**Annex III.3:** **Guidance Note 1: Attributing Outcomes to the Project**

1. The definition of “impact evaluation” used in the 2012 GPS is “an evaluation that quantifies the net change in outcomes that can be attributed to a specific project or program, usually by the construction of a plausible counterfactual. “[[1]](#footnote-2) Thus, impact evaluation focuses on quantifying the incremental contribution to results that is attributable to the intervention.
2. In theory, a comparison with a counterfactual can be done for any level in the project’s results chain.[[2]](#footnote-3) But the question of attribution becomes trivial at lower levels of the results chain: the counterfactual of “what would have happened in the absence of the project” becomes “before-project”. A good example is the impact of a water supply project on the time household members spend collecting water.[[3]](#footnote-4) The average water collection time falls after the project. The only plausible explanation is the improved proximity or predictability of water. In this case, the counterfactual (what would have been the time spent gathering water, without the project) is that the time would have remained the same as before the project. A before-and-after approach is sufficient to determine the change in outcome attributable to the project. At higher levels in the results chain, however, construction of a counterfactual is more difficult.
3. A variety of quantitative methods can be used in impact evaluations (see below). Even if it is not possible or desirable to conduct an impact evaluation, qualitative methods can be used to construct a plausible counterfactual and make an informed judgment (but not quantify) the additionality of the project to the intended outcomes.
4. Establishing a causal relationship between the project and its outcomes starts with the project’s results chain that links project activities with intended outputs and outcomes. In other words, the starting point is to build up the program theory. This is sometimes referred to as a “theory-based” evaluation framework. This approach maps out the channels through which the activities, inputs, and outputs are expected to result in the expected outcomes. It is a systematic testing of all of the links (assumptions) in the results chain. It also allows for the identification of unintended effects.
5. The Network of Networks on Impact Evaluation (NONIE) provides the following overall guidance for impact evaluation:

*Carefully articulate the theories linking interventions to outcomes. Address the attribution problem. If possible, use quantitative approaches, embedding experimental and quasi-experimental designs in a theory-based evaluation framework. Qualitative techniques should be used to evaluate attribution issues for which quantification is not feasible or practical. Preference is to use mixed-methods designs. Use existing research relevant to the results of the intervention.[[4]](#footnote-5)*

1. The remainder of this Guidance Note discusses various quantitative methods that can be used to attribute results to project activities, what to do when quantitative techniques are not feasible, or practical, and how to construct a counterfactual for policy-based operations.

**Quantitative Methods**

1. The main designs for impact evaluation include the following:[[5]](#footnote-6)

* Randomized assignment: Randomized assignment of treatment essentially uses a lottery to decide who among the equally eligible population receives the project treatment and who does not. Under specific conditions, randomized assignment produces a comparison group that is statistically equivalent to the treatment group.
* Difference-in-differences: Estimates the counterfactual for the change in outcome for the treatment group by calculating the change in outcome for the comparison group. This method takes into account any differences between the treatment and comparison groups that are constant over time.
* Matching Estimators: Uses statistical techniques to construct an artificial comparison group by identifying, for every possible observation under treatment, a non-treatment observation (or set of non-treatment observations) that has the most similar characteristics possible. These “matched” observations then become the comparison group that is used to estimate the counterfactual. A common way to match different units is to model how likely each unit is to be treated, based on observed variables, and then to match treated and untreated units based on this likelihood (or “propensity score”).
* Regression approaches: An alternative to matching, usually done by observing both treated and control units and to “control” for as many pre-program covariates as possible. This method is similar to matching techniques in that it uses observed characteristics of treated and untreated units to try to make them “similar”.
* Instrumental variables: A method used to control for selection bias due to unobservables. Certain variables are chosen that are believed to determine program participation but not outcomes. These instrumental variables are first used to predict program participation; then, the predicted values are used to see how outcome indicators vary with the predicted values.
* Regression Discontinuity: An impact evaluation method that can be used for programs that have a continuous eligibility index with a clearly defined cutoff score to determine who is eligible and who is not. The regression discontinuity measures the difference in post-intervention outcomes between the units just above and just below the eligibility cutoff.
* Modeling the theory: The determinants of outcomes are estimated using regression models. The determinants of these determinants are also modeled, working down the results chain until the link is made to project inputs.

1. Experimental and quasi-experimental methods should be used to construct a comparison group when they are appropriate, feasible, and practical. In many situations, however, they are not possible – for example, when the project is comprehensive in scope (such as economy-wide policy reforms) or works with a small number of entities (such as institutional reforms). Random assignment also may not be possible for political or ethical reasons.[[6]](#footnote-7)
2. Often, baseline data are not available. Possible alternative designs include (i) single difference methods (after-project comparisons of participants and non-participants), if the groups are drawn from the same population and some means is found to address selection bias; and (ii) using another dataset to serve as a baseline.

