certification program). Observer Y infers that farmers who were not intending to cut their canopy in the absence of the certification program are most likely to participate and thus we observe most of the canopy intact (implicit assumption: the same amount of canopy would exist had there not been the certification program). Observer Z notes that farm-gate coffee prices dropped around the time the program was implemented and thus infers incentives to cut down canopy to plant sun-loving coffee disappeared (implicit assumption: the same amount of canopy would have been intact had there not been the certification program).

The interpretation of the canopy monitoring data matters. The explanations of observers Y and Z imply that the program created no additional environmental benefits because it did not change anyone's behavior. However, Observer Y's explanation (self-selection by farmers not planning to cut down their canopy) implies that the program could have been better designed to increase its effect (e.g., better targeting), whereas Observer Z's explanation (coffee prices dropped) implies little could have be done by project designers to create additional environmental benefits.

How do we adjudicate between these rival explanations? To eliminate the explanation of Observer Z, we could measure canopy before and after the project on non-participant farms

whose behavior would have been also affected by coffee prices. If non-participant canopy changed by as much as participant canopy, the data are consistent with a price-based explanation rather than a project-based one. If there were, however, a difference in the changes in canopy, we still cannot eliminate Observer Y's explanation. Why? Every year, many farmers with canopy choose not to cut their canopy because of personal or financial reasons. Because the costs of participating in a certification program are low for such farmers, they are much more likely to participate in the program and thus may comprise much of the participant group. A farmer's intention, or more precisely what he or she would choose to do in the absence of the project, is unobservable. Eliminating Observer Y's explanation may therefore seem impossible. However, with additional effort in project design, attributing a portion of the change in a certification project's monitored indicators to the project rather to rival explanations is indeed possible. The next section explains how experimental project designs provide this potential, as well as help do what none of the three observers do: estimate how much additional canopy is generated by the project.

Box 1. Measuring Success: Monitoring versus Impact Evaluation

Monitoring, which all GEF projects are designed to do, differs from **impact evaluation**, which few GEF projects are designed to do. **Monitoring** is the process of measuring the trends and status of project performance indicators. **Impact evaluation** is the process of attributing changes in the trends and status of project performance indicators to the project actions separate from other factors. For credible impact evaluations, data are collected to identify counterfactual trends and status of project performance indicators. Thus, contrary to text that is sometimes written in GEF proposals, one does not "monitor the impact of the project." The goal of monitoring is to determine whether project indicators are moving in the desired direction or have achieved a pre-defined target. A project can meet its targets, however, without having any causal effect. The goal of impact evaluation is to eliminate rival explanations for the observed data in order to be confident about attributing effects to the project. A project can fail to meet its targets but have a large effect on the indicators of success.

The difficulties in attributing changes in indicators to project interventions and in quantifying project effects are not unique to certification programs. All environmental policies and programs have the same difficulties. These difficulties have been identified as weakening the evidence base in a variety of environmental areas, including energy, pollution, ecosystem conservation, fisheries, and land and forest management (Pullin and Knight, 2001; Sutherland et al., 2004; Saterson et al., 2004; Bennear and Coglianese, 2005; Stem et al., 2005; Frondel & Schmidt, 2005; Ferraro and Pattanayak, 2006; Smith et al. 2006; Greenstone and Gayer, 2009; Ferraro, 2009; Herring and Sorrell 2009; Blackman et al. 2010; Pattanayak et al. 2010; Pullin et al. 2010).

Box 2. Criteria for Causality

1. Temporal Precedence. *The cause precedes the effect.*

2. Covariation of the Cause and Effect. *If the treatment is present, there is an effect, and if the treatment is not present, there is no effect.*

3. No Plausible Alternative Explanation. *One can eliminate rival explanations to a specified degree of certainty (never 100%).*

2. What is experimental project design?

Impact evaluations assess the degree to which changes in outcomes can be attributed to a program, policy, or intervention rather than to confounding factors that also affect the outcomes (see Box 1). In other words, "Did the intervention cause the outcome?" (see Box 2) We will use the word "treatment" to describe any potential project component about whose effects the GEF wishes to learn more. Treatments refer to a broad range of potential GEF-funded actions, from minor project components (e.g., different ways to transmit information to an industrial sector) to the entire proposed project (e.g., a project designed to reduce carbon emissions from manufacturing firms). We will use the phrase "treatment units" to refer to actors affected by the treatment. They may be individuals, communities, firms, sectors, geographic areas, species or any other construct whose exposure to a GEF project could be controlled by the project implementer.

Many projects focus only on the treatment units. By only focusing on outcomes among the treated units, a crucial question is neglected: What would have happened to the treated units' outcomes if they were not exposed to the GEF project?" Impact evaluations answer the question, "Does the treatment work better than no treatment or better than an alternative treatment?" An answer requires knowing what outcomes would have looked like in the absence of the treatment. This counterfactual world, however, can be inferred only indirectly: one cannot, for example, observe the same farmer both participating and not participating in a GEF project.

