IDB
Development Effectiveness

Impact-Evaluation Guidelines
Technical Notes
No. IDB-TN-157
August 2010

Evaluating the Impact of Regional Development Programs

Paul Winters
Susana Sitja Rubio

Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank

Evaluating the Impact of Regional Development Programs

Impact-Evaluation Guidelines

Paul Winters
Susana Sitja Rubio

Inter-American Development Bank
2010

Paul Winters. Email: winters@american.edu

Susana Sitja Rubio. Email: susanasi@iadb.org

Evaluating the Impact of Regional Development Programs

Abstract

Paul Winters[*] and Susana Sitja Rubio[**]

The purpose of this guideline is to discuss the objectives and approaches of regional (subnational) development programs in order to provide guidance on issues related to evaluating the impact of these programs. Regional development programs are designed to improve the income-generating capacity of, and reduce poverty in, a focus region within a country. The primary and unique characteristics of these programs lie in promoting a broad range of productive activities in a predefined region. To do this, regional development programs often need to address the institutional structure under which decisions are made as well as how economic resources should flow from the center to the local level. As such, the programs involve both productive and institutional transformation. Evaluating the impact of regional development programs is complicated by the need to assess both the impact of productive investment as well as the institutional transformation. As with all impact evaluations, evaluating regional development programs is thus much more likely to be successful if planned along with the design of the program. This guideline provides a summary of the options for setting up evaluations of regional development programs, while carefully considering the need to go beyond evaluating the impact on beneficiaries alone.

JEL Classification: C81, O12, O18, O22, R11, R58

Keywords: regional development, impact evaluation, development effectiveness, rural livelihoods, decentralization.

[*] Paul Winters, Associate Professor. Department of Economic, American University, Washington, DC. E-mail: winters@american.edu

[**] Susana Sitja Rubio, Sector Lead Specialist. Strategy Development Division, Inter-American Development Bank Washington, DC. E-mail: susanasi@iadb.org

The authors are grateful to Isabelle Mulin who obtained and synthesized the background material on Panama and other regional development projects supported by the IDB. We are also grateful to Carola Alvarez and John Horton for providing helpful comments on earlier versions of this guideline.

Table of Contents

1. Introduction

As part of a general strategy to reduce poverty, economic development programs often focus on a particular geographic region within a country with either high levels of poverty or a set of industries with potential for being engines of economic and social development, or both. The geographic region may be defined in different ways including rural areas, a region within the country, or a geopolitical unit such as a state, province or municipality. Regional development programs that focus on rural areas are generally designed to move beyond agriculture and are distinct from agricultural projects in that they recognize that rural households tend to be involved in multiple income-generating activities including nonagricultural wage employment and self-employment (FAO, 1998; Reardon, Berdegue and Escobar, 2001). In fact, recent evidence from a cross-country study suggests that about half of income generated by rural households is earned in the nonagricultural economy (Davis et al. 2010). Development programs that focus on a particular region or geopolitical unit usually include both rural and urban areas. This approach recognizes that the links between rural and urban development are often quite strong, particularly for intermediate-size towns and especially in the case of resource-based industries such as agriculture, resource extraction and tourism.

The objectives of regional development programs are to improve income-generating capacity and reduce poverty in the focus region. The primary and defining characteristics of these programs are therefore promoting productive activities, although programs often include investments in nonproductive activities that provide necessary infrastructure for economic activity and provision of services. Much of the challenge in designing regional development programs is that regional economies include multiple industries, and households living in those regions generate income from various activities. For a program based solely on a single industry, such as agriculture or tourism, it is possible to focus the components of the program on promoting that sector, although this in itself can be challenging. A truly regional program, on the other hand, needs to be sufficiently flexible to ensure that funds can be allocated to sectors with the potential to spur economic growth and raise the incomes of poor households. If it is sufficiently clear that a region has potential in a certain industrial area, a development program can be designed to promote that industry. Promoting a certain industry may be particularly appropriate if this develops an industrial cluster that takes advantage of agglomeration

economies. Of course, when there are multiple options for developing a region or where picking a clear winning industry in the program-design stage is difficult, it is necessary to consider carefully the institutional design of the program so that it allows identification of potential investment opportunities that promote development.

Identification of high-potential economic activities usually requires local information. Furthermore, promoting such high-potential activities requires a degree of local participation. As such, the expectation is that programs promoting regional development will include an institutional mechanism that incorporates local input in one form or another. When regional development programs involve substantial institutional transformation, they are often referred to as territorial development programs, and much of the focus of this guideline is on these particular types of projects. For a given program, the manner in which institutional change occurs depends largely on what current institutional arrangements exist and the degree of regional decentralization. In some cases, there may be an institutional vacuum if there is no clear government entity that should manage the program (Zezza et al. 2009). For example, programs focusing broadly on the rural economy may find that multiple line ministries (e.g. agriculture, commerce, tourism, etc.) and/or subnational governments should play a role in the program, but there is no single entity in a position to manage the program and no mechanism by which these agencies can coordinate activities. For a province or municipality, while such mechanisms may exist on paper, they may have limited political or fiscal authority. In such cases, developing the institutional structure to manage the regional development program becomes necessary. At the other extreme, the region in which a program is to be implemented may have complete political and fiscal authority and may be the logical counterpart for managing these programs. Depending on the current situation, a regional development program may need to address the institutional structure under which decisions are made as well as how economic resources should flow from the center to the local level.

In some ways, a regional development program might be a step in creating a decentralized structure for managing regional economic development. That is, the institutional structures set up for the program may eventually be provided with the political and fiscal authority to manage regional economic development. Alternatively, these structures may actually hinder decentralization by creating a parallel institutional structure. The debate over the role regional development programs play in the process of decentralization is similar to the debate

over social investment funds which provide funding to localities for social services through a demand-driven process that is managed by the national government. Social funds may be viewed as either a step toward full decentralization of social service provision or as a mechanism for centralized management of local investment (Faguet and Wietzke, 2006).

The purpose of this guideline is to discuss the objectives and approaches of regional (subnational) development programs in order to provide guidance on issues related to evaluating the impact of these programs. In doing so, the guideline proposes possible methods for assessing the logic model and identifying the impact of these programs. At present, the Inter-American Development Bank (IDB) has a variety of programs that promote regional development through differing mechanisms. In particular, this guideline focuses on those that can be defined as taking a territorial approach, which includes productive and institutional transformation, although much of the discussion is relevant for a broader range of projects. Given the importance of these programs and the continuing trend toward decentralization, assessing the impact of these programs is critical for identifying their success rate, for determining what models have been most successful, and for establishing how these programs link to the decentralization of economic development.

Towards this end, Section 2 of the guideline provides the theoretical underpinnings of regional development programs and the issues and trade-offs to consider in promoting these types of programs. The section concludes with a series of hypotheses to be tested with respect to regional development programs. Section 3 gives an overview and categorization of the types of programs that fit the mold of regional development programs. Using IDB-financed program examples from Panama, it discusses the expected impact of these programs and highlights the evaluation challenges. Section 4 discusses the methodological issues in identifying impact and proposes specific steps for evaluating programs and collecting the necessary information. The final section provides conclusions and key messages.

2. The Logic of Regional Development Programs and Expected Impact

The overall objective of regional development programs is to improve the capacity of inhabitants in certain geographic regions of a country to generate income. A regional, rather than sectoral, focus is based on evidence that households receive income from multiple sources, including from agricultural and nonagricultural activities, from wage and self-employment, and from public and private transfers (Davis et al. 2010). Focusing on a single sector, such as agriculture, may limit opportunities for some households or push households towards the promoted activity even when alternative options may be better. By recognizing that households employ different strategies to improve their livelihoods and that many households diversify their range of economic activities, regional development programs aim to facilitate the use of different pathways towards improved well-being.

In addition to acknowledging that households vary in their livelihood strategies, regional development programs also recognize that certain localities may have a comparative advantage in a particular economic activity. Further, it may be difficult for the national government to identify these activities or to have sufficient understanding of local conditions to know how to best invest public resources. Regional development programs seek means to provide public investment that facilitates the development of industries in localities according to their comparative advantages, with the vision of spurring growth poles in other areas. Of course, the expectation is that the private sector will also invest in these activities and that the role of public investment is to help promote investment through public-private sector partnerships. While certain social objectives—such as poverty reduction—may be achieved through such public investment, it is also motivated by the desire to create clusters of economic activities when possible.

The basic premise of the literature on clusters, industrial districts and agglomeration economies is that location matters since the geographic concentration of economic activities assists in making those activities competitive. Clusters promote competition through (1) increasing the productivity of firms through better access to specialized inputs and employees, information, customers, marketing channels and institutions and public goods; (2) a greater ability to innovate through better contact with suppliers and customers; and (3) lower barriers to entry due to better information and opportunities in existing clusters (Porter, 1998). Firms within

an industrial district (cluster) can benefit from agglomeration economies—that is, the increasing returns generated by the positive externalities of the activities of one firm on other firms—through exchanging specialized labor, specialized intermediate inputs and knowledge (Krugman, 1991).

