From Outputs to Outcomes: A Global Review of Impact Assessment Methods in Public Development Banks

October 2023
Produced by the Montreal Group
Authored by Raphaël Cottin and Mathieu Galliot
Executive Summary

This document analyzes the different impact measurement methods used by 15 National Promotional Banks and Institutions (NPBIs) across the world (complete list on page 6).

The main findings of this report are the following:

1) All NPBIs worldwide conduct some sort of impact measurement. The trend towards impact measurement has accelerated in the past three years, following the Covid-19 pandemic.

2) There is considerable variation in the scope, approach, and frequency of impact assessment among The Montreal Group (TMG) members and other NPBIs surveyed. This diversity of methodology reflects the great heterogeneity of context and mandates among the group.

3) NPBIs focus mostly on evaluating the impact of their financing products on revenue, firm growth and employment creation. While many other outcomes indicators are measured, these represent a common ground for most institutions. Financing products usually cover loans or guaranteed loans programs. Some institutions also assess the impact of equity financing. Among those offering non-financial support, only a few assess the impact of this type of service.

4) The absence of appropriate data for a target population usually comprised of SMEs and the high cost of conducting impact assessments are the main issues faced by NPBIs. These factors drive the type of methods used and the number of assessments performed.

5) Counterfactual analysis is perceived as the gold standard by most banks. This type of analysis is implemented through the application of micro-econometric techniques on observational data. However, the complexity, the problem to access data and the time-delayed inherent to counterfactual analyses do not make it always the most appropriate tool. Randomized Controlled Trials (RCT) have been proven hard to implement, and in some cases, deemed impractical. The level of precision of the method used is variable, with most institutions aiming for an intermediate level of scientific precision.

6) The other methodologies, besides counterfactual analysis, are more eclectic: input-output analysis, and outcome measurement mostly. Some institutions use ad hoc scoring methods. Often, banks do not just use one method, but a mix of different approaches.

7) Most banks are looking at improving or developing methods to measure their impact on Greenhouse gases emission (GHG) reduction.
8) Banks using impact assessment mentioned that this method had a lot of positive outcomes and led to program improvements in several cases.

9) There is also an absence of international benchmarks to compare the performance of similar programs across countries. The development of such benchmarks could be a fruitful area of cooperation for The Montreal Group.
# Table of Contents

1. **Introduction** 5

2. **Measuring Impact: A panorama of National Promotional Banks and Institutions worldwide** 7
   2.1. All institutions members measure their impact, somehow 7
   2.2. A diversity of approaches coming from different contexts and mandates 8
   2.3. Impact on what? Outcome vs. Program coverage 11
   2.4. Challenges, limits and benefits of impact measurement 13

3. **From counterfactual analysis to scoring: no one-size-fits-all method** 17
   3.1. Ex-ante assessments: scoring and macro-level analysis 18
   3.2. Ex-post assessment, descriptive 19
   3.3. Counterfactual analysis through propensity score matching 24
   3.4. Going further than PSM: refinements and complements 29

4. **Conclusion** 32

5. **References** 33
List of Figures

1. Since when are you conducting impact measurement? (other than monitoring) 7
2. Stylized timeline of an impact assessment 24
3. The Maryland scale of scientific methods and TMG impact assessment 29

List of Tables

1. A typology of impact measurement methods 10
2. A typology of indicators for impact measurement 13

List of Boxes

1. What is impact measurement? 8
2. What is Process evaluation? 20
3. Before-after comparisons and event study methods 22
4. What is counterfactual analysis? 24
5. What is Propensity Score Matching? 26
6. Alternative methods of randomization 31
1. Introduction

National Promotional Banks and Institutions (NPBIs) play a key role in the socio-economic development of their countries and in building inclusive, prosperous, resilient, and sustainable economies. Their activity includes direct financing and indirect lending (through loan guarantees or financial intermediation), equity investing and non-financial support, targeted mainly towards SMEs. The gradual extension of the activity of NPBIs has increased the need to understand the impact of their intervention from a public policy point of view. This need has become important in the wake of the Covid-19 pandemic, which has caused a significant increase in the need for government intervention to face the downturn. Because of the fiscal effort made by governments to offer these products and the subsequent strong increase in the number of SMEs supported, the need for impact measurement has become more important.

The interactions between The Montreal Group (TMG) economists during the Covid-19 pandemic have brought to light a common need to share and exchange knowledge related to impact measurement. All participants have expressed a desire to understand the different experiences of individual members, highlight successful approaches, and understand the limits of various methods; they have a clear will to share the experiences between themselves and outside the group.