One useful way of inferring counterfactual outcomes of the treatment units in the absence of the project is to construct a control group comprised of units not exposed to the GEF project. The outcomes of the control group stand in for the unobservable counterfactual outcomes of the treatment group. To make a credible comparison, however, the treatment group and the control group should be, on average, nearly identical in terms of their expected outcomes in the absence of the GEF project. Such similarity is more likely when the treatment and control units have similar characteristics that affect the outcome and inhabit similar economic and physical environments.

Unfortunately, many GEF project implementers screen potential project areas or participants and base their choice on pre-defined characteristics, such as the willingness to participate, which may

affect outcomes even in the absence of the GEF project. The eco-certification example from Section 1 had this potential bias: farmers who are not planning on clearing their forest canopy are much more likely to participate in a shade-grown coffee project. Thus they would have more forest canopy than non-participants even in the absence of the eco-certification program. Potential bias from administrative selection or participant self-selection is rife in environmental programs, as it is in most social policy programs.

Box 3: Creating Experimental Variation in Project Designs

1. *Simple Randomization*. Randomize eligible candidates into one or more treatment groups and a control group.

2. *Randomization in Oversubscribed Projects*. When the number of eligible and interested units exceeds project capacity, the project can select participants through a lottery among eligible and interested candidates.

3. *Randomized Phase-in Projects*. When a project will be phased in over time, the project can select which candidates will enter the program first through a lottery among interested and eligible candidates.

4. *Randomized Encouragement*. Rather than randomize candidates into treatment and control groups, the project randomizes encouragement of candidates into the treatment group.

5. *Discontinuous Eligibility Criterion*. Rather than randomize candidates into treatment and control groups, the project selects an eligibility criterion, such as a cut-off score, which generates treatment and control groups around the cut-off score that may not differ substantially in characteristics that affect the measured outcomes.

To reduce the potential for creating an incomparable control group, one should generate some variation in the areas, firms or people that are exposed to a project and ensure this variation is not related to potential outcomes in the absence of the project. In the certification example, measuring the effect of the project on forest canopy would be easier if the project implementers could ensure that the probability that a farmer participates in the certification program is unrelated to whether the farmer was going to cut or keep the canopy on the farm during the life of the project.

The most straightforward way to create this source of variation is through randomizing which eligible farms are allowed to participate, but there are other ways (see Section 3 and Box 3). **The key to experimental project design is to take advantage of the fact that in many cases, the project implementer has some control over the temporal and spatial assignment of the treatment**. That control can be exploited to help disentangle the effect of a project treatment on

the desired outcomes from all the other factors that also affect those outcomes. Box 4 describes some of the other key features of an experimental project design.

The case for experimental project design is best expressed by Imbens (2010) where he notes that "in a situation where one has control over the assignment mechanism, there is little to gain, and much to lose, by giving up this control through allowing individuals [or areas] to choose their own treatment regime…..I do not want to say that, in practice, randomized experiments are generally perfect or that their implementation cannot be improved, but I do want to make the claim that giving up control over the assignment process is unlikely to improve matters." Imbens then goes on to point out that opponents of experimental designs have not made a case where a non-experimental project design would have improved on an experimental design, conditional on the question lending itself to an experiment design.

Box 4. Key Components of Experimental Project Designs

1. A theory that specifies, as elaborately as possible, a causal relationship between some aspect(s) of the project (*treatment*) and some measurable outcome(s).
2. One or more clearly defined treatments.
3. One or more treatment groups and control groups.
4. Experimental control over the assignment of treatment (see Box 3 and Section 3 for examples).

Note that experimental project designs do not obviate the need for theory (see Box 4). Good experiments use theory and data to go beyond a simple contrast of a treatment and control group outcome. For example, the most powerful experimental project designs can show effects where they should be expected as well as no effects where they are not expected.

Furthermore, it is worth emphasizing that a counterfactual outcome in an experimental project design does not have to be only the outcome in the absence of the project. It could also be the outcome under a different project intervention. For example, a biodiversity project may define the treatment as efforts to strengthen local community institutions in the management of marine protected areas and the control condition as efforts to strengthen local government institutions (a second treatment could be to combine the two approaches). The downside of such a design is that if there were no difference between the treated and control group outcomes, one could not be sure if each approach were equally effective or equally ineffective. Nevertheless, the design can offer important insights about relative effectiveness of two important strategies.

Finally, one should remember that the experimental design in a project can be quite modest. Instead of experimenting with the main project intervention, one can experiment with features of the intervention. For example, a payment for environmental services project may determine that it cannot randomize the payment itself, but it can randomize how the payment is delivered: some communities get household-level payments, some get community-level payments and some get a mix of both. A pollution-prevention program may not want to randomize information provision, but it might want to randomly assign complementary incentive payments to some firms but not others (or to some sectors, but not others). Or the same program may test whether regulatory

compliance is better with random government audits compared to third-party audits. A community forest management program may have already identified the communities with which it will be working, but it might be able to ensure that some communities receive technical assistance and financial incentives, while the other communities receive only technical assistance.

3. How can GEF projects be designed as experiments?

Box 3 summarizes the main experimental project designs that are described in more detail in this section. Boxes 5 and 6 illustrate examples of some of these designs.