These arguments suggest that the synergies created by the geographic concentration of economic activities enhance the competitiveness of a region. Government investment in developing clusters is justified since there are positive externalities in private investment, yet the private sector will likely invest less than the social optimum. There is on-going debate regarding the role of the public sector in the creation of existing clusters and in the formation of new clusters, with some arguing they develop organically without government intervention. Nevertheless, Porter (1998) maintains that governments, in addition to providing a sound macroeconomic environment for economic activity, may need to take specific actions to upgrade existing clusters or foster the development of new clusters, particularly those emerging from established ones. This provides a justification for development programs that use public-private sector partnerships to develop clusters.

The basic logic of regional development programs is therefore to improve income generation in the specified region through public investment that recognizes that households and localities may specialize in different activities and that seeks to facilitate, when possible, the development of clusters of related economic activities. Of course, industrial cluster formation is not promoted solely through regional development programs and there are other programs, often referred to as competitiveness programs, which are designed specifically to promote such clusters, usually on a broader scale than regional development programs. Furthermore, as discussed below, regional development programs involve more than just cluster promotion and involve institutional transformation for purposes other than cluster development.

The challenge in proceeding with supporting this type of investment is that local information and participation are critical both in identifying the best opportunities for investment and in providing complementary private-sector investment. In fact, a key to regional cluster formation is noting the uniqueness of regions not only in terms of resources available but also in terms of political institutions and cultural factors such as the level of social capital (Hospers and Beugelsdijk, 2002). Furthermore, the supported activities may range across a number of line ministries and thus require coordination with those ministries as well as between national,

regional and local governments. The ability of regional programs to do this depends largely on existing institutional structures. It may be that there is an institutional vacuum where no government entity is responsible for the development of a region, making public investment unlikely to occur unless that vacuum is filled (Zezza et al. 2009). Regional development programs, therefore, require a carefully constructed institutional structure to identify and manage investment in regional economic activities. Part of the logic of these programs, then, is productive investment combined with institutional transformation. This institutional transformation is based on the theory of decentralization in the provision of goods and services. Previously focused mostly on fiscal considerations, decentralization is now regarded as a process with multiple vectors. It incorporates issues of political and fiscal decision making, with a planned devolution of responsibilities and deconcentration of economic activity from a strong pole of growth at the center to regional poles to develop clusters of activity.

The decision of how, when and what to decentralize when looking to promote regional economic development is one of maximizing efficiency while addressing the indirect effects of these decisions. In the Decentralization Theorem (Oates, 1972, p. 35), the efficiency argument for decentralizing the provision of certain goods and services is formalized:

> "For a public good the consumption of which is defined over geographical subsets of the total population, and for which the costs of providing each level of output of the good in each jurisdiction are the same for the central or for the respective local government it will always be more efficient (or at least as efficient) for local governments to provide the Pareto-efficient levels of output for their respective jurisdictions than for the central government to provide any specified and uniform level of output across all jurisdictions."

For regional economic development, the assumptions behind the theorem are that: (1) the local population exercises choice in demanding public investment; (2) local populations know their preferences and priorities better than regional or national levels of government; (3) the funding of investments follows the "Benefits Principle of Public Finance", that is that user fees should fund investments as much as possible, and basic community investments are funded through local taxes; and (4) investments mandated by higher levels of government are funded through intergovernmental transfers (grants).

The theorem above ignores three issues: spillover effects (externalities), population mobility, and the fact that the local government would provide the same level of services as the

central government, and provide those services uniformly across jurisdictions (Oates, 2007). Since residents of one jurisdiction can benefit from the public sector activities of another, spillovers can take place across jurisdictions leading to underspending as jurisdictions fail to recognize these externalities or attempt to "free ride" on other jurisdictions. Population mobility occurs when there are different levels of investment in different areas motivating individuals to move to a jurisdiction that provides their preferred level of services.[1] The uniform provision of services assumption fails to recognize that local governments know local preferences better than central governments and are not likely to provide a uniform level of service. While central provision internalizes spillovers, decentralized provision creates other sets of variables to consider when determining the Pareto-efficient level of public expenditure on any given public good (Koethenbuerger, 2008).

Along with these three limitations, a decentralization model must recognize the political system under which it is taking place, the fiscal constraints within a region or country, the institutional capacity of local governments to manage new responsibilities, and the regional disparities both in needs, resources and capacity that can make decentralization actually lead to greater, not less, inequality. A decentralized model does not necessarily provide for redistribution in a way that leads to better outcomes for the general population. In fact, in a decentralization process that implies autonomy at the local level for managing revenues and expenditures, there is a conflict between local autonomy and national redistribution (Bird, 2003). In this respect, a regional economic development program needs to be viewed from a national perspective of deconcentrating economic activity in ways that spur growth but do not exacerbate inequities between regions.

Because of these issues, decentralization inevitably involves trade-offs and there is thus no perfect model to implement it. However, there may be an optimum equilibrium somewhere between pure centralized control and total local autonomy; what this optimal level of decentralization is will depend on current conditions in a country or region. Further, the expectation is that this may change over time and, in pursuing a decentralization policy, countries may take piecemeal steps to reach that optimum level. Since decentralization also incorporates economic, political and administrative factors, there may be different degrees of

[1] Of course, this requires an assumption that mobility has no cost, and that there is no discrimination between long-term and newer residents.

decentralization for each of these areas. Finally, note that what may be an optimal level of decentralization for service provision may not be the same for the management of regional economic development. In the case of service provision, the focus has been on communities and municipalities which have led to the strengthening of relatively small administrative units. This leads to the emergence of a strong central government with strengthened municipal or local governments with little in between. While this approach is possibly appropriate for service provision, regional economic development and the development of industrial clusters may require a broader perspective. The central government, given its location in the capital and its limited information on local conditions, is not always in the best position to manage regional economic development. Yet municipal governments are often small in size and population and, particularly in rural areas, are generally not of a sufficient size to take on the task or to have the sufficient incentives or resources to invest. Underinvestment is particularly likely to occur if there are spillovers in economic activities. Thus, for regional economic development to be successful under a decentralized system, the decentralization must be at a level that takes into account the trade-off between local information and participation and the regional scale of economic activities.

In general, the institutional transformations that are promoted within the structure of regional development programs take into account the need to identify local preferences and develop local capacity to manage resources, but do not usually include any mechanism for local entities to generate their own resources. As such, the institutional structure tends to be a partial step to decentralization as defined by the above theorem. This may mean that regional development programs are a step towards a more complete decentralization program, but are temporary in nature. The institutional approach used in these programs may also be an attempt to balance the benefits of decentralization (better information) with the possible costs in terms of efficiency (spillover effects, redistribution, etc). Regional development also tends to encompass larger administrative units than municipalities, based primarily on the argument that these units are in most cases too small to manage regional economic development and a larger unit of intervention is required.

2.1. Expected Impacts

Given the logic of the regional development programs, a few clear hypotheses on their potential impact can be identified although the approach to evaluating these programs is complicated by the need to consider not just whether they had an impact on beneficiaries in a target region, but whether they promoted the development of clusters and whether the institutional approach used was appropriate given the country context. For some of these hypotheses, standard impact evaluation approaches can be used while others require a more specific approach. Generally, the hypotheses that can be tested are as follows:

1. *H_0: The regional development program identified productive activities for investment from different sectors including both agricultural and nonagricultural activities.*

   The expectation is that the portfolio of investments funded through regional development programs represents a range of economic activities that are appropriate for local conditions. While similar activities may be found in certain areas, if the diversity of activities is minimal, there seems to have been limited benefit in using local information to determine the best investment.
2. *H_0: The investments in individual projects improved the welfare of members of beneficiary localities through enhancing their ability to generate income.*

   Even if the chosen activities are appropriate, if the investment itself did not improve the well-being of participants in recipient localities, it provides insufficient justification for investment.
3. *H_0: The benefits of funded investments were sufficiently high to justify the costs associated with funding the program.*

   Assuming the welfare in recipient localities is improved, the question is whether the cost of achieving those benefits is worthwhile (a positive rate of return). Could these benefits be achieved in an alternative way at a lower cost (was this the most cost-effective alternative)?
4. *H_0: The particular institutional structure used in the program was the best manner in which to deliver funds for productive investment.*

   Even if the rate of return is positive and sufficiently high to justify investment, there remains a question of whether the institutional structure used is the best mechanism by

which to choose and make those investments—meaning the relative return exceeds that of an alternative institutional structure for identifying investment opportunities.

5. *H_0: The particular mechanism of delivering funds allows for public investment that facilitates private investment and helps to develop clusters of economic activities that benefit from agglomeration economies.*

   While public investments may have a sufficiently high return to justify them and may induce private investment, this does not necessarily imply that they help develop clusters of activities. For this to happen, the expectation is that there might be a pattern of investment in related activities in certain areas. If there is, it may be possible to try and test in the longer run if agglomeration economies exist (although this may be difficult to attribute to the program investment alone).

Given the different challenges in testing these hypotheses, the approaches to evaluating each hypothesis are discussed in Section 4. Prior to that discussion, in order to put the evaluation discussion in context, Section 3 considers regional development programs supported by the IDB.

3. Evaluating Regional Development Programs

Over the last decade, the IDB has had a number of programs that can be classified as taking a regional development approach in that they involve a predefined geographic space within a country, include a focus on multiple productive activities, and incorporate institutional transformation. The regional development programs span all types of countries that borrow from the Bank, from large federal states such as Brazil, to low income, small countries like Honduras. In some cases, the IDB has developed these programs in multiple regions within a country to spur growth in distinct areas of that country. There are also cases of regional, multicountry structured programs, such as the *Sustainable Management of the Lempa River Basin* in Central America (2001), which incorporated within a single program loans to three countries (El Salvador, Guatemala, and Honduras) and had local, municipal and regional activities.