This study explores the different ways members of TMG and comparable institutions outside of the group measure the economic impact of their SME-focused operations. It presents the main methods used by NPBIs, highlights commonalities and seeks to explain differences.

About de Montreal Group

The Montreal Group is an international forum created to foster peer group exchanges, identify best practices and innovative solutions on key topic issues, and develop thought leadership material contributing to the global SME ecosystem. The Montreal Group provides a global collaborative environment where members can share and exchange best-in-class knowledge and information on important SME-related topics.

TMG members share a similar objective of supporting entrepreneurship, SMEs, and overall economic development. To a variable degree, they also tend to support regional development to ensure economic prosperity outside of large urban centres. Additionally, most institutions receive a clear mandate to help businesses that could have not received funding without them.
The Montreal Group's head office is located in Montreal, Canada. As of October 2023, members of The Montreal Group are:

- **Banco Nacional de Desenvolvimento Econômico e Social (BNDES)** - Brazil
- **Bpifrance** - France
- **British Business Bank (BBB)** - United Kingdom
- **Business Development Bank of Canada (BDC)** - Canada
- **China Development Bank (CDB)** - China
- **Development Bank of Nigeria (DBN)** - Nigeria
- **Finnvera** - Finland
- **Nacional Financiera (Nafin)** - Mexico
- **Saudi Industrial Development Fund (SIDF)** - Saudi Arabia
- **Small Industries Development Bank of India (SIDBI)** - India
- **Tamwilcom** - Morocco
- **Wallonie Entrepndre** - Belgium

**Methodology**

This study is based on interviews conducted between April and June 2023 with economists from the 12 public institutions that are members of The Montreal Group.

Three NPBIs, not members of TMG, were also interviewed:

- **Cassa Depositi e Prestiti (CDP)** - Italy
- **Instituto de Credito Oficial (ICO)** - Spain
- **Japan Finance Corporation (JFC)** - Japan
These in-person interviews were complemented by an examination of the documents produced by TMG members and a review of secondary literature.

This report is organized as follows:

- Section 2 is a high-level overview of the impact measurement practices of TMG members, the relation to their mandates and the common challenges faced by all institutions.
- Section 3 is a deep dive into the methodologies of impact measurement, highlighting the pros and cons of each method and critically assessing their uses.

Disclaimer: The view and opinions expressed in this study are those of the authors and do not necessarily reflect the views or positions of TMG members.


### 2.1. All institutions members measure their impact, somehow

**A recent push towards impact measurement**

Most of the institutions interviewed have started recently to conduct impact measurement. Only one third of TMG members have been doing impact evaluation for more than five years (Figure 1). This reflects in part the relatively recent creation of these institutions. But this seems mostly to be the result of an increase in prominence in impact evaluation methods: in many cases the impetus for conducting impact measurement comes from an executive or strategic decision, and everyone would endorse the importance of the process.

*Figure 1: Since when are you conducting impact measurement? (Other than monitoring)*

- More than five years: 33%
- Less than five years: 67%
Box 1: What is impact measurement?

For development banks, impact measurement is about estimating the effect that they have on the society in which they operate.

The words “impact” and “evaluation” can have different meanings according to the organizational culture and the academic discipline (economics, public administration...).

This report uses “impact measurement” as an umbrella term covering all the methods used by our respondents to assess the effect of their financing and non-financing SME programs on these businesses and the economy. The method can be ex-ante or ex-post, micro or macro.

“Impact assessment” is generally used to refer to methods that seek to isolate the causal effect of a program. Nevertheless, the notion of impact measurement is often broader. Some methods presented in the report measure an “economic footprint” while others estimate an “economic impact.” They are all part of an overall impact measurement process, which takes different forms and includes a range of indicators, from monitoring disbursements to attributing specific effects through statistical analysis.

Reasons for conducting impact measurement

TMG members do not measure their impact for the same reasons:

- Most institutions are looking to know ex-post if their products made a difference for their clients in order to understand the efficiency of a whole program and assess the need to continue or redesign their offer. The practice is mostly requested informally, although for some institutions impact evaluation is a regulatory requirement.
- Other institutions are using impact estimation to determine eligibility for their funding. For example, SIDF uses a scoring system to evaluate the potential effect of a project before granting funding. This system is a great way to facilitate decision-making when choosing between several projects. This approach can be qualified as an ex-ante approach, embedding a forecasting approach to help decision-making on a project-by-project basis or to assess an expected macroeconomic impact of a program.