Box 5. Experimental Design in the GEF

The United Nations Environment Programme has a medium-size project entitled, "Developing an Experimental Methodology for Testing the Effectiveness of Payments for Ecosystem Services to Enhance Conservation in Productive Landscapes in Uganda" (GEFSEC Project ID: 3682). The project is designed explicitly to test the effect of conservation performance payments on deforestation and poverty. The project will randomly select treatment and comparison communities by:

(a) identifying areas at risk of deforestation;

(b) collecting baseline information on deforestation levels, forest use, and local institutions governing forest management; and

(c) randomizing villages into treatment and comparison groups and initiating the payment scheme.

In the group of forty treatment villages, the option of payment will be offered to individual landholders in return for contractually agreed activities such as maintaining forest cover or actively patrolling forest areas or other activities such as planting of indigenous tree species.

Forty comparison villages will neither be offered payment nor be expected to undertake conservation actions. The differences in environmental and social outcomes between treated and control villages provide estimates of the environmental and social effects of the payment scheme. Data on how sub-groups (e.g., poor villages) respond yield further insights that can be applied to other areas of Uganda and Africa. The Government of Uganda plans to use the evidence generated by the project to develop a replication strategy in other areas at risk of deforestation and to attract ecosystem service buyers. For example, the project evidence may help to position Uganda as a credible supplier of carbon credits in a future international scheme for avoided deforestation and reduced forest degradation.

3.1 Simple randomization

In the simplest randomized experimental project design, candidate units are assigned to treatment and control groups on the basis of a chance mechanism, like a random number generator, and their outcomes are compared. Only chance determines who among the candidate units receives the treatment. Because only chance determines which candidate units are assigned to treatment

and control groups, each experimental group has the same expected outcomes in the absence of the project (they also have the same expected values of all characteristics, observable or not). Thus the control group outcome is a valid counterfactual outcome for the treatment group. Randomization of a given sample may produce experimental groups that differ by chance (particularly when the number of units in each group is small), but common statistical tests and confidence intervals were developed precisely to quantify the potential effects of chance and to distinguish them from a true project effect.

Randomization can take place at the level of the actor whose behavior the project is attempting to influence (e.g., area, individual, household, firm, government agency) or it could be at a higher-order aggregation of these actors (e.g., larger areas, villages, administrative units, industrial sectors). See Box 5. When randomization is conducted at higher-order levels, more care must be taken in the design and analysis.

Box 6: Experimental Design in the GEF: a hypothetical example

A GEF project includes a pilot component to offer firms subsidies to adopt energy-efficient technology in the manufacturing sector. In order to be able to design the 'optimal' program, the project implementers do not just want to know if an incentive leads to reduced energy consumption. They also want to know the incentive amount that is most cost-effective. They thus randomize the size of the incentives offered to firms, stratifying across a few firm characteristics that they believe are the most important determinants of energy-efficient technology adoption. By looking at how responses vary with firm-level characteristics, the project implementers will also understand better how to target "smart incentives" at particular kinds of firms, or how to avoid the trouble of marketing incentives to firm types that, on average, will be unresponsive to the incentive.

3.2 Randomization in Oversubscribed Projects

A common opportunity for introducing randomization occurs when project resources are limited, and thus demand for a program or service exceeds supply. In this case, a natural and fair way to ration resources is to select treated units by lottery among eligible candidates. For example, an energy efficiency project may include a component that includes vouchers for citizens to replace energy inefficient technologies with more energy efficient technologies (e.g., lighting). Demand for these vouchers may exceed supply and thus one fair way to allocate the vouchers would be through lotteries, thereby ensuring that the reason why one interested citizen receives a voucher and another does not has nothing to do with their expected future energy use (real-world examples of such designs have been done with school and housing vouchers).

3.3 Randomized Phase-in Projects

Financial and administrative constraints often lead NGOs to phase-in programs over time. Randomization will often be the fairest way of determining the order of phase-in. Thus candidates who enter the program later can be used to estimate the counterfactual outcomes for candidates who enter the program earlier. All candidates will eventually receive the program and

thus no one is denied project benefits. For this design to be successful, one would have to measure outcomes that could reasonably be expected to be affected by the program prior to all the control candidates entering the program. If a randomized phase-in is too rapid relative to the time it takes for project indicators to change, it will be impossible to detect effects.

3.4 Quasi-experimental randomization of encouragement

When a program must remain open to all eligible participants or areas, but participation in the program will not be universal, project implementers may be able to use an encouragement design to create experimental control. An encouragement design does not randomize the treatment, but rather randomizes efforts to encourage participation. Thus it is often considered a quasi-experimental design. For example, in Kenya, Duflo, Kremer, and Robinson (2006) evaluated the effect on future fertilizer adoption by farmers after witnessing fertilizer demonstration on another farmer's plot. They set up fertilizer demonstrations on a random sample of farmers' plots and then invited a randomly selected subset of the farmers' friends to view the demonstration. While a farmer's other friends were also welcome to come, the fraction who attended was much larger among those invited than those not invited. Since the invitation was randomly assigned, it provides experimental variation in who is exposed to the project treatment (demonstration farms) that is unrelated to the measured outcomes (fertilizer adoption). This source of variation is called an "instrumental variable" in the statistical literature.