3.1 Panama case study

In order to evaluate the impact of any program, it is critical that the evaluator have a clear understanding of how the program was designed and implemented. To highlight the issues in evaluating regional development programs, we have chosen to use the regional development programs supported in Panama. The IDB has provided loans covering all of the regions of the country over the last decade. The first of these programs was the Darien Phase I, followed by Bocas del Toro Phase I, Chiriquí, Darien Phase II, Bocas del Toro Phase II, Colon Phase I, and Central Provinces. Five of the seven programs were chosen as they were assessed recently by CONADES, a public organization created by the National Government to enhance development objectives in the country. These five programs, implemented in three regions, include Darien Phase I and II, Bocas del Toro Phase I and II, and Chiriquí. The basic design of these programs is noted here to provide insight into the challenges of conducting an evaluation of a regional development program.[2]

The Panama case is an interesting one because regional governments were not elected at the time the programs were designed, but appointed by the national government, and therefore an extension of the national government. Nevertheless, they recognized the benefits of knowing

[2] This set of programs in Panama was chosen for largely because of questions raised regarding whether they were effective. The information presented here was collected as part of an effort to evaluate these programs although an evaluation was never implemented.

local preferences, and therefore had institutional structures that accounted for this factor. The programs aimed to strengthen the institutional structure at the regional and local levels to promote decentralization; to improve productivity and economic diversification; to enhance the development of industrial clusters in the region while emphasizing sustainability and resource management; and to improve basic services and revamp the transportation infrastructure. Though each program had unique subcomponents, the institutional mechanisms for program conception and development are similar across the board, and probably a reflection of Panama's political system. The ability of these programs to promote an optimal decentralization mechanism was limited by the political-administrative structure of the Panamanian government. Furthermore, management of resources remained at the national level, therefore diminishing the impact of decentralization. Nevertheless, a national strategy to develop new development poles through the programs was likely to lead to reducing the effects of spillovers, migration and uneven service provision.

To understand how productive activities were chosen within the programs, a brief description of how the programs were designed is necessary. In general, the regional development programs took the following steps to indentify productive investments.

1. *Diagnostic and baseline collection:* The first stage of the preparation of the regional development programs consists of the government, with IDB support, preparing a regional development strategy based on the largest economic sectors for each region and taking an inventory of all the potential natural resources that could be used towards income generation. The most prominent local development agents, public and civil leaders, and vulnerable groups participate in this process. This baseline information is then incorporated in the initial diagnostic of the development potential of the region.
2. *Analysis of development scenarios:* The second stage focuses on the creation of various development scenarios for the region. With the assistance of consultants, a series of development scenarios are presented that take into account provincial, regional, national and international perspectives (depending on the location of the borders). Ideally, an economic analysis is conducted as an input into this process. In the case of the Panama programs, these were not included in the program documentation, or in the subsequent evaluations, therefore the degree to which this analysis was done is uncertain.

3. *Formulation of sustainable development strategy*: The next step is to formulate the principle strategic guidelines of a regional development strategy in coordination with civil society, the private sector and regional and local governments, while also specifically including vulnerable groups (indigenous and others). In parallel, Provincial Territorial Plans that subdivide the province, depending on composition, resource differentials, and political alignments, are created. The strategic needs within the region are then prioritized.
4. *Action plan to implement strategy*: The Strategic Provincial Action Plan for Investments is then created based on 5, 10 and 15 year time horizons. It includes investments in infrastructure and basic service projects, institutional strengthening, natural resource management, rural development programs, and conservation of protected areas as well as other regional-specific projects.
5. *Formulation of bank-loan program that will finance part of proposed Action Plan*: The last step is the formalization of the Bank Loan Program that considers all the planning instruments and defines the elements of the Action Plan to be financed with funding via the national and local governments.

These steps were taken prior to the implementation of the loan in order to create the regional framework for program implementation. This was done through a consultative process with both national and local actors, both with existing formal institutional structures as well as informal channels. Out of this process came an institutional mechanism for decision making and for the selection of specific investment projects once the overall programs were implemented in each region. This is discussed next.

For each of the Panamanian regional programs, similar institutional mechanisms were used for investment project selection, implementation, and fund disbursement. The executing agency for the programs consisted of the National Council for Sustainable Development (CONADES), which represented the Ministry of the Presidency (MP), and a Program Executing Unit. Together they were responsible for coordination, general administration, financial and accounting management, and monitoring and evaluation. For Darien and Chiriqui, the executing agency was assisted by a Program Technical Team, which (1) coordinated program execution, including hiring consultants and contractors for all components; (2) managed and supervised activities related to program implementation; (3) organized product presentations for the MP, the

Ministry of the Economy and Finance, the provincial technical board, and the provincial council; and (4) provided ongoing assistance to the subregional committees and other community organizations to help them develop their subregional action plans.

At the provincial level, the national government was assisted by the Provincial Council (PC), made up of representatives from each *corregimiento* (county). The PC promoted, coordinated and harmonized official activities at the provincial level, and saw to the promotion, coordination and consensus-based approval of the programs' annual work plans (AWP). The Provincial Technical Board, led by the appointed governor and consisting of the regional offices of the ministries and national institutions with sector authority in the province, supported the council. They: (1) ensured that investment initiatives were consistent with sector regulations and specifications; (2) consolidated and coordinated the program AWPs, based on the programs given priority; (3) oversaw the monitoring and execution of the AWP; and (4) reviewed the program's annual reports and financial statements and submitted them to the PC. Local governments and municipalities participated in activity execution within their sector, and were responsible for the dissemination of program information throughout their respective communities. Co-executing, decentralized government agencies also participated in the program by providing support to the Program Executing Unit to: (1) devise, review and participate in the bid evaluation commissions; (2) perform technical and environmental supervision of the investments, as well as periodic inspections of works and monitoring of their operation and maintenance in accordance with generally accepted technical standards; and (3) prepare the monitoring reports for activities under their responsibility. They were also expected to participate as members of the Provincial Technical Board in preparing and building consensus for the AWPs, jointly with the Program Executing Unit.

With these institutional structures in place, the following step-by-step process was how projects were selected:

1. Communities and local municipalities prioritized their needs and interests with regard to productive and infrastructure activities and projects, preparing the respective profiles or complete projects, based on the criteria set in the operating regulations.
2. Local beneficiaries submitted their proposed projects to their respective municipality.
3. Municipalities turned this information in to the Provincial Technical Board, which was responsible for evaluating the feasibility of the projects.

4. If the evaluation was positive, technical experts helped the beneficiaries ensure the project complied with sector regulations and specifications.
5. Feasible projects were sent to the Program Executing Unit for approval.
6. Once approved, the projects were funded.

As part of this process, the institutional mechanism encouraged community members to form activity associations in order to select a project that would assist them in promoting their activity throughout the region. This assumed that each region had specialized economic activities that could benefit from additional capital to enhance efficiency and productivity through technological and other inputs. In Darien, for example, funds were used for technology transfers to associations engaged in the agriculture, agroforestry and the fishery sector. The regional programs in Chiriqui and Bocas del Toro were aimed at helping banana producers reach exportable status. Tourism activities in Darien and Bocas del Toro were also included in promoting economic opportunities that benefit regional development. Overall, the process therefore led to investments in a range of agricultural and nonagricultural activities as well as infrastructure projects to strengthen these activities.

As seen through this model, projects were identified based on local demands and prioritized at the municipal level. In this manner, each program had an element of decentralization, at least for economic and small infrastructure projects. Nevertheless, these projects only account for a small fraction of the budget. The major projects continued to be implemented using a top-down hierarchical framework. These projects included provision of major infrastructure works and institutional-strengthening services. On the one hand, the rigid institutional mechanism seemed to go against the need for regional programs to be flexible in order to ensure that funds were allocated towards sectors, specific to individual regions, which would promote desired economic growth. However, it appeared that local committees provided the necessary input to national heads so as to allocate budgets according to prioritized local needs. As municipalities were still strengthening their governance to support a decentralized model for provision of public services, the need for national support through decentralized national agencies and government oversight appeared to be both intended and logical. The model that combined national decision making with voiced local demands is potentially a preliminary step towards a decentralization approach for regional development programs, particularly when decentralization implies a new governance model.

3.2 Challenges in evaluating regional development programs

The Panama regional development programs led to a series of direct investments in rural areas with clear beneficiary populations. The challenge in identifying the impact of the investment projects on these direct beneficiaries lies in addressing the standard evaluation issue: what would have happened to these beneficiaries in the absence of the program? The need is to identify a counterfactual. To do this involves following standard evaluation practice and, as seen below, in this particular case mirrors the evaluations that have been undertaken of Social Investment Funds (See Rawlings, Sherburne-Benz and Van Domelen, 2004).