2.2. A diversity of approaches coming from different contexts and mandates

A typology of impact measurement methods

There is considerable variation in the breadth, depth, and frequency of impact assessment among TMG members. Table 1 summarizes the broad families of methods used by the respondents.
The first axis of differentiation between methods is whether the impact measurement takes place before or after the intervention: ex-ante vs. ex-post measurement. Ex-ante methods do not require measurement of outcomes after the program has taken place; they require only minimal information from the companies affected by an intervention. They may rely on expert modelling of the economy and the interaction between sectors. In contrast, ex-post measurement does not necessitate a full model of the economy to deliver results but relies on extensive information on outcomes by participant firms (and, depending on the precise method, non-participants as well).

Ex-ante methods can be further subdivided into scoring methods for individual interventions, such as the one implemented by SIDF, and macro-methods inspired by the input-output model. The choice between one or the other depends on the objective of the analyst: scoring methods are used as an input in decision-making, while input-output modelling is more useful at a strategic level and for reporting purposes.

Ex-post impact measurement comes in a variety of forms. We distinguish between three broad families of ex-post methods: program monitoring, outcome evaluation, and counterfactual analysis. Program monitoring is concerned with how and for whom the program funds are disbursed. Outcome evaluation, by contrast, focuses on the economic results achieved by the firms and sectors targeted by the intervention, such as employment, turnover, growth, etc. Finally, the counterfactual analysis seeks to tease out the causal impact of the program on the beneficiaries, net of windfall effects and crowding-out mechanisms (see part 3 for more detail).

The choice of a particular method within ex-post studies is mostly governed by a trade-off between the expected timing of the assessment, the precision of the results, the availability of data and the amount of resources devoted. As we move further down the table, methods can deliver more precise estimates of the banks’ impacts but require more resources and data to be performed. For instance, counterfactual analysis requires, at a minimum, data on outcomes for participants as well as non-participants, which is not easy to come by. Furthermore, this type of analysis takes time to perform. In comparison, performance indicators or outcome evaluation can deliver results relatively quickly, allowing for rapid feedback into operations.

As evidenced by table 1, there is a cumulative nature in ex-post evaluations. Institutions conducting outcome evaluation generally perform program monitoring as well, and those who do counterfactual analysis also publish indicators related to the first two methods of ex-post impact evaluation. The reason for this is that the indicators of the various families are, for practical purposes, nested within each other: when performing counterfactual evaluation, indicators pertaining to outcome evaluation become readily available as a by-product.
### Table 1: A typology of impact measurement methods

<table>
<thead>
<tr>
<th>Timing</th>
<th>Type of impact measurement</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex-ante or Ex-post</td>
<td>Scoring</td>
<td>BNDES, JFC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SIDF</td>
</tr>
<tr>
<td></td>
<td>Macro modelling</td>
<td>BBB, BNDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CDP, ICO</td>
</tr>
<tr>
<td>Ex-post</td>
<td>Program monitoring</td>
<td>CDP, SIDF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAFIN, Wallonie Entreprendre + all institutions below</td>
</tr>
<tr>
<td></td>
<td>Outcome evaluation</td>
<td>DBN, ICO</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JFC, Tamwilcom + all institutions below</td>
</tr>
<tr>
<td></td>
<td>Counterfactual analysis</td>
<td>BBB, BDC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bpifrance, BNDES, Finnvera</td>
</tr>
</tbody>
</table>

**Differences in the mandates and structure of the banks influence the practice of impact measurement**

TMG members have different experiences with impact measurement, which can be traced back to variations in the economic and political environment. These differences are influenced by TMG members' legal status, obligations toward their government, and the range of products they offer.

Overall, institutions providing a significant number of subsidies either directly or through low-interest loans tend to be more scrutinized and are more likely to develop an impact measurement process. On the other hand, for institutions that provide funding at or above market interest rates, the need for impact measurement, while still relevant, is perceived as less pressing.

The nature of the funding also plays a vital role in governing the extent of impact assessment. Institutions relying on state resources or external funding for their operations face higher demands for impact indicators from their stakeholders. For instance, the fact that Bpifrance operates programs that are backed by national budget resources (subsidies,
guarantees) was instrumental in the decision to put in place impact indicators from its onset. Similarly, institutions receiving funds from international organizations, like DBN, are also expected to provide impact analysis.

Some institutions issue specific debt instruments such as Green Bonds or Social Bonds for part of their funding. These instruments typically entail reporting obligations on the use of proceeds or the general impact of the financing. For example, advances in evaluation analysis (estimation of employment via input-output, CO2 emissions avoided...) were very useful in fulfilling reporting requirements in ICO’s social and green bonds.

The size of the bank, in terms of assets or the number of clients, or the diversity of product offerings, also influences the level of precision and the type of impact measurement conducted. Larger organizations have more easily resources to devote to an internal team dedicated towards impact measurement. In some countries, the various functions of NPBIs (loan guarantee, export credit, venture capital....) are split between various institutions. This fragmentation may create coordination costs that make impact measurement more difficult.