Analyzing and interpreting results from an encouragement design is a bit more complicated than a simple randomized experiment because the experimental variation only increases the probability that a treatment is received, rather making the probability either zero or one. Nevertheless, the design can provide the requisite experimental control to help uncover policy-relevant causal relationships between GEF project components and outcomes.

3.5 Quasi-experimental Discontinuity Designs

Many projects create eligibility rules that determine in which areas the project will operate or which firms, households or individuals can participate. When the eligibility rule is in the form of a sharp cutoff value, such as a poverty score or number of employees, the discontinuity in eligibility formed at the cutoff value creates a potential source of variation that is unrelated to the measured outcomes. Although, overall, the treated units are unlikely to be directly comparable to the control units, the treated and control units that are in the immediate neighborhood of the discontinuity (i.e., just on either side of the cutoff) are likely to be very comparable. Therefore, within this neighborhood, treatment is assigned 'as if random.' In the scientific literature, this design is called a "discontinuity design."

Consider a hypothetical example of the Amazon Region Protected Areas Project in Brazil. The country provides a much larger number of potential sites for protection than GEF and other donors will have the ability to fund. Conservation personnel often use a quantitative scoring metric to compare each potential site in planning exercises. If the project can define an eligibility cutoff based on this scoring metric, enforce the cutoff rigorously (i.e., no sites with scores below the cutoff receive funding, or they receive less funding), and conduct a baseline survey of units

close to the cutoff, the differences in forest cover and other outcome indicators on either side of cutoff give an estimate of the project effect for those units. If the cutoff were not enforced rigorously, one could still estimate the effect of the program, but the methods would be a bit more complicated (a so-called "fuzzy discontinuity design").

3.6 Non-experimental Evaluation Designs

In many cases, experimental project designs are not feasible or they may yield less credible results than a non-experimental design. In these cases, a wide range of non-experimental project designs are available for GEF projects. This advisory document offers guidance on experimental project design and some forms of quasi-experimental project design in which the project designers can influence the probability of treatment assignment (exposure to the project). We direct the reader to other sources of references for non-experimental designs (Frondel and Schmidt, 2005; Ferraro and Pattanayak, 2006; Ferraro, 2009). The goal of these non-experimental evaluation designs is the same as the goal of experimental designs: to anticipate potential rival explanations of changes in the measured indicators and to collect data before and after a project begins with the intention of eliminating these rival explanations.

4. Issues in Experimental Project Design

This section summarizes some, but not all, of the important issues that are often raised when considering an experimental project design (for more in-depth treatments see Shadish et al., 2002; Duflo et al., 2008). Some common concerns about such designs are listed in Box 7. An important concern to highlight is the ethics of experimentation. Opponents of experimental project designs see ethical implications when a project is described as an experiment, but see no such implications when a non-experimental project encourages individuals, firms or species to participate in an unproven initiative. In other words, the people and the environment in GEF projects are already acting as experimental subjects. They are simply participating in poorly designed experiments. In the author's opinion, the real concern with the current trend towards credible causal inference in environmental and development policy, and towards experimental project designs in particular, is that it may lead project designers to avoid projects where experimental control is difficult, or even conceptually impossible. But the GEF, with almost no experimental project designs in its portfolio, is no danger at this point of going down this road.

4.1 Validity

As with all evaluation designs, one must consider not only their internal validity (i.e., whether one is actually estimating a causal relationship rather than hidden biases), but also their construct validity (whether one is actually measuring the outcome and treatment one reports to be measuring) and external validity (whether the results would be the same for other people, places, or times). These issues, however, are largely context-specific rather than design-specific.

However, it is worthwhile to remember that inferences drawn from experimental project designs are particular to certain units at certain times under certain circumstances. Although

experimental designs and methods can control and measure uncertainty associated with measuring the effect of a treatment in a given sample (internal validity), they cannot account for the uncertainty associated with generalizing an effect estimate beyond the experimental sample (external validity). To increase the external validity of an experimental project design, one would want to ensure, as much as feasible, that the sample of units in the experiment are representative of the relevant populations (random sampling of candidate units can thus help with increasing external validity). One could also measure observable characteristics of the experimental population, with which others could determine how applicable the results of the project are for other areas. In areas where reliable evidence is scarce, knowing with a high degree of confidence the effects of an intervention on a population similar to the population of interest may be extremely valuable, even if the two populations are not identical in many respects. Moreover, an experiment that demonstrates a new intervention is cost-effective may generate substantial policy

4.2 Heterogeneous effects and mechanisms

Box 7. Five Common Concerns about Experimental Project Designs

Concern #1: *Experimental designs are unethical because they deny project benefits to some groups*. Approaches to ethical experiments with human subjects are well established. Furthermore, (a) few GEF projects have universal coverage and thus some candidate units are denied access in most GEF projects; and (b) when a project intervention is based on little or no empirical evidence, and theory suggests it may be ineffective (or worse), exposing individuals, firms, ecosystems and species to a non-experimental version of the project also has ethical implications.