However, as the Panama programs highlight, much of the innovation in regional development programs is reflected in the institutional structure that is designed to deliver public investment. The programs can be viewed as an attempt towards regional decentralization that improve the capacity of inhabitants of the geographic region to generate income and that assist regional institutions in improving their ability to service the needs of their constituents. The interesting question to address is thus not just whether the programs had an impact on beneficiary populations, but whether that impact would have been the same under an alternative institutional structure. These programs then have the added challenge of needing to carefully consider and assess the programs' institutional structure.

4. Approaches to Evaluating Regional Development Programs

In this section, possible approaches to testing the hypotheses at the end of Section 2 are noted. The hypotheses are laid out in order; that is, the first hypothesis requires understanding what projects were funded, which is necessary for evaluating the impact on beneficiaries (the second hypothesis). The second hypothesis identified the direct impact of the program, which plays a part in assessing the returns to the program (the third hypothesis). And so on. Each of the hypotheses is then discussed in order.

As a general rule, designing and setting up an impact evaluation *ex ante*—prior to the implementation of the program—facilitates conducting an evaluation and in particular determining impact (testing the second hypothesis). While it is possible to set up an evaluation *ex post*—after the program is in execution—creating a proper counterfactual is much more challenging. Although it would be too late to design an evaluation for the Panama programs since they are in execution, future regional development programs could be designed to facilitate an evaluation. As such both *ex ante* and *ex post* evaluation designs are considered.

4.1 The first hypothesis: Identification of productive activities

The expectation is that the portfolio of investments funded through regional development programs represents a range of economic activities that are appropriate for local conditions. While similar activities may be found in certain areas, if there is limited diversity, there seems to have been little to gain from using local information to determine the best investment. For example, in the case of Panama the expectation is that within each province there would be some diversity of activities funded both within each province as well as across provinces. If the same types of projects are being funded everywhere, it suggests that using a decentralized approach did not help in selecting projects. Further, a limited number of activities may reduce the likelihood of achieving gains through agglomeration economies. Up to a degree, the more diversified the portfolio, the more likely that synergies are developed. Of course, excessive diversification is also unlikely to bring about any synergies in activities so the expectation is that activities will be related in some way. Precisely defining the optimal level of project diversification is challenging, but a key consideration is whether the projects are related and

whether they are likely to create greater economic activities together than if funded in isolation—that is, whether the whole is greater than the sum of the parts.

The process of identifying projects for funding and the logic of the choices made should assist in determining if the sets of projects were viewed as related. As such, it is important to know not just which projects were funded, but also those that were proposed and rejected. This helps to determine whether the process facilitated investments in related projects and rejected those that were both of lower return and unrelated to other projects. The proposed first step in the evaluation is then to carefully identify the projects that were considered for funding as well as those that were actually funded or will be funded if the program is still in execution. These activities can be assessed to see whether the program appears to have taken appropriate advantage of local information and helped develop clusters of activities.

Note that an added reason to have information on both funded and rejected projects is that they represent a potential counterfactual that could be used for comparison. Whether this is a legitimate counterfactual depends on the process by which projects were determined for funding. The specific information required is as follows:

1. The intended and actual process for identifying projects for funding including an assessment of the costs associated with the process.
2. The list of individual projects funded through each program, the amount of funding, the timing of the funding, the specific location of the project investment, and the type of project that was funded.
3. The list of individual projects that requested funding through each program but were rejected, the amount of funding requested, the specific location of the requested project investment and the type of project that was requested.

If the need for this type of information is recognized *ex ante*, a careful monitoring system can be put in place as part of the program that captures all the relevant information. As seen below, this can greatly facilitate an impact evaluation. If the monitoring system is not designed to collect this information, it must be collected *ex post*. This generally requires trying to track down program documents and government or IDB staff who worked on the program to reconstruct what was done and what types of investments occurred. In the case of the five Panamanian programs, an *ex post* assessment of the programs did find that the selection process led to investments in a range of agricultural and nonagricultural activities as well as infrastructure to support their

development. The specific productive investments ranged from fisheries commercialization, artisanal diversification, forestry, agricultural commercialization, farming, tourism capacity and small-scale infrastructure. In some cases though, identifying the industry of investment was difficult since records only showed the amounts provided for investment, but the original proposals were not found. Further, it was difficult to determine what projects were proposed and rejected in the process and the reasoning for the choices made. Thus, while it was possible to identify some projects that were funded, it was difficult to understand the process by which they were chosen and those that were not chosen. This makes it very difficult to provide an assessment of the impact of the program and highlights the importance of *ex ante* evaluation planning.

4.2 The second hypothesis: Impact on beneficiary localities

Even if the chosen activities appear appropriate, if the investment itself did not improve the well-being of individuals in recipient localities, it provides insufficient justification for investment. Testing this hypothesis requires conducting a standard impact evaluation. This type of evaluation mirrors those done of Social Investment Fund programs (SIFs). These evaluations sought to determine if SIFs achieved their objectives with respect to improving the well-being of final beneficiaries within communities that received funds (Rawlings, Sherburne-Benz and Van Domelen, 2004).

The key to evaluating the impact of a program is to establish a counterfactual—that is, an estimate of what would have happened in the absence of the intervention. In evaluating the impact of a program on intended beneficiaries (the treatment group), creating a counterfactual is usually accomplished by identifying a group of individuals who did not receive the program but are similar to the treatment group in all other ways (the control group). The control group then represents what would have happened to the treated group in the absence of the program. With an established counterfactual, comparing treatment and control groups using indicators of program success is possible. The ability to establish a counterfactual and the method of conducting an impact evaluation depends on the type of data that is collected and whether the evaluation is designed *ex ante* or *ex post* as noted below. We consider two broad categories of impact evaluation: experimental and nonexperimental.

4.2.1 Impact evaluation design using experimental methods

The advantage of designing an evaluation prior to program execution is that it provides the ability to obtain baseline information prior to execution and to adjust the implementation of the program to facilitate an evaluation. In the case where funds are allocated based on demand, as in regional development programs and social investment funds, the best way to set up an evaluation is by using a "pipeline comparison". In these programs, a fund is set up for investment and a process by which those funds will be used for specific investments is determined. The process should include the parameters for funding projects (the types of project eligible for investment) as well as how to prioritize those projects. Of the proposed projects, presumably only some will be selected for funding based on some predefined criteria. This same process should rank the projects from the most to least desirable and should include some cutoff by which certain ones will or will not be funded.

Assuming there are a sufficient number of projects included in the "pipeline"—that is, the list of projects that will ultimately receive funding—it should be possible to randomly assign those that will receive funding in the first stage of the program and those that will be funded at a later stage. Those that receive immediate funding can be used as a treatment group and those that receive later funding as the control group. With enough projects to allow a sufficient sample size, the random assignment should be sufficient to ensure that the treatment and control groups are alike in all ways except for having received the funding. With the collection of baseline data (prior to program funding) from households in the treatment and control groups, this can be verified.

In practice, a number of issues might come up in randomly assigning projects. First, it may be the case that there are a limited number of projects to be funded. Even if this is the case, attempting to randomly assign the projects to control and treatment is worthwhile. But in such a case, there is a value to stratifying the projects and randomly assigning by strata. If projects are ranked by expected returns, it may be worthwhile to divide them by level of return and randomly assign from each strata to ensure a range of high to low return projects in both control and treatment. If projects vary by sector—for example agriculture, tourism, etc.—and there is a logical grouping of projects, they may be stratified by these groupings to ensure that a range of sectors is included in control and treatment. Of course, stratification has implications for the sample size required to conduct the analysis. If measuring the impact of the project within each

grouping, or strata, the size of the sample must be sufficiently large to ensure that the "power" of the sample is such that it can capture the impact of the project within each of these groupings. These issues are addressed more fully elsewhere and should be carefully considered.

An additional problem with using experimental methods in practice relates to the fact that in some cases not all projects to be funded are submitted and selected at once. This means that not all eligible and funded projects are known at the outset of the program. Doing a pipeline comparison using the projects funded at different stages of the program is still possible, but concerns will be raised over whether the projects that are funded initially are somehow fundamentally different from those that are funded later. This is likely to be the case if there is some reason why certain groups or certain areas were able to put together projects for funding more quickly. If earlier- and later-funded projects differ, the control may not be a perfect comparison group to the treated group and the manner in which the analysis of impact is done must be carefully considered. Nonexperimental methods, noted below, must be used to conduct the analysis.

If random assignment of the pipeline to control and treatment is possible, the best method of evaluation is to use a double-difference approach. This requires collecting baseline information from a random sample of control and treatment households. A follow-up survey should also be administered after treatment in the treated communities, but prior to the treatment of the control group. Of course, a sufficient amount of time must have transpired between the treatment of the treated group and the control group to allow for the project to have had an impact. The final data set then has detailed information on the treated group (P=1) and the control group (P=0) prior to program implementation (baseline, so T=0) and after program implementation (post-treatment, so T=1). For an indicator y_i, the impact of the program can then be determined by estimating the following equation:

$$y_i = \alpha + \gamma P_i + \lambda T_i + \delta(P_i * T_i) + e_i \qquad (1)$$

In equation (1), γ estimates the initial difference between the two groups so

$$\gamma = E[y_i \mid P=1,\ T=0] - E[y_i \mid P=0,\ T=0],^{3}$$

[3] Note that E [.] is the expectations operator and gives the average value for the object to which it is applied. In this case it refers to the average value for the indicator y_i for the treated group at baseline minus the average value for the control group at baseline.