2.3. Impact on what? Outcome vs. Program coverage

Because of this diversity of situations, institutions have different priorities and look at different indicators when measuring their impact. These indicators can be grouped into two categories: program coverage indicators and outcome indicators.

Indicators of Program coverage: sectors and firms supported

The first category of indicators used to assess the impact of NPBIs is related to program coverage. These measure whether the institution reaches the strategic goals on which it has direct control. This can include the number of SMEs supported, or the number of loans disbursed and the corresponding amount.

Many TMG members also have other implicit objectives in terms of geographic rebalancing: it is frequent for them to report KPIs by region or geographic area. Some institutions have explicit goals in terms of support of strategic sectors (e.g., manufacturing). In these cases, TMG members look at very specific indicators: clients in tech sectors, clients owned by women, etc.

Outcome Indicators: firm revenue and employment growth

Outcome indicators measure whether the institutions reach their ultimate goals in terms of the effects expected from their interventions. This can include a contribution to value added or job creation but can also extend to effects further removed in the causal chain such as innovation.
The most common metric used to gauge the impact is the firm turnover. For institutions conducting equity investment, it is common to look at the valuation of supported firms, as the revenue is generally not significant for early-stage, high-growth firms.

Employment growth is also a very common metric followed by all TMG members. Estimates of jobs created or supported are heavily used in political instances as it is easy to communicate and relatable to the population.

However, depending on the economic context, job creation is not always a priority for TMG members. This is the case for BDC which does not value employment creation as much as in the past because of labour shortages. It shifted its focus towards productivity growth and impact on minority groups. SIDF also echoed the concern.

One indicator that has an intermediate status is the knock-on effect on the amount of funds raised. When the goal of the institution is to act as a catalyst for other private investment, the amount raised by the fund or the firm with other investors is of interest per se.

**New concerns and new indicators**

Regardless of the type of indicators used, institutions regularly adjust them to better fit their current objectives. For instance, many TMG members are looking at adding environmental indicators to track their contributions to GHG reductions.

Other indicators used by TMG members include:
- Number of SMEs supported
- Characteristics of entrepreneurs supported (women, minority groups, nationals, etc.)
- Number of patent applications from clients
- Private capital leveraged
- Impact on government revenue
- Contribution to exports
- Economic diversification
- Productivity growth
- Etc.
Table 2: List of indicators used by participants

<table>
<thead>
<tr>
<th>Program coverage indicators</th>
<th>Outcome indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Number of clients (direct, indirect, by program)</td>
<td>- Revenue growth</td>
</tr>
<tr>
<td>- Percentage of businesses (or SMEs) that are clients of the institution</td>
<td>- Gross value added</td>
</tr>
<tr>
<td>- Characteristics of firms supported (size, location, sector)</td>
<td>- Firm valuation</td>
</tr>
<tr>
<td></td>
<td>- Profitability</td>
</tr>
<tr>
<td></td>
<td>- SMEs’ access to financing</td>
</tr>
<tr>
<td></td>
<td>- Firms’ survival rate</td>
</tr>
<tr>
<td></td>
<td>- Clients taking action to reduce their environmental footprint</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MSMEs</th>
<th>Institutional</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Private capital leveraged (can be considered outcome indicator, as well)</td>
<td>- Level of business investment</td>
</tr>
<tr>
<td>- Share of credit from the institution going to SMEs</td>
<td>- Government revenue</td>
</tr>
<tr>
<td>- Amount lent to businesses (by program, etc.)</td>
<td>- Economic diversification</td>
</tr>
<tr>
<td></td>
<td>- Investors' confidence in investing in SMEs</td>
</tr>
<tr>
<td></td>
<td>- Financial institutions' willingness to lend to SMEs</td>
</tr>
<tr>
<td></td>
<td>- Products or services introduced by financial institutions to support SMEs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Community</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Characteristics of entrepreneurs supported (women, minorities, nationals)</td>
<td>- Employment growth</td>
</tr>
<tr>
<td>- Indicators of additionality (e.g., % of clients who would have not received financing without the institution)</td>
<td>- Quality of jobs</td>
</tr>
<tr>
<td></td>
<td>- Labour productivity</td>
</tr>
<tr>
<td></td>
<td>- Capital productivity</td>
</tr>
<tr>
<td></td>
<td>- Export growth</td>
</tr>
<tr>
<td></td>
<td>- Greenhouse gas emissions (GHG)</td>
</tr>
<tr>
<td></td>
<td>- Reduction of GHG emissions</td>
</tr>
<tr>
<td></td>
<td>- Local content in the production</td>
</tr>
<tr>
<td></td>
<td>- Product complexity</td>
</tr>
<tr>
<td></td>
<td>- Level of innovation</td>
</tr>
</tbody>
</table>

2.4. Challenges, limits and benefits of impact measurement

Irrespective of the method used, the program being evaluated, and the objective of the impact measurement, there are some commonalities in the challenges faced by NPBIs. In this section, we highlight two high-level challenges: getting and using reliable data and communicating the results. We also highlight the benefits mentioned by the respondents of conducting impact measurement, and the outlook for further development.