Concern #2: *The experimenter must have control over all variables that affect the outcome, as in laboratory experiments*. Such control is neither necessary nor possible. Moreover, experimental designs do not require that experimental units be homogenous or be a random sample from a population of units (see "internal validity" in Section 4).

Concern #3: *A project needs perfect compliance with the experimental protocol*. Although compliance makes drawing inferences about effects easier, there is a substantial literature on methods for drawing inferences in the presence of noncompliance (e.g., units refuse treatment).

Concern #4: *Experimental project designs are research-oriented and impede action*. They do not impede action, but rather shape action to answer valuable project implementation questions. They are a form of action research and implementation science.

Concern #5: *Experimental project designs are expensive*. There are two sources of additional costs: (1) design expertise; and (2) data collection on control units. Whether failing to collect data on non-participants in a non-experimental design is a cost-savings is debatable. With a design that makes inferring project effects easier, one could save money by reducing the need to monitor a large set of indicators. Thus whether total costs of experimental designs are substantially larger is context-specific.

In many contexts, we do not simply want to know "does it work" but rather we also wish to know "for whom does it work and how does it work?" In other words, we would like to understand the heterogeneity of program effects, so we can better understand distributional issues and potential improvements in program targeting. And we would like to understand the

mechanisms through which a project works, so that we can better understand potential improvements in the project design for future efforts. Experimental project designs are sometimes accused of being "black boxes," which tell you whether an outcome indicator changed as a result of project actions, but not much else. But there is nothing intrinsic about an experimental project design that prevents project designers from considering questions about heterogeneity and mechanisms. One can collect data on units in the treatment and control units and analyze effects conditional on subgroups (e.g., poor versus rich; large firms versus small firms; steep slope versus flat slope). With an elaborate causal model that identifies potential mechanisms, one can collect data on the mechanisms and combine the experimental data and non-experimental statistical methods to examine the causal paths between the treatment and the outcomes (e.g., Morris and Gennetian 2003). Or one can create an experimental design to explicitly estimate a mechanism's effects (e.g., Ludwig et al. 2011). As more experimental project designs are implemented, by the GEF and others, methods of research synthesis (meta-analysis or multilevel models that pool primary data) can be used to estimate how intervention effects vary with project implementation, sample characteristics, and local context.

4.3 Potential Biases in Experimental Project Designs

Related to issues of validity (4.1) are the many things that can go wrong in an experimental project design, thus reducing the credibility of the evidence it generates. See Shadish et al. (2002) for an in-depth treatment of the potential biases that can creep into even the best designed project experiments. For example, an experiment may not be a good estimate of the effects of a treatment if people behave differently when they know they are being observed, if they believe the experimental intervention is temporary but a scaled-up version would be permanent, or if the scaled-up version would create spillovers from participants to non-participants that were not present in the experiment. In well designed experiments, there are often departures from experimental control (e.g., treated units refuse treatment; control units are exposed to treatment) and attrition of units (e.g., firms go out of business; villagers move away). However, there are methods for addressing some of these biases and in many cases, their presence bounds the effect estimates from an experiment rather than invalidates them (e.g., one can say that the estimate is a maximum because potential biases inflate the effect estimate). Moreover, when a potential problem, such as units refusing treatment, would be part of any non-experimental project, it is not a bias. For example, if people refuse to participate in both experimental and non-experimental versions of a project, then we may want to measure the effect on people invited to participate (called the *intent to treat treatment effect*), not just the effect on participants. Likewise, for many environmental outcomes, monitoring (watching people) is an integral part of any project and thus any effects of monitoring in the experiment would also be experienced in the scaled-up version and should be part of the estimate of the project effect.

Likewise, the most likely important source of bias for GEF projects is not peculiar to experimental design: spillovers from treated units to control units. An implicit assumption in simple analyses of experimental data is that the outcome of one unit should be unaffected by the assignment of treatments to the other units. But in some GEF projects, this assumption is likely to be false. For example, in an incentive-based project, control units may change their current behavior in anticipation that such a change will increase their likelihood of receiving an incentive

in the future. Or a project aiming to reduce deforestation in one community may simply increase it in a neighboring control community. To reduce the potential for spillovers to bias inference in an experimental project design, one can either (a) try to control for them (e.g., select units that are spread out so that, for example, control units are too far away from the project to have heard about it) or (b) try to measure them (e.g., through a randomized saturation or a randomized distance design). Controlling for them would be warranted if spillovers are a function of the pilot nature of the project rather than the project itself (e.g., a scaled-up project would be available to everyone and thus spillovers would no longer be relevant). Measuring them would be appropriate if the spillovers are an important potential component of the program effect at any scale.