λ estimates the change in the outcome that occurred over time in the control group due to other factors unrelated to the program or

$$\lambda = E[y_i \mid P=0, T=1] - E[y_i \mid P=0, T=0]), \text{ and}$$

δ estimates the impact of the intervention so

$$\delta = E[y_i \mid P=1, T=1] - E[y_i \mid P=1, T=0]\} - \{E[y_i \mid P=0, T=1] - E[y_i \mid P=0, T=0].$$

Estimating δ through (1) assumes that both groups would experience the same time trend (λ) in the absence of the program so that once initial differences (γ) and the time trend are accounted for, the remaining difference between the two groups can be attributed to the intervention. This approach is referred to as a double difference or difference-in-difference since it controls for initial differences across space (between control and treatment) and differences over time, making any remaining difference due to the program. If the control and treatment are randomly assigned, this is generally considered a reasonable assumption.

Even with random assignment, however, there is the possibility that the control and treatment will not be identical, particularly if the sample size is not very large. Of course, using the baseline data, the characteristics of the control and treatment groups can be compared to see the results of the experiment and this is always recommended. Further, including additional variables (*X*) in equation (1) can help control for any preexisting observable differences that may exist between control and treatment and helps in the precision of the estimates (Stock and Watson, 2003). In practice, even when experimental data is available the following regression is often estimated to get impact:

$$y_i = \alpha + \gamma P_i + \lambda T_i + \delta(P_i * T_i) + \beta X_i + e_i \qquad (2)$$

where X_i is a vector of baseline characteristics

β is the influence of those variables on the outcome variable of interest,

and the other variables and parameters of the model are as previously noted. This approach is referred to as a double difference with conditioning variables. In general, including conditioning variables can help improve estimates of impact (since they help deal with preexisting differences), but cannot hurt (since if there are no preexisting differences the results are as they would be without the conditioning variables).

4.2.2 Impact evaluation using nonexperimental methods

When implementing a program to facilitate an evaluation through the use of an experiment is not possible, nonexperimental methods of evaluation must be used to establish program impact. As when using an experiment, the key to accomplishing this task is establishing a counterfactual. Creating a counterfactual depends on what data can be collected and, closely linked to this, the ability to employ a method that helps to identify the counterfactual and avoid biased estimates of impact. Since the data collected and method employed are closely linked, we begin by considering the options for data collection.

For regional development programs, data collection for establishing impact is going to involve administering surveys in treatment and nontreatment areas (at least at the time of the survey). The more closely the nontreatment area represents a good counterfactual, the easier it is to establish impact. Three options for nontreated communities to include in the data collection are: (1) *pipeline communities* who will receive a project in the future but have not yet started receiving program benefits; (2) *rejected communities* who applied for funding but did not get it; and (3) *nonprogram communities* who did not apply or were not eligible for the program.

Pipeline communities represent a good counterfactual since they are clearly eligible for the program and thus must have some characteristics that make them similar to those communities that were already treated. Yet since they did not receive the project early on, there is a concern that there is some reason that they received it later. For example, the program may decide to treat those communities with the highest-expected returns first and those with lower-expected returns later, indicating that those communities later in the pipeline have lower-return investments. Without controlling for this, the impact of the overall program might be overestimated since it reflects not only the benefits of the program, but also the fact that these early investments are high-potential communities. Therefore, it is critical to understand the reasoning for the order of treatment within the pipeline. If the ordering is arbitrary—reflecting, for example, political considerations rather than returns to projects—it may be easier to use the project pipeline as a counterfactual.

It may seem that rejected communities should never be used as a counterfactual by virtue of the fact that they were rejected for a program—they must not have met program criteria for some reason. But rejected communities did decide to apply for a project, suggesting that they at least perceived they met the criteria. They also must have a degree of organization, which

allowed them to apply. Whether they can be viewed as a reasonable counterfactual depends on the reason for their rejection. If the expected returns to the proposed project were not sufficiently high, it may be still possible to use the communities as a counterfactual, especially if they were near the borderline of acceptance. If the proposal was not approved because of a lack of funding or political consideration, again they may be a useful counterfactual. Once more, understanding the process for accepting and rejecting the proposed projects of communities is important.

Finally, there is the option of using nonprogram communities. In doing so, the first consideration is why they are not part of the program. Are they ineligible because they could not have applied for funding or are they in fact eligible but chose not to apply? If they are ineligible, the reason for this becomes important. It could be simply that they are not in the geographic area considered for inclusion in the program. This could mean that there are similar communities in a neighboring geographic region that were not part of the program. If they were within the eligible geographic region but chose not to apply for funding, there is the question of why they chose not to. Is it that they did not have a good project to propose, lacked the organization to put forth an application, or were unaware of the project? Whether it is because a community is ineligible or chose not to apply, the reason for them not receiving a project needs to be clear.

As can be seen from the discussion of the options, what is critical in all cases is to establish (1) whether any nontreated communities can be considered sufficiently similar to treated communities to be a legitimate counterfactual, and, (2) if these communities are used, what potential differences between the treatment and the identified control might lead to biased estimates of impact. Considering these potential sources of bias helps to decide whether any of the evaluation approaches are suitable to address this bias and allow for a reasonable estimate of impact. Some of the commonly used nonexperimental methods of impact evaluation are discussed below, but prior to considering these methods a discussion of baseline data and its role in evaluation is warranted.

While it is possible to conduct an impact evaluation without baseline data using nonexperimental methods, one must assume that the control group has been nearly perfectly created. More importantly, without baseline data it is impossible to test whether this assumption is reasonable. Even when using an experimental approach, as noted above, checking the baseline to see if the experiment created a good counterfactual is recommended. Given the difficulties noted above in creating a good data set, in the case of nonexperimental evaluations, verifying the

pretreatment similarity of the treatment and control groups is even more important. Along with verifying whether the control and treatment are reasonably comparable, baseline data is also useful in identifying impact since it can be used to control for any preprogram differences in control and treatment groups that remain after the creation of the data. Recall that a good counterfactual is one in which the control group is similar to the treatment group in all characteristics except in receipt of the program. If there are any differences between these groups, the baseline may be used to remove any of these remaining differences.

Having baseline data assumes that the decision to conduct an evaluation has been made *ex ante* and program leaders had the foresight to recognize the value of baseline data. If there is no such foresight or the decision to evaluate the program is taken *ex post*, the evaluation may have to be done without baseline data. In such cases, carefully selecting the control group remains critical and the data collection should be done in such a way as to solicit as much information about the characteristics of beneficiaries and nonbeneficiaries that could not have been influenced by the program. It may also be possible to use secondary data from previous surveys or censuses to assess treatment and control groups prior to the program. Any information that facilitates the ability of the evaluator to understand potential differences between control and treatment groups is helpful. Finally, even if true baseline data cannot be collected, panel data—that is data from two points in time for the same set of observational units—can be useful in identifying program impact, as it may help in controlling for preprogram differences in control and treatment communities.

The appropriate nonexperimental method for conducting an evaluation depends on the data collected and should be determined in conjunction with data collection. Providing a detailed description and the appropriate background literature on nonexperimental methods is beyond the scope of this guideline. Instead, a basic overview is provided to help assess the possibilities in conducting an evaluation. In employing any of these methods, readers should refer to more detailed descriptions of the approaches.

We begin by considering the most common nonexperimental approach: propensity score matching or PSM. The basic idea of PSM is that while the available data from control and treatment communities may not be perfect, within this data there exists a set of treatment and control observations that are similar in all ways except that the treatment group received the program. The key then is to come up with a way of identifying, or matching, these two groups

together. If they can be matched, it is almost as if an experiment was done, which is why this approach is often referred to as a quasi experiment.

In practice, PSM uses the predicted probability of participation in the program, referred to as the propensity score, to match treatment observations to control observations. The propensity scores are calculated using a probit or logit regression on program participation using all the data (treatment and control) and employing as explanatory variables a set of characteristics expected to adequately predict participation. The variables should be preprogram or, if collected after program implementation, should be variables that could not have been influenced by the program (that is, exogenous variables). Since these characteristics are observable, PSM assumes that program participation is related to observable characteristics of participants as opposed to unobservable characteristics. As a general rule, when a program determines who can participate there is less concern that unobservables influence participation since the program had to have observed the characteristics of the participants to allow them to participate.[4] On the other hand, when participants self-select into the program there is more likely to be concern that some unobservable characteristics—e.g. motivation, entrepreneurial ability, leadership, etc.—is likely to influence participation. Ultimately, it becomes the judgment of the evaluator to determine if this assumption is reasonable.

Once propensity scores are determined for control and treatment observations from the regression on participation, a quick assessment of the scores obtained is recommended. Since the idea is to match treatment and control observations, the distribution of these scores must overlap at least to a degree. The region of overlap is referred to as the area of "common support" and there should be enough of an area of common support to allow for matching to be done. It is common for the distribution of scores for the treated and control groups to be graphed to get a visual representation of the distribution of scores. Scores outside the area of common support may be trimmed from the sample.[5] Along with checking the common support, the "balancing properties" of constructed data set can be checked. In theory, the treatment and control group in the constructed data set should balance in the sense that there should be no differences between the two groups in terms of the preprogram variables. To check the balancing properties of the

[4] This assumes that the evaluator knows the factors that influence participation and has information on those factors.
[5] Some evaluators often trim control scores that are less than the lowest treatment score and the treatment scores higher than the highest control score. Others trim the top and bottom 1 percent of 5 percent of scores. There is some debate about the appropriateness of trimming observations and the interested reader should refer to the literature in this area.

constructed data, tests of differences of mean value of the variables used to create the scores are often done for the entire sample and even for subsamples such as each quintile of propensity scores. This latter step is to ensure that similar scores do not occur from different combinations of characteristics. In these tests of difference, the expectation is that there will be limited differences between the constructed control and treatment groups in these exogenous variables.