Good Data is hard to come by

Regardless of the type of analysis conducted, accessing data is a common challenge among all TMG members. The data on counterparts may be difficult to collect and access for several reasons, and the quality may be subpar.

Outcome evaluation requires information on the characteristics of the company under consideration. Some of this information, such as the size of the balance sheet, the turnover and the number of employees, may be routinely collected as part of the commercial
relationship between the bank and the firm. More advanced forms of assessment such as counterfactual analysis often require additional data that are less common (e.g., innovation or greenhouse gas emissions).

Data about firm characteristics and outcomes is typically collected when a commercial relation is initiated or renewed between the development bank and the client, but it may be challenging to get comparable data in the time interval between contacts. Several correspondents echoed the idea that “the day when the loan is signed is the last when we hear from the firm.”

In many cases, measuring impact was not a priority at the creation of the institution which means that many programs were not built to be evaluated ex-post. For example, BDC collects data on its clients mostly for operational purposes but not for statistical impact measurement.

Compliance burden on businesses can also be an important issue. While a detailed database leads to better impact analysis, complying with mandatory data collection can be burdensome for businesses. Too much paperwork during loan application, tax reporting or mandatory surveys from statistical agencies can have negative impacts on economic growth. Demanding too much data from clients can also be seen as red tape by entrepreneurs and deteriorate the business-friendly image of the institution. A balance must be found between the quality of the data collected and the burden imposed on businesses.

Many institutions operate indirectly, through third-party providers, or funds, at least for some part of their activity. This poses an additional challenge for data collection. Portfolio data can be filled partially by account managers for whom it is not a priority. The process is generally not automated and implies back-and-forth with numerous partners, and ex-post harmonization of data, which is time-consuming.

The data collected may not be reliable. Lack of reliability can come from different sources. Self-reported data provided by clients either during loan application or through surveys, may be subject to desirability biases. Additionally, data collected through online surveys generally has a low response rate; this is a problem if the response is not random, as it may introduce a selection bias in the analysis.

Collecting and accessing data are two different steps in the process. Even when extensive data is collected on counterparts by the operations in the banks, it does not follow automatically that the impact analyst will have access to it. Confidentiality issues may preclude the wide sharing of company data; in any case, an internal mobilization exercise is necessary to access and be able to work on such data.

Third-party data providers may be mobilized to avoid the problems of data collection. However, this does not completely assuage the problems linked to insufficient coverage or
data quality. The best practice is to use data collected by national statistical institutions, but this data does not always exist or is not always accessible to external researchers (see part 3.2 for more details on this).

The problem is especially important for institutions in Africa. Both Tamwilcom and DBN mentioned it as an important challenge. DBN struggles to conduct online surveys among micro-businesses which do not always use the internet on a regular basis. Additionally, in-person surveys are not always feasible as they are very costly because of their duration and their labour-intensive nature. Because DBN provides indirect support, entrepreneurs do not always know the institution and can also be reluctant to give information. Another example of ad hoc data collection is Finnvera, which teams up with another Finnish institution to conduct a survey of SMEs for the purposes of impact measurement, pending access to confidential information from a national register.

**Communicating evaluation results**

Communicating the results of an impact study can be challenging but is one of the keys to maximizing its usefulness.

Some TMG members mentioned that some impact measurement reports were not comprehensible by all readers because of their complexity. Differences between terminologies are not always clear for all audiences, such as jobs supported vs. jobs created. While authors of impact reports will differentiate them, some other stakeholders might use them interchangeably. In this context, the added value of a complex counterfactual analysis over an I-O model or an economic footprint is not always perceived by decision-makers.

Finally, while economists and academics are interested in developing complex impact studies, other stakeholders do not always see what the benefit is. Several institutions mentioned that they needed to do important work of convincing before being able to go further into impact analysis.

Complex impact studies do not always have a lot of readers among key decision-makers. Ideally, measuring the impact of a program leads to better fund allocation and program designs. In practice, there is not always a clear connection. This lack of interest can be explained by the backward-looking nature of most impact studies while decision-makers are forward-looking. Additionally, most teams in charge of measuring the impact are not involved in the process of designing new programs.