**Qualitative Methods**

1. If quantitative methods are not feasible, or practical, the evaluation should employ “causal contribution analysis” by building a strong descriptive analysis of the causal chain. The evaluator attempts to provide evidence that assumed links in the chain in fact occurred, or identify breaks in the chain so as to argue that expected results did not occur. Arguments can be strengthened by triangulation, i.e., drawing on a variety of data sources and approaches to confirm that a similar result obtained from each.
2. To analyze the links in the causal chain, the evaluator:

* Assesses the causal chain in relation to the needs of the target population, collaborating with stakeholders and experts.
* Examines the critical assumptions and expectations inherent in the project’s design, reviewing the logic and plausibility of the results chain. Again, this is done in collaboration with stakeholders.
* Uses available research evidence and practical experience elsewhere, comparing the project with projects based on similar concepts.
* Observes the project in operation, focusing on interactions that were expected to produce the intended outcomes.[[7]](#footnote-8)

1. Beneficiary surveys, focus groups, structured interviews, and other instruments are other techniques commonly used to provide qualitative evidence for causal contribution analysis.
2. Case studies are useful as a complementary method. They can describe what the implementation of the project looked like on the ground and why things happened the way they did. Not only are case studies more practical than large national studies, they also provide in-depth information that is often helpful to decision makers.
3. Most evaluation textbooks and guidelines advocate a mixed-method approach, combining quantitative and qualitative methods when possible. This is because some impact evaluation methods give results out of a “black box” – i.e., they can be used to quantify the results of a project but do not necessarily explain why the results occurred. For other reasons, it may be useful to compare the results of before-and-after comparisons with the results of using other methods to determine causality.

**Policy-Based Lending[[8]](#footnote-9)**

1. It is more difficult to assess and attribute the results of PBL operations than investment loans. PBLs support a program of policy and institutional changes, and often operate at the economy-wide level. Assessing PBL outcomes is complicated by the interaction of IFI-supported reforms with contemporaneous changes in other public policies, shocks, cyclical factors, and changes in market conditions. Isolating and attributing change to any particular set of PBL-supported policy and institutional actions is information-intensive and analytically demanding.
2. Quantitative approaches are available to isolate PBL outcomes and to compare performance with a counterfactual scenario. Some PBL evaluations have employed simple growth decomposition methods to isolate the effects of policy change from major shocks and changes in the terms of trade. Others have used cross-country regression models to distinguish the effects of policy change from starting points and structural characteristics of borrowers.
3. A number of qualitative approaches also may be useful for separating the effects of the program supported by the PBL from other factors, and assess the influence of the PBL on program outcomes. These include:

* A review of performance indicators, activity surveys, and structured interviews with key stakeholders can be used to assess whether or not the implementation of PBL-supported measures actually gave rise to the outputs and outcomes expected of them.
* Beneficiary satisfaction surveys can be conducted, with the results of policy and institutional change being “scored” directly by stakeholders.
* In some cases, PBL appraisal reports contain a “without reform” scenario for certain key outcome variables.
* It may also be possible to use the performance of policy and impact variables in similar countries that did not undertake PBL-supported reforms as a baseline comparator.
* Observed outcomes can be benchmarked against regional or international standards of public policy and institutional performance to assess the significance of PBL-supported actions to transforming policy settings.
* Advantage should also be taken of previous evaluations and research, including comparative studies of experiences with structural adjustment. They can suggest factors that have been associated with successful adjustment.
* The insights obtained from other sources of information, including key informant and group interviews and mini-surveys, can shed further light on attribution issues. Individuals intimately involved in a reform process can often identify the counterfactual.

1. In some cases, a qualitative assessment of the linkages between the PBL and the desired outcomes is sufficient to identify what elements were missing, or could have been better designed. With adequate benchmarks and *ex post* performance information, simulations, cost-benefit, cost-effectiveness, and other quantitative techniques can be used to inform such judgments. Evaluations of PBL poverty outcomes draw on a variety of techniques and survey instruments to assess changes in living standards, livelihoods, and benefit incidence.
2. Precisely attributing the contribution of any single PBL is nearly impossible when many stakeholders have a hand in policy change, but evaluators can assess what additional value the PBL had to the policy change process, beyond the provision of financial support. For example, the additionality of a PBL can be evaluated in terms of whether or not it (i) accelerated (or delayed) reform, (ii) strengthened the hand and credibility of reformers, (iii) raised the perceived political returns to reform in terms of easing budget constraints and positive reputation effects, (iv) fostered policy learning, (v) built domestic capacity to design policy, and (vi) spurred debate and dialogue on new approaches to meeting development objectives.

1. Note that this definition of impact evaluation is not the same as “an evaluation that focuses on the final level in the causal chain” (e.g., social and economic outcomes such as poverty reduction, which are sometimes called “impact”). [↑](#footnote-ref-2)
2. See White, Howard (2007). *Evaluating Aid Impact*. World Institute for Development Economics Research, Research Paper No. 2007/75, November. [↑](#footnote-ref-3)
3. Example from White, Howard (2009). *Some Reflections on Current Debates in Impact Evaluation.* International Initiative for Impact Evaluation Working Paper No. 1, April. [↑](#footnote-ref-4)
4. Leeuw, Frans and Jos Vaessen (2009). *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation*. Network of Networks on Impact Evaluation. [↑](#footnote-ref-5)
5. See Paul J. Gertler and others. (2001), *Impact Evaluation in Practice*. World Bank; and Yuri Soares (2011), *Note on the Practice and Use of Impact Evaluation in Development: Reflections for the Evaluation Cooperation Group (ECG) Conference in Manila*; March. [↑](#footnote-ref-6)
6. See Soares (op. cit.) for the limitations of randomized control trials and of impact evaluation in general. [↑](#footnote-ref-7)
7. Morra-Imas, Linda, and Ray C. Rist (2009). *The Road to Results: Designing and Conducting Effective Development Evaluations.* World Bank. [↑](#footnote-ref-8)
8. Tabor, Steven R., and Stephen Curry (2005). *Good Practices for the Evaluation of Policy-Based Lending by Multilateral Development Banks.* Asian Development Bank report prepared for ECG, March. [↑](#footnote-ref-9)