With the observations to be used for identifying impact, a procedure must be used to determine how the matching will occur. Assuming there is common support, the simplest approach is to use the "nearest neighbor"—that is, to match each treatment observation with the control observation that has the closest score. The logic of this is that households with similar scores would have similar preprogram characteristics and would thus be comparable. To ensure that the score is sufficiently close, a caliper width can be determined before matching and matches will only occur between scores within that width. This implies that some treatment and some controls may not match with any observation and thus may not be used. Once the nearest neighbor is identified, the outcome variable is compared for each matched treatment and control observation. These differences are then averaged over the entire sample for an outcome y_i, as follows:

$$\delta = \frac{\sum_{i=1}^{NT} \left(y_{iT} - y_{iC} \right)}{NT} \tag{3}$$

where y_{iT} is the value of the impact indicator for treatment observation i

y_{iC} is the value of the impact indicator for control observation that is matched to treatment observation i

NT is the number of treatment observations $i=1...NT$

δ estimates the impact of the intervention.

As an example of this approach consider an indicator, such as expenditures per capita, which is used as an overall welfare measure. Since matching is used to create the counterfactual, if it worked as expected expenditures per capita prior to treatment would be, on average, the same for the constructed control and treatment groups. Suppose that after matching, on average, the constructed treatment and control groups both spent $100 per capita prior to treatment. After treatment, suppose the treatment group spent on average $120 per capita and the constructed

control $110 per capita. Using equation (3), the estimated impact of the program will be $10 per capita ($120-$110==$10).

As noted above, the problem with this approach is that it assumes that differences between control and treatment are observable. If there are unobservable differences then the calculation from equation (3) can capture both program effects and these preexisting differences, leading to biased estimates of impact. One solution to this problem is to use baseline data. If baseline data is available, instead of comparing the value of the indicator after treatment, the change in the indicator value can be used—that is, for outcome y_i the difference in the outcome over time $\Delta y_i = y_{i1} - y_{i0}$ where y_{i1} is the value of the indicator after treatment and the y_{i0} the value before treatment. The reasoning behind taking this approach is precisely that it helps control for pretreatment differences in the constructed treatment and control groups whether they are observable or unobservable. Following the example above, suppose that due to some unobservable differences in the constructed treatment and control groups, the control actually spent on average $105 per capita prior to the program. Equation (3) would still provide an impact of $10 even though the control group's expenditure per month only went up $5 and the treatment group's expenditure went up $20 suggesting a relative increase of $15 per capita. By looking at changes in the indicator, it is possible to remove any pretreatment differences in the constructed control and treatment. This can be seen in Table 1 below.

Table 1: Difference-in-Difference Example

	Before	**After**	**Before-after difference**
Treatment	100	120	20
Control	105	110	5
Treatment-Control difference	-5	10	15

Formally, this can be written as:

$$\delta = \frac{\sum_{i=1}^{NT} \left(\Delta y_{iT} - \Delta y_{iC} \right)}{NT} \qquad (4)$$

where $\Delta y_{iT} = y_{iT1} - y_{iT0}$ *is the change in the value of the impact indicator for treatment observation i*

$\Delta y_{iC} = y_{iC1} - y_{Ci0}$ *is the value of the impact indicator for control observation that is matched to treatment observation i*

Note that if there are no preexisting differences in the constructed treatment and control equations (3) and (4) will give the same result ($10 per capita in the above example). Thus, equation (4) is the safer, and preferred, approach since it controls for any potential preexisting differences but if there are none still gives the same estimate.

Matching to the nearest neighbor assumes that the closest match is the best and does not use information on other control observations which might be almost as similar. Instead of simply matching with the one nearest neighbor, matching treatment observations to more control observations, such as the five or ten nearest neighbors, is also possible and common. The argument for doing this is that it uses more information and provides more robust results (although there is some question about whether this is the case). As with one-to-one matching, the closest propensity scores (in absolute value) are used for comparison if within the predefined caliper width. In this case, the difference in the outcome indicators between each of the chosen control observation and the treatment observation is calculated. The overall difference between the individual treated and controls is the weighted mean of the individual differences where each "neighbor" is equally weighted. Alternatively, instead of equally weighting the nearest observations, weighting can be based on the difference between the treatment propensity score and the score of the control observation with closer scores being weighted more than further scores. In this case, the information from different control observations is used, but a greater emphasis is placed on closer observations. Different ways of calculating these weights can be used, such as those based on a kernel function. Using this scheme, the impact estimate can be written as follows:

$$\delta = \frac{\sum_{i=1}^{NT}\left(\Delta y_{iT} - \sum_{j=1}^{NC} W_{ij}\Delta y_{ijC}\right)}{NT} \quad (5)$$

where Wji is the weight placed on control observation j that is associated with the treatment observation i

$\Delta y_{ijC} = y_{ijC1} - y_{ijC0}$ *is the value of the impact indicator for control observation j that is matched to treatment observation i*

NC is the number of control observations j=1…NC matched to treatment observation i i=1…NT.

There are no clear rules about what approach or weighting scheme to use and in practice evaluators often use multiple approaches to test the robustness of the results across specifications.

Finally, note that because of the manner in which PSM estimates impact, impact estimates do not come with straightforward standard errors that can be used to test the statistical significance of results. Of course, it is desirable to obtain not only an estimate of impact, but also to determine if that estimate is significant in a statistical sense. The common approach to obtaining tests of significance is to use "bootstrapping" to calculate standard errors. The details of this are beyond the scope of the guideline, but it is important that a sense of the significance of results be obtained.

There are two other nonexperimental approaches that could potentially be used to evaluate regional development programs which we briefly describe below. They are regression discontinuity design and instrumental variables. Although these approaches might be considered for regional development programs, they are less likely to be suitable for reasons discussed below.

Regression discontinuity design uses the fact that there is some cutoff for eligibility of a program. For example, a program may target the poor and use a minimal score on a poverty index to decide who is eligible. Alternatively, the cutoff could also be a geographic limitation since programs may be only within certain local political units. Those right above and right below a cutoff point are likely to be very similar and thus those below the score may represent a good counterfactual for those above the eligibility cutoff. The approach assumes then that there

is no discontinuity in impact indicators—that is, those on either side of the cutoff would have behaved in a similar manner with or without the program or any difference that can be controlled for. If this assumption holds, it is possible to use this fact to identify the impact of the program.

The approach is particularly useful for programs where sets of individuals or households are likely to be near some cutoff. For regional development programs that target communities, there are less likely to be cutoffs or at least an insufficient number of observations on either side of a cutoff. However, there is the possibility that there are geographic limitations that create a discontinuity across political boundaries or that projects are chosen based on being above a certain expected rate of return. In such cases, using regression discontinuity approaches may be appropriate. Note that this approach depends on the design of the data collection noted above. If it is reasonable to assume a discontinuity in program design but not in outcomes, it is possible to use this method and it should be explored.

The final nonexperimental approach considered is instrumental variables. The primary benefit of an instrumental variable approach is that it relaxes the exogeneity assumption of the previous methods—that is, among other things it helps to deal with the possibility that program participants differ from nonparticipants in unobservable ways. To see this, consider a standard regression on an impact indicator y_i as follows:

$$y_i = \alpha + \partial P_i + \beta X_i + e_i \tag{6}$$

where P_i is equal to one for participant observations and zero otherwise,

X_i is a set of exogenous variables, and

e_i is the error term.

As noted above, if the data collected for the evaluation is through an experiment, and the experiment leads to a treatment and control group that are alike in all ways, δ provides an estimate of the impact of the intervention. With nonexperimental data, we are concerned that due to program placement or self-selection, those who are in the program may be fundamentally different from nonparticipants and δ might capture these differences along with program impact, leading to biased estimates of the influence of the program. If these sources of bias are observable, PSM may be used. If they are unobservable, it might be possible to use baseline data to remove any preexisting differences, assuming these differences do not change over time—that is, they are time invariant. However, in the event neither of these options works, an instrumental

variables approach may be the best option. With such an approach, an instrumental variable or set of instruments, *Z*, is required that is exogenous, matters to program participation, but does not influence the variables measuring program impact. The instrumental variables then help predict program participation, but do not suffer from the same problems as the participation variable. Under such cases, equation (6) could be estimated through a two-stage procedure as follows:

$$P_i = \sigma + \eta Z + \pi X_i + u_i \tag{7a}$$

$$y_i = \alpha + \partial \hat{P}_i + \beta X_i + e_i \tag{7b}$$

where $\hat{P}_i$ is the predicted probability of participation in the program estimated from (7a).

Since P_i is exogenous then $\hat{P}_i$ is also exogenous and should provide an unbiased estimate of impact. Because P_i takes the value of zero or one it might be estimated by a probit or logit although there are arguments for using a simple linear probability model.