Some impact assessments, while able to estimate a causal effect of the program on outcomes, do not provide a clear explanation of the mechanisms behind the results. This is problematic because it can give the impression that impact assessment is a “black box” and is damaging for the credibility of the exercise.
Recently, when the United Kingdom left the European Union, several programs active in the UK but funded by the EU came to an end. An evaluation of a specific EU program administered by The British Business Bank made it possible to safeguard its funding by the UK.

An impact study from the Development Bank of Nigeria helped identify issues with the length of the loans offered as many entrepreneurs wanted longer loans with more flexibility.

An evaluation from Tamwilcom showed that their loan guarantee program helped reduce the collateral required by banks.

Bpifrance and BDC also demonstrated that their activities had a positive impact on the growth of SMEs which reinforced the political support for the institution.

Finally, there is also a lack of international benchmarks to compare the performance of development banks. For instance, BNDES can find that one of its programs has a positive economic impact but cannot say if this impact is on par with what could be expected from such a program across the world. Because of differences in terms of programs, objectives, databases and methods, it is not possible to determine if an institution performs better than another one in terms of outcomes.

**Despite challenges, SMEs and public institutions benefit from impact measurement**

Impact studies can bring positive change despite not being used at their full potential.

Several institutions that have been doing impact measurement for several years have successfully used the results to improve their practices and show their contribution to the economy. Impact studies are especially useful during periods of budget allocation, at the institution or the government level.

Regardless of the methods used, impact studies often lead to further strategic discussion among decision-makers and in some cases helped improve some program designs.

- Recently, when the United Kingdom left the European Union, several programs active in the UK but funded by the EU came to an end. An evaluation of a specific EU program administered by The British Business Bank made it possible to safeguard its funding by the UK.
- An impact study from the Development Bank of Nigeria helped identify issues with the length of the loans offered as many entrepreneurs wanted longer loans with more flexibility.
- An evaluation from Tamwilcom showed that their loan guarantee program helped reduce the collateral required by banks.
- Bpifrance and BDC also demonstrated that their activities had a positive impact on the growth of SMEs which reinforced the political support for the institution.

When used to their full potential, impact assessments have several benefits for the government, the institution carrying it and the SMEs.

**Benefits for the government:**

- Better information to design public policy
- More transparent public spending
- Better budget allocation
Benefits for the institution:

- Better program designs
- Better understanding of their clients
- Can help identify a market gap

Benefits for SMEs:

- Better programs benefit SMEs
- More funding can be attributed if a program is proven to work well

**Most respondents are planning to do more to measure their impact**

Many banks are looking at improving their method. SIDF is planning to move from a scoring model to a computable general equilibrium model. Other banks are looking at implementing some type of counterfactual analysis in the future. Others are looking at improving their current counterfactual analysis method or experimenting with RCT.

Additionally, evaluating contribution to GHG reduction is on the radar of many banks across the world. Some are already doing it in some aspect, but most are looking at developing methods to do so. This seems to be one of the next big challenges for development banks.

**3. From counterfactual analysis to scoring: no one-size-fits-all method**

TMG members use a variety of methods to assess the impact of their interventions, reflecting the diversity of their mandates and organization, as documented in part 2. The precise choice of the impact assessment method also depends on external constraints (notably the availability of good quality firm data) and strategic choices.

This section provides a “deep dive” into the various methods used by TMG members and partner institutions. We use the same typology of methods used in part 1: ex-ante vs. ex-post methods; among ex-post methods, we distinguish between descriptive methods and counterfactual methods.

Within each class of method, there is some variability in the implementation between different institutions. We highlight the main sources of differences and, when possible, try to highlight what we perceive as being the “best practice” within each class of method.

The main findings of this section are the following:

1) There is no “one-size-fits-all” methodology for doing impact assessment. The choice of the type of impact assessment is primarily driven by data availability.
2) Counterfactual analysis is used by about half of TMG members and is judged very satisfactorily by its practitioners. It strikes a good balance between objectivity, credibility, and communicability of the results.

3) Most institutions conducting counterfactual analysis are doing it at an intermediate level of robustness, as defined by the Maryland Scientific Method Scale.

4) The main drawbacks of counterfactual analysis are the high data requirement and the time lag between the intervention and the results. The additionality of the intervention is often presumed rather than proven.

5) Descriptive statistics come in two forms: process evaluation and dashboards of outcome variables. They are used when the possibility of doing a more thorough impact assessment is not possible, but also in complement to counterfactual analysis.

6) Data issues are common to all types of methods. Common problems include the proper identification of beneficiaries and the difficulty of accessing outcome data for indirect clients.