4.4 *Long-term versus short-term effects*

One criticism of the use of experimental project designs in the environmental arena is the slow pace at which effects often materialize. Why would one bother to go through the trouble of using an experimental project design if, at the end of the project, one could not hope to detect an effect? This is a reasonable criticism, but it applies to monitoring and evaluating efforts in any project. One response is that often there are intermediate variables that can be measured during the project lifetime. Under the assumption that the causal model that connects actions to final outcomes is correct, changes in these intermediate outcomes will be correlated with the final outcomes. For example, a program designed to reduce chemical emissions through the diffusion of production practice changes may take many years to have an effect on emissions. One might, however, be able to measure changes in production practices or production investments during the project lifetime. If there is no project effect on these intermediate variables, one would be skeptical that emissions will eventually change. If there is a project effect on these intermediate variables, one is left with either the conclusion that emissions will eventually change or the causal model proposed during project design is incorrect. For some projects, such evidence may be better than nothing. A second response is that if the experiment is well designed, publicized and documented, someone, if not the GEF, will surely return to the area and do follow-up research. Such follow-up studies are routine in other policy fields like public health and education, whose populations are often much more mobile than those in GEF projects.

4 Conclusion

If the goal of GEF projects is to affect global environmental outcomes and to catalyze investments from other donors and host-country governments, experimental project designs can contribute to realizing this goal. The evidence generated by such designs can be credible and transparent, two attributes that are necessary for evidence to spur action. Experimental project designs, such as conditional cash transfer programs (e.g., *Progresa*) and antiretroviral drug distribution programs (e.g., *Bangkok Collaborative Perinatal HIV*), have had huge policy effects on governments and donor agencies. Such designs could have similarly large effects in environmental policy. They would be less focused on testing whether a specific project "worked" and more focused on providing insights about the validity of the implicit and explicit causal models that underlie the global environmental investment portfolio.

Like any organization dependent on external funding, the GEF wants to be able to tell a good story about what it does and why it matters. Experimental project designs offer simple, credible stories. They therefore not only generate knowledge with which the GEF can improve its operations; they also can catalyze investments directly and indirectly into GEF initiatives.

Experimental project designs are consistent with the GEF mandate and are applicable to some of the many projects funded by the GEF and its partners each year. Experimental project designs could be financed through full-size and medium-sized projects, and through global and regional exclusion funds. Indeed, the Biodiversity Focal Area Set-Aside Programming Strategy includes a window for incentive funding for projects that are "contributing to global conservation knowledge through formal experimental or quasi-experimental designs that test and evaluate the hypotheses embedded in project interventions." Focal area learning objectives can be used to guide the questions that are posed in experimental project designs and ensure that experimental project designs are part of an overall portfolio of activity oriented toward learning and knowledge management. Even better would be to build into the GEF portfolio incentives that encourage experimental project designs (e.g., create a Programmatic Window for Targeted Research on Implementation Science and Experimental Project Designs).

Experimental project designs are most needed in the presence of three conditions, common to many GEF projects:

(1) the empirical evidence base supporting a project intervention is weak;

(2) the measured outcomes are spatially and temporally variable <u>in the absence of the project</u> (e.g., at any point in time or space, there is variation in the number of trees cut, pounds of pollutants emitted, investments made in soil or water management, etc.); and

(3) selection into the project treatments is systematically related to characteristics that also affect the expected outcome.

Condition one implies credible evidence is critically needed, and conditions two and three imply that simple predictions based on engineering models, before-after comparisons, or participant-nonparticipant comparisons are likely to yield substantially biased effect estimates.

This advisory document does not advocate that every GEF project use an experimental project design. It merely advocates that *some* of the hundreds of GEF projects in a replenishment cycle use such a design. While the analysis of data from experimental projects can be straightforward, their design requires some expertise. All of the hard work falls at the project design stage rather than during or at the end of the project, as is typical in common evaluation designs. One must understand to what interventions one can successfully apply experimental control in a given setting, and then build a project design around this understanding. The text in Box 8 describes conditions under which experimental project designs will be most feasible at the GEF.[1] Not every condition has to be present for an experimental design to be worthwhile.

[1] Projects would benefit if they also have the ability to conduct a baseline survey on pre-intervention values of the outcome indicators and a few key covariates that affect the outcome. Although such data are not critical to

Experimental project designs will be most useful if they test fundamental behavioral questions (e.g., how does an energy-efficient technology lending program change energy use? how do land-users respond to financial incentives? how do local government decision makers respond to information or capacity building?) or test popular classes of policy approaches (e.g., devolution of resource management authority to communities).

Programs targeted to individuals, firms, local communities or municipalities are likely to be strong candidates for experimental project designs. For example, (1) incentive and lending programs (e.g., for energy efficiency technology adoption, sustainable land use practices, sound chemical management and biodiversity protection); (2) certification programs (e.g., energy

accomplish the goals outlined in the text, they would increase the statistical precision of impact estimates, and would allow one to check how well the randomization worked (and control for small biases that may stem from randomization that fails to balance treatment and control groups on factors that affect the outcomes).

Box 8. Favorable Experimental GEF Project Design Conditions

1. The intervention is popular, or increasingly popular, or the required behavioral change is common to many projects. Thus the results of an experimental project design would have broad application to the GEF portfolio and beyond.

2. The intervention will be conducted on a sufficient number of units (typically more than thirty), which permits sufficient statistical power to detect a policy-relevant effect should one exist. Thus experimental units like individuals, households, villages, firms or areas will be easier to incorporate into an experimental project design than regions, nations or entire ecosystems.

3. Factors that affect access to the intervention are well understood. Thus project designers understand how to inject experimental variation into the project implementation.