While an instrumental variable approach is appealing conceptually since it appears to deal with the primary problems of identifying impact, the problem lies in coming up with instruments that meet the criteria of being exogenous, correlated with participation, and uncorrelated with the outcome variable. Such instruments are hard to come by and testing all the assumptions underlying the validity of an instrument is not possible. Arguing for the exclusion of Z variables from the estimate of program indicators is particularly difficult. In practice, an instrumental variable approach is used most often in evaluation, when there is an instrument somehow related to program design such as geography or politics of program placement or even partial treatment.

4.2.3 Evaluating mechanisms of impact, spillovers and heterogeneity

Although the focus of this section is on the approaches to identifying the impact of regional development programs on beneficiary localities, a brief discussion of the indicators to use and the manner in which to conduct the analysis is merited. In general, programs funded by development banks and other donors include logical frameworks or results matrices. If done carefully, these should include indicators not just of the overall expected impact of the program, but also the mechanism of impact. In conducting an analysis of program impact, it is important not to focus just on the final outcomes such as improved expenditures or income per capita, but

on identifying the mechanism by which this outcome occurred. By carefully selecting indicators to evaluate, it is possible to understand the process by which an impact occurred and potentially to improve this process in subsequent programs. For regional development programs, this means trying to understand how investment in locally requested projects influenced the economic decision making of beneficiaries and the effect this had. To be able to capture such effects, the investment projects must be well understood and the data collection designed in such a way as to collect the right type of information.

A second consideration is whether the program might have had indirect effects within a community or on neighboring communities. Depending on the project, these "spillover" effects can be substantial.[6] In a regional development program, if there is only a subgroup of individuals within a community that participate in a project, the spillover effects may be on their neighbors within the same community. If the investment is in training or in infrastructure, there may also be benefits to neighboring communities. To capture these effects requires identifying both direct and indirect beneficiaries of the projects prior to data collection. Data must then be collected on both of these groups as well as on the control group. The mechanism by which spillover effects are expected to occur must also be considered so data should be collected to check whether this mechanism did in fact cause the spillover effects or not.

The focus above has been on obtaining an unbiased estimate of what is referred to as the average treatment on the treated—that is, the average effect of the program on treated observations. While this is of interest, the focus on just the average effect of the program is generally insufficient and considering whether the impact differed by subpopulations is important. For example, in regional development programs certain types of investment, such as training versus infrastructure projects or agricultural versus nonagricultural projects, might be expected to have a different impact, and assessing this possibility is important. Further, certain groups within communities, such as the poorer population, may benefit more or less from the program. There may also be differences in effects across region. Having some idea of how the impact of the program might vary across subpopulations prior to data collection allows the relevant information to be obtained and the heterogeneity of the program impact to be assessed. This in turn provides insight into the manner in which future programs might be designed or existing programs modified.

[6] See Angelucci and Di Maro (2010) for a discussion incorporating spillover effects in impact evaluations.

4.3 The third hypothesis: Returns to investment

One general limitation of impact evaluation is that it tends to focus on impact indicators and says nothing about whether the investment in achieving that impact was worthwhile. This, however, does not have to be the case. Impact evaluation carefully determines the impact of a program on key indicators using a counterfactual analysis. Using the information obtained from the analysis, estimating at least some of the benefits of a program is possible. In this section, we focus on how to assess whether the cost of achieving those benefits is worthwhile. As with impact evaluation, there is an extensive literature on cost-benefit analysis (CBA), and examining that literature is beyond the scope of this guideline. In conducting a detailed analysis, this literature should be utilized.

Before discussing the details of the analysis, it should be clear that unlike standard impact assessment, CBA can be done prior to a program implementation (*ex ante*) or at the end of the program (*ex post*). *Ex ante* CBA estimates whether an investment in a program is justified based on its estimated net effect on welfare. *Ex post* analysis considers whether the returns to a program were sufficiently high to merit the prior investment (and to potentially invest in a similar program). *Ex ante* and *ex post* analysis are similar, with the primary difference being the means by which the value of benefits and costs are calculated. For *ex ante* CBA, costs are the expected or planned costs and benefits are estimates of what benefits should emerge from the program. For *ex post* CBA, costs are the actual program costs and benefits can be estimated through a well-constructed evaluation strategy. It is therefore the *ex post* analysis that is most closely linked to impact evaluation and here we focus on this type of CBA.

CBA can be done in seven steps as follows:

1. Identification of the program
2. Identification of all the social costs and benefits of the program
3. Valuation of the social costs and benefits of the program for each year of the program
4. Discounting of future values into present values
5. Assessment of the program based on investment criteria
6. Conducting of sensitivity analysis
7. Determination of whether the investment in the program was worthwhile.

Identification of the program refers to having an in-depth understanding of the program and how it was implemented. As with impact evaluation, understanding program details is critical. In the case of regional development programs, this refers to the details noted above regarding how projects were selected and administered. In this case, the information is required to identify the costs and benefits of the program rather than to identify a counterfactual. As noted in the second step, all social costs and benefits of the program should be identified. By *social* costs and benefits, we refer to all the costs and benefits to society, including the government, firms, and individuals. These costs and benefits may be financial and nonfinancial. Most costs in regional development programs are known since they are identified and valued in program documents; so both planned and actual expenditures should be available. Of course, there could be additional costs that are borne by members of society that need to be identified. For example, if a regional development program includes investment in infrastructure, such as a road, there may be costs to those living near that road. These costs should be included. As with the costs, a well-documented development program should also identify most of the benefits.

Once the costs and benefits are identified, the next step is to determine a procedure for estimating values for each of these. Impact evaluation is most useful in valuing program benefits on direct, and sometimes indirect, beneficiaries of a program. The procedures outlined above for impact evaluation provide an estimate of the average effect of the program on key indicators of program success. If the data used in the evaluation is representative of the beneficiary population, this can be used to estimate total benefits using information on the total number of beneficiaries. If these are monetized values, the average benefits can be simply multiplied by the total number of beneficiaries to get the total benefits. If they are not monetized, and include indicators such as increases in agricultural production, they need to be put in monetary terms so that the total value can be calculated. In identifying benefits, it is important to avoid double counting and ensuring that some values do not subsume others. For example, the impact on total income per capita is likely to be the result of increased agricultural production and both should not be counted. Given this is the case, for CBA, the impact of programs on overall welfare measures are the most useful for calculating benefits. For other types of benefits (and costs) of the program beyond those provided to beneficiaries, there is a large literature on methods for placing value on both market and nonmarket goods.

Since regional development programs usually take place over many years and the benefits occur over time, it is necessary to identify the costs and benefits in every period over some time horizon. Because we are interested in assessing if the program is currently worth the investment spent, future (and past) values need to be put into present-value terms. By present values, we mean the value in current monetary terms of past, present and future costs or benefits. For a number of reasons, including time preference, uncertainty, opportunity, cost of capital and so on, future values are discounted to put them into present-value terms. The appropriate rate of discounting the future is subject to debate and largely depends on assumptions about the reasoning for discounting. The key however to calculating the current value of future costs and benefits is the discount rate, which is the rate at which future monetary values are valued in the present. A higher discount rate discounts future monetary value more than a lower rate. In conducting a CBA, some rate must be decided upon and the IDB uses 12 percent.

Once benefits and costs are identified, valued and put in present-value terms, an initial assessment of the program is possible using some investment criterion. There are three principle investment criteria used to assess programs: net present value (NPV), benefit-cost ratio (BCR) and the internal rate of return (IRR). The NPV is a measure of the absolute monetary gain of a program and is calculated by subtracting the present value of costs from the present value of benefits. For a given discount rate, ∂, and costs and benefits in time $t=0...T$ indicated by C_t and B_t respectively, the NPV of a program is measured by:

$$NPV = \sum_{t=0}^{T} \frac{B_t - C_t}{(1+\partial)^t} \tag{8}$$

If this value is positive, the program is justifiable while if it is negative it is not worth the investment. The problem with this measure is that it gives an absolute monetary value with no consideration of how much was invested. Therefore, for example, two programs could have an NPV of $1 million and be viewed as equivalent on this criterion even if one only costs $2 million and the other costs $200 million.

The BCR criterion attempts to address this by using the ratio of the present value of benefits and the present value of costs as follows:

$$BCR = \frac{\sum_{t=0}^{T} \frac{B_t}{(1+\partial)^t}}{\sum_{t-0}^{t} \frac{C_t}{(1+\partial)^t}} \tag{9}$$

This provides an estimate of the return to a dollar invested. A value greater than one for the BCR implies benefits exceed costs and provides an estimate of the return to one dollar invested. A value of less than one implies each dollar invested returned less than a dollar in benefits. For this criterion, a ratio of greater than one suggests a program is worthwhile.

Along similar lines is the internal rate of return (IRR) which provides the rate of return of benefits to costs. Calculating the IRR requires finding the discount rate, ∂, for which the NPV is equal to zero. This can be done using alternative values for ∂ in equation (8) until the NPV=0. Whether the IRR is sufficiently high to justify the investment depends on whether the IRR exceeds the discount rate since if it does it implies that the investment exceeds the manner in which the future is discounted. For development programs, the tendency is to use the IRR as the investment criterion and see whether it exceeds a predetermined value, such as 12 percent. Programs that exceed this threshold are viewed favorably while those below this value are viewed as offering insufficient returns.