7) Ad-hoc methods such as qualitative interviews and scoring methods are used in the absence of representative outcome data.

8) Randomized controlled trials are not used routinely for impact assessment within TMG members. The previous experiences of RCTs have proven disappointing.

3.1. Ex-ante assessments: scoring and macro-level analysis

Scoring methods

Two institutions mention the use of “scoring models”, but the term describes two different practices. In the case of JFC, “scoring” is a qualitative assessment of the objectives of the different branches and sub-divisions of the corporation. The dimensions of the assessment are related to the operational functioning of the units and their adequacy to the overall organization objectives: “business operation plans” and “organizational plans”.

In the case of SIDF, the scoring method refers to an ex-ante evaluation model that feeds directly into the decision process. The goal of this exercise is to value the supposed positive social and economic contribution of the SMEs directly in the loan generation process. As part of the application process, firms self-report on a certain number of dimensions, which are then confronted with a pre-existing scoring matrix. The output is a score that will be taken into account for the acceptance of a project. The dimensions considered in this scoring process are the following: gross value added, capital productivity, labour productivity, product complexity, local content, export promotion, and indirect impact. Their goal is to capture the additional value added of an investment in the Saudi economy, in terms of broader economic development.

A detailed examination of these methods lies outside the scope of this paper. The methods are highly institution-specific and make use of criteria that cannot be easily transposed to other institutions. In the case of SIDF, the precise criteria, weights and scores are proprietary and confidential. They constitute frontier cases between monitoring and evaluation, and decision-making.
Ex-ante macro methods

Several institutions use ex-ante macro assessments to estimate their impact on production and employment. For instance, ICO employs this type of analysis in the framework of the report linked to the issue of a social bond.

Ex-ante macro assessments can take various forms. Within the institutions interviewed, the most common approach is a variation based on an analysis using Input-Output tables. These tables are derived from national accounts and provide information for each sector and final use, including the intermediate consumption of other products.

By manipulating input-output tables, analysts can determine the induced and indirect effects of each additional € of production from a sector, considering the entire value chain of that sector. When combined with macro estimates of the additional production per € financed and the number of jobs per additional unit of production, this type of analysis can deliver direct, indirect, and induced impacts in terms of both job creation and additional production resulting from financing provided by NPBIs.

This general methodology can be refined further to integrate more precise factors. For example, ICO separately computes the induced effects for productive investment and the financing of working capital, which requires additional assumptions and data. However, it allows for a more realistic estimation of the institution's impact.

In general, Input-Output assessments are more precise than analyses based on economic footprint. They require less data from counterparts than descriptive statistics on outcomes. However, the method does involve manipulating national accounts and several macro parameters, such as the incremental capital-output ratio or the marginal propensity to consume, which require input from specialists. Once these data and parameters are known, relatively little input is required from the bank.

The ability to deliver estimates of indirect effects is also valuable to analysts and the institution's stakeholders. On the other hand, this type of analysis has several limitations that should be clear in the minds of its users. The results obtained by manipulating input-output tables are of a short-term nature. Moreover, in this case, the additionality of the bank is assumed rather than measured (as in any ex-ante assessment).

3.2. Ex-post assessment, descriptive

Process evaluations allow to get rapid feedback on the operational process

Although generally not considered to be a proper causal impact assessment, several institutions use process evaluation as part of their monitoring and evaluation process. Process evaluation is concerned with the implementation of a program (as opposed to the ultimate outcomes) and does not entail the construction of a counterfactual.
Process evaluation is a structured way to conduct client satisfaction, for the purpose of rapid feedback on the operation of a program. It seeks to answer the question, “Is the process for delivering funding (or any other intervention) going as planned”. It is best used as an early stage of the operation of a program, as a way to get an early warning when a problem occurs. In our sample, BBB and DBN have mentioned conducting this type of assessment.

The main sources of information for process evaluation are questionnaires sent to the beneficiaries of the program. The themes covered include the ease of accessing the funds, the process of applying for the program and the appropriate targeting of the program. The output of this kind of assessment includes profiling of the beneficiaries: type of firm (size, age, sector of operation), geographic dispersion of the beneficiaries, and statistics on the identity of the founder. For instance, BBB publishes statistics and the number of beneficiaries located outside London as well as the share of beneficiaries by gender and ethnic minority status. DBN also presents the share of program recipients located outside Lagos (the most important city) and the share of program recipients who are youth or women.

Some institutions try to gather information on non-beneficiaries as part of the process evaluation: either on firms that have applied to the scheme but have been denied access or on firms that could have been eligible but who have declined to apply. The goal of this step is to gain insight into the reasons that deterred target firms from applying to a particular program and identify the presence of roadblocks in the application process, in order to feed back quickly into operations. Getting answers from non-beneficiaries may prove challenging, as the response rate is likely to be low.