4. Access to the intervention is well controlled by the project implementers and there are few similar interventions in study area (unless these similar interventions are the counterfactual sought by the designers). Thus project designers can clearly separate units into treatment and control groups.

5. Final outcomes, or important intermediate outcomes, can be observed by the end of the project. Experimental designs can indeed be used to measure post-project effects, but uncertainty over funding for future monitoring costs make conditions less favorable.

6. Outcomes across units are relatively independent, or measuring or controlling potential spillover effects can be built into the design (e.g., using larger units, such as villages, or randomized saturation designs, in which the proportion of units exposed to the program is varied randomly across space).

7. Opportunities exist to include real policy alternatives within the experimental project design so that the results can inform best practice, rather than just answer the question, "Does it work?" The best project designs can answer, "In which circumstances do interventions work best?" and "Which policy levers maximize effects?"

efficiency, eco-friendly production practices); (3) information provision to farmers, firms or local governments (e.g., energy audits, ecosystem service valuation, sustainable agriculture, pollutant release and transfer registries); (4) community natural resources management; and (5) policing and compliance auditing strategies. Particularly appropriate would be pilot programs or programs implemented by nongovernmental organization partners, which are not expected to serve everyone and may have more flexibility with regard to where and with whom they operate. Project components that focus on national-level regulatory change or national-level capacity building are generally not appropriate for experimental project designs, nor are projects that attempt to effect broad, but diffuse, change.

In summary, experimental project designs are part of a comprehensive strategy for the GEF to become a leader in the production of environmental policy evidence that spurs innovation, investment and impacts across the globe. The GEF is ideally placed to be such a leader. It is well respected and invests in popular classes of interventions in multiple nations, thus affording

opportunities for replicating experimental project designs at different sites and with design variations. Experimental social policy, which uses experimental project designs to complement other sources of evidence, is becoming standard in other social policy fields and will eventually become standard in environmental policy. The GEF can be a leader at the front of this movement, or it can wait and struggle to catch up ten years from now when experimental designs as a component of environmental policy portfolios are the norm.

References

Bennear, L. S., & Coglianese, C. (2005). Measuring progress: Program evaluation of environmental policies. *Environment, 47*(2), 22–39.

Blackman, A and J Rivera. 2010. Environmental Certification and the Global Environmental Facility, A STAP Advisory Document. Washington, DC.

Bowler, D, L Buyung-Ali, JR Healey, JPG Jones, T Knight and AS Pullin 2010. Community Forest Management as a Mechanism for Supplying Global Environmental Benefits and Improving Local Welfare, A STAP Advisory Document. Washington, DC.

Duflo, E, R Glennerster, and M Kremer. 2008. Using Randomization in Development Economics Research: A toolkit. *Handbook of Development Economics* 4: 3895-3962.

Duflo, E, M Kremer, and J Robinson. 2006. Understanding Technology Adoption: Fertilizer in Western Kenya: Evidence from Field Experiments. Working Paper.

Ferraro, PJ. 2009. Counterfactual Thinking and Impact Evaluation in Environmental Policy. In Special Issue on Environmental Program and Policy Evaluation, M. Birnbaum & P. Mickwitz (Eds.). *New Directions for Evaluation* 122: 75–84.

Ferraro, PJ and SK Pattanayak. 2006. Money for Nothing? A call for empirical evaluation of biodiversity conservation investments. *PLoS Biology* 4: 482-488.

Frondel, M, and CM Schmidt. 2005. Evaluating environmental programs: The perspective of modern evaluation research. *Ecological Economics* 55: 515–526.

Greenstone, M, and T Gayer. 2009. Quasi-experimental and experimental approaches to environmental economics. *Journal of Environmental Economics and Management* 57: 21-44

Herring, H. and S. Sorrell, eds. 2008. *Energy Efficiency and Sustainable Consumption: Dealing with the Rebound Effect*. Palgrave Macmillan, Basingstoke.

Imbens, GW. 2010. "Better LATE Than Nothing: Some Comments on Deaton and Heckman and Urzua. *Journal of Economic Literature*, 48(2): 399–423.

Ludwig, J, JR Kling, and S Mullainathan. 2011. Mechanism Experiments and Policy Evaluations. *Journal of Economic Perspectives* 25: 17–38

Morris, P, and L Gennetian. 2003. Identifying the Effects of Income on Children's Development: Using Experimental Data. *Journal of Marriage and the Family* 65: 716-29.

Pullin, AS, and TM Knight. 2001. Effectiveness in conservation practice: Pointers from medicine and public health. *Conservation Biology, 15*(1), 50–54.

Saterson, K. A., Christensen, N. L., Jackson, R. B., Kramer, R. A., Pimm, S. L., Smith, M. D., and Wiener, JB. 2004. Disconnects in Evaluating the Relative Effectiveness of Conservation Strategies. *Conservation Biology, 18*, 597–599.

Shadish, WR, TD Cook and DT Campbell. 2002. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA, US: Houghton, Mifflin and Company.