For regional development programs, calculating each of these measures is fairly straightforward. The challenge is in identifying and valuing the costs and benefits. Before making a recommendation on whether a program was worthwhile, the final steps require assessing the sensitivity of the analysis to key assumptions, including those made in calculating the benefits and costs. Quite often, calculating benefits and costs requires certain assumptions about the magnitude of effects—for example, the number of direct or indirect beneficiaries, the average impact on beneficiaries, price changes on products, sustainability of effects over time and so on. Checking how important these assumptions are in calculating the return to the program is important so that any recommendation can be qualified. Once this is carefully done, a recommendation on whether the program was or would be worthwhile can be made.

4.4 The fourth hypothesis: Assessment of institutional structure

Even if the rate of return is positive and sufficiently high to justify investment, there remains a question of whether the institutional structure used is the best mechanism by which to make

those investments—meaning the relative return exceeds that of alternative ways of investing. It may be that an alternative approach might have achieved the same types of investment but at a lower cost (i.e. been more cost effective); or have chosen better productive projects (i.e. had a higher rate of return); or have been more sustainable (i.e. facilitated future investment activities).

To assess whether an alternative approach might have been more effective, an alternative institutional structure must be identified. In the case of regional development programs, one alternative is to use the structure that existed prior to the program's implementation. Since regional programs by definition involve institutional transformation, the pre and post structure can be compared. Alternatively, the program's institutional structure can be compared to some other alternative structures, such as full decentralization. Of course, any of these comparisons require a degree of speculation regarding what type of results would have been obtained under this alternative structure. Unless alternatives have been tried elsewhere and evidence is available, it can be difficult to determine what benefits might have been obtained. However, going through this process of considering what an institutional structure might look like and the type of benefit it would have produced, it might become obvious that an alternative structure would have worked better. For example, a regional development program may create a complicated and costly institutional structure that identifies projects that would have almost certainly been identified even without the complicated institution—even if the project had a sufficiently high internal rate of return, it could have been higher with a less costly institutional set up.

If information is available on the costs of an alternative structure and on potential benefits, CBA could be conducted, in a manner described above, to determine which institutional structure would have achieved the highest returns. Alternatively, cost-effectiveness analysis could have been done. Cost-effectiveness analysis looks at alternatives to determine which alternative can bring about a given result in the least costly manner. Of course, the desired result needs to be defined and for regional development programs it might be increased per capita income among beneficiary households or some similar metric. In this case, cost-effectiveness analysis might examine the costs of alterative institutional structures in bringing about a certain percentage gain in per capita income.

As should be evident, the best approach to assessing institutional structure of regional development programs depends on a number of factors, especially the information that can

reasonably be made available. Equally clear should be the importance of going beyond impact evaluation to considering whether the approach used was the best possible alternative.

4.5 The fifth hypothesis: Assessing agglomeration economies

While investments may have a sufficiently high return to justify them, this does not necessarily imply that they help develop clusters of activities. For this to happen, the expectation is that there might be a pattern of public investment in related activities in certain areas (if a cluster is at its early stages of development) that will create the incentives necessary for private funding to enter the particular area and develop a cluster of activity that generates the greatest return for any particular investment. Proposals for public investment might also note this as an objective of providing funding, arguing that the benefits of locating an investment in a specific area far outweigh any added costs because of the functional linkages between activities. Thus, a first step when considering if a regional development program induced agglomeration economies is to look over the pattern of investments and the investment proposals to see if they are consistent with creating agglomeration economies. This step is, of course, linked to the first hypothesis noted above that examines the range of investment in productive activities. An assessment of private investment complementing this public investment would also suggest clusters are being formed and could be examined although this may be difficult to attribute to the program investment alone.

It may also be possible to try and test in the longer run if agglomeration economies exist. Again, this may be difficult to attribute to the program, but at least it provides a sense of whether the program investment was consistent with the creation of agglomeration economies. The details of how this can be done are beyond the scope of this guideline. However, the basic logic is to determine whether the productivity, or similar metric, is influenced by the location of firms near other similar or related firms—that is, by controlling for other factors determining whether firms obtain benefits from being in the proximity of other similar firms. In the case of regional development programs, this requires having detailed data on the firms that exist within the region of interest including spatial information on the location of the firms and the proximity to other firms. Such information is unlikely to be available unless specifically collected by a program as part of an evaluation.

5. Conclusion

Regional development programs have a certain logic in that they recognize that households and localities within a geographic area may specialize in different activities and that there may be a value in public sector investment in these activities to facilitate the development of clusters of related economic activities. However, there is an inherent challenge in supporting this type of investment in that local information and participation are critical both in identifying the best opportunities for investment and in providing complementary private sector investment. Additionally, the model used by a country to move towards decentralization can generate inequities and externalities that need to be recognized so that regional economic development can address them, rather than exacerbate them. Regional development programs, therefore, require a carefully constructed institutional structure to identify and manage investment in regional economic activities. The programs then can combine productive transformation with institutional transformation.

While based on certain logic, the programs create what can be viewed as a complicated institutional structure whose merits may raise questions. As such, these programs require careful evaluation to determine if they are effective in bringing about development in their target region. As seen in this guideline, a complete evaluation of these programs needs to move beyond standard impact evaluation to include additional analysis that assesses whether the benefits exceed costs and whether the institutional approach is the most cost effective. While possible, it requires careful planning of the evaluation strategy as noted in this guideline. The key lessons to learn in such planning are as follows:

1. As with all impact evaluations, evaluating regional development programs is much more likely to be successful if planned along with the design of the program. Only in this manner can the right type of information be collected.
2. One way to determine if the institutional structure used in the program is successful, is to carefully review the proposed and selected programs. This allows an assessment of whether clusters are being formed and if the process brought about investments that might not have been identified if the institutional structure was different.
3. The best way to conduct an evaluation of impact on beneficiaries is to initially randomly assign investment projects to both a treatment and a control group. The control group can

receive the investment later in the program, but in the meantime can act as a reasonable counterfactual to show what would have happened to the treatment group in the absence of the program. This allows an unbiased estimate of the impact of the program on the beneficiary population.

4. Beyond impact evaluation, for regional development programs it is critical to carefully consider whether the method of identifying the investment projects was the best possible approach. This requires conducting cost-benefit analysis or cost-effectiveness analysis and considering alternative institutional structure. If included as part of the program design, the more likely this is to prove a valuable exercise, since it is likely to lead to better information being collected.

References

Angelucci, Manuela and Vincenzo Di Maro. 2010. "Program Evaluation and Spillover Effects". *Impact-Evaluation Guidelines. IDB-TN-136*, Strategy Development Division. Washington, DC, United States: Inter-American Development Bank. http://www.iadb.org/document.cfm?id35173297

Bird, R. 2003. "Taxation in Latin America: Reflections on Sustainability and the Balance between Equity and Efficiency". *International Tax Program Papers 0306*. Toronto. International Tax Program, Institute for International Business, Joseph L. Rotman School of Management: University of Toronto.

Davis, B., P. Winters, G. Carletto, K. Covarrubias, E. Quiñones, A. Zezza, K. Stamoulis, C. Azzarri, and S. DiGiuseppe. 2010. "A Cross Country Comparison of Rural Income Generating Activities". *World Development* 38(1): 48-63.

FAO. 1998. *The State of Food and Agriculture 1998: Rural Non-Farm Income in Developing Countries.* Food and Agriculture Organization of the United Nations. Rome: FAO

Faguet, J., and F. Wietzke. 2006. "Social funds and decentralisation: optimal institutional design". *Public Administration and Development* 26(4): 303-315.

Hospers G., and S. Beugelsdijk. 2002. "Regional Cluster Policy: Learning by Comparing". *Kyklos* 55:381-402.

Koethenbuerger, M. 2008. "Revisiting the 'Decentralization Theorem'—On the role of externalities". *Journal of Urban Economics* 64(1): 116-122.

Krugman, P. 1991. *Geography and Trade*. Cambridge, MA, United States: MIT Press.

Oates, W. 1972. *Fiscal Federalism*. New York, United States: Harcourt Brace Jovanovich.

Oates, W. 2007. "On the Theory and Practice of Fiscal Decentralization". *CREI Working Paper no. 1/2007*. Rome: Centro di Ricerca Interdipartimentale di Economia delle Istituzioni.

Porter, M. 1998. *On Competition*. Cambridge, MA, United States: Harvard University Press.

Rawlings, L., L. Sherburne-Benz, and J. Van Domelen. 2004. *Evaluating social funds: a cross-country analysis of community investments.* Washington, DC, United States: The World Bank.

Reardon, T., J. Berdegue, and G. Escobar. 2001. "Rural Nonfarm Employment and Incomes in Latin America: Overview and Policy Implications". *World Development* 29(3): 549-560.

Stock, J., and M. Watson, 2007. *Introduction to Econometrics*. Boston: Addison Wesley.

Zezza, A., G. Carletto, B. Davis, K. Stamoulis, and P. Winters 2009. "Rural Income Generating Activities: Whatever Happened to the Institutional Vacuum? Evidence from Ghana, Guatemala, Nicaragua and Vietnam". *World Development* 37(7): 1297–1306.