Stem, C., Margoluis, R., Salfasky, N., & Brown, M. (2005). Monitoring and evaluation in conservation: A review of trends and approaches. *Conservation Biology, 19*(2), 295–309

Sutherland, W. J., Pullin, A. S., Dolman, P. M., & Knight, T. M. (2004). The need for evidence-based conservation. *Trends in Ecology and Evolution*, *19*(6), 305–308.

For Further Reading

Baker, JL. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Washington, DC: The World Bank.

Duflo, E. R Glennerster, and M Kremer. 2008. Using Randomization in Development Economics Research: A toolkit. *Handbook of Development Economics* (4): 3895-3962.

Orr, L. 1999. *Social experiments: evaluating public programs with experimental methods*. Sage Publications: Thousand Oaks, CA, USA.
Ravallion, M. 2008. Evaluating Anti-Poverty Programs. *Handbook of Development Economics* (4): 3787-3840.

Annex 1

Glossary
All glossary terms are defined as the words are used in this guideline document.

Baseline: measures of indicators of project *outcomes* or of characteristics of the *treatment* and *control* units prior to project implementation.

Causal Model (also called theory of change): A precise and elaborate causal theory of the way in which the experimental *treatment* leads to *effects* (a causal pathway). A causal model requires one to articulate assumptions that can be tested and measured, and select indicators of *outcomes* (final or intermediate).

Comparison group: See *control group*.

Control group: The group of people, communities, ecosystems or firms that do not participate, or are not exposed to, the project component of interest. Also called a *comparison group*. For example, firms that do not receive GEF-funded assistance in a renewable energy project or households that do not receive payments to protect biodiversity are potential control groups. Contrast with *treatment group*.

Confounding factors: Factors that mimic or mask a *treatment*'s effect. These factors may include characteristics of the *units* exposed to the program (e.g., a firm's intention to pollute, a farmer's intention to cut down forest) or of the environment in which a program operates (e.g., input prices, weather).

Counterfactual Outcome: A *unit*'s *outcome* in absence of *treatment*. "Absence of the treatment" does not imply absence of any intervention. It implies the next-best alternative. This alternative could be the status-quo interventions or it could be another treatment.

Effect (causal effect): The difference between the observed *outcome* for a treated unit and the *counterfactual outcome* for the same unit. Typically we can only infer average effects on a group of individuals. *Effect* rather than *impact* is used in this document to avoid confusion with the way in which "impact" is used by the GEF-Evaluation Office in its Monitoring and Evaluation Policy to imply long-term, or ultimate, causal effects (http://www.gefieo.org/evaluations/gef-monitoring-and-evaluation-me-policy-2010).

Experimental project design: A project design in which *project implementers* directly manipulate variables to test cause-and-effect relationships (e.g., manipulate the eligibility criteria for project participation). In an experiment, the assignment of *treatment* and *control* groups is a process controlled by the project implementers and, importantly, makes it unlikely that the expected outcomes in treatment and control groups in the absence of the project would be different. *Experimental project design* is not synonymous with *pilot project*, although pilot projects can often be implemented with experimental designs.

External validity: Whether the *effects* would be the same for other people, places, or times.

Indicator: A measured variable that proxies for the outcome of interest. For example, one might select change in forest cover as an indicator of changes in biodiversity, or changes in energy consumption as an indicator of changes in greenhouse gas emissions.

Internal validity: Whether the estimated *effects* indeed describe a causal relationship.

Outcome: The variable that a *treatment* is hypothesized to affect. These may be final outcomes (e.g., emissions) or intermediate outcomes (e.g., energy consumption). Examples include deforestation, species populations, chemical emissions, and soil carbon content.

Project implementer: An agent or group of agents that implements a GEF project and thus controls how the project unfolds in the field. Implementers typically comprise individuals from government agencies, non-governmental organizations, or multi-lateral or bi-lateral organizations.

Quasi-experimental project design: A project design in which assignment to *treatment* and *control* groups is not controlled by the project implementers, but which can, under certain assumptions, can allow one to infer *causal effects* of the project.

Random Sample: A sampling method in which each member of a set has an equal and independent probability of being selected. The purpose of a random sample is to more credibly generalize the inferences drawn from the sample to the population from which the sample comes.

Randomization: Randomly assigning units into *treatment* and *control groups*. The purpose of randomization is to increase the credibility of inferences drawn about cause-and-effect relationships (internal validity). See main text for more explanation.

Treatment: The project component that is hypothesized to have a causal effect on an outcome of interest (i.e., a cause of an effect). In an experimental design, the treatment is the project element that is manipulated by the experimenter. Examples include regulations (e.g., protected area), incentives (e.g., subsidies for energy efficient technology adoption), information transfer (e.g., farmer training on climate adaptation skills), and decentralization (e.g., community forest management), or the delivery mechanisms for these interventions.

Treatment group: The group of people, communities, ecosystems or firms that participate, or are exposed to, the project component of interest (the *treatment*). For example, firms that receive assistance in a renewable energy project or households that receive payments to protect biodiversity are treatment groups. Contrast with *control group*.

Unit: Units may be individuals, households, communities, species, geographic areas, firms or other organizations. They are the actors whose behaviors the GEF projects are attempting to influence