**Box 2: What is Process evaluation?**

According to the US CDC, process evaluation “determines whether program activities have been implemented as intended and resulted in certain outcomes”. Closely related to program monitoring, process evaluation is concerned with “left-side” variables in the logic model, in contrast to outcome evaluation which deals with the right side (the ultimate outcomes that the program addresses). Process evaluation shows the following elements:

- How well the program is working
- The extent to which the program is being implemented as designed
- Whether the program is accessible and acceptable to its target population.

It is useful mainly for its ability to provide an early warning for any problems that may occur. The World Bank classified process evaluation as a complementary approach to impact assessment. They are not suited to answer cause-and-effect questions but can be conducted quickly and at a reasonable cost. Indeed, they complement rather than replace
impact assessment, and may sometimes be sufficient for operational purposes: “Applying an impact evaluation to a program whose operational processes have not been validated poses a risk that either the impact evaluation resources are misspent when a simpler process evaluation would have been sufficient, or that needed adjustments in program design are introduced once the impact evaluation is underway, thereby changing the nature”.

References: adapted from
Gertler et al. (2016), pp. 13-18

Outcome evaluation: descriptive statistics, dashboards, and before-after analyses

Several institutions collect and report descriptive statistics of outcome variables among beneficiaries. This type of analysis is conducted by 4 institutions in our sample and is the second most popular technique for assessing impact.

Descriptive statistics frequently come in the form of a “dashboard” of indicators that detail a list of outcomes year after year. The outcome variables are usually those mentioned in part 1, with special attention to jobs created and turnover growth in supported businesses. These descriptive statistics are frequently disaggregated by the gender of the owner of the business, less often by the ethnic minority status (for BBB), and its youth status (for DBN). Often, the statistical dashboard includes a breakdown of beneficiaries by sector of activity and by region, to address concerns about spatial equity and territorial development that are often part of the mandates of public development banks.

Descriptive statistics may be standardized, and their production automated to a large extent. The teams responsible can produce the corresponding quickly and according to a predefined template, which allows the reader to grasp the main message in a single glance. The standardized nature also allows to chart the evolutions of these descriptive indicators year after year and to spot quickly big changes in outcome variables or in the environment. The source of the information for this type of analysis is nationally representative surveys, commercially available data, or surveys specially commissioned by the institution. This is the case of DBN or Finnvera for instance.

This approach (descriptive statistics) cannot be used to infer causality. The number of jobs in supported businesses cannot be attributed to the supporting institution unless a specific causal claim is made, which is generally not advisable. Hence, it is not possible to infer value for money or to conduct a cost-benefit analysis based on descriptive statistics alone. This approach is not suited to answer questions related to the good use of public funds or the efficiency of the policy.

Some commercial institutions (viz. Crédit Mutual Arkéa in France, Banking for Impact) are increasingly publishing the metrics mentioned above, as part of their ESG communication.
This is both an opportunity and a challenge for public development banks. Public institutions should not be judged by the same standards as private institutions; there is a risk of undermining the legitimacy of public development banks if they are judged by the same metrics as private banks.

Some institutions (viz. Tamwilcom) show descriptive statistics of outcome variables at several points in time (Before-After study). Before-after comparison is not enough to reach causal conclusions on the effect of a program, except under specific circumstances. An increase in the outcome variable is often judged as a “first line of evidence” in the absence of better knowledge. The drawback of before-after comparisons is that they require more data, compared to simple descriptive statistics. Indeed, if the data is sufficiently good to perform good quality before-after analysis, it may be good enough to be leveraged for a difference-in-difference design (i.e., counterfactual analysis), which may explain why relatively few of the participating institutions are using before-after analysis to assess impact.

Box 3: Before-After comparisons and event study methods

Before-After comparisons are generally not sufficient to conclude a positive impact of a program. Beneficiary firms may self-select into the treatment, and the outcome (employment, firm growth) may have increased without the help of the public development bank (selection bias). There may also be other reasons that explain why the firm’s outcome has progressed the way it has, which may have nothing to do with the bank’s intervention (confounding factors).

In the absence of comparison group data, it may be possible to recover causal impacts from only the treated units, if there is significant variation in the timing of the treatment among the treated units and if the precise timing of the treatment is “as good as random”: this is known as an “event study design” and is increasingly popular in academic economics (Currie et al., 2020). In these kinds of studies, the units that are treated later constitute a counterfactual for the units that are treated at an earlier date, and vice versa. Nevertheless, the data requirements and validity conditions are rather more stringent than in the case of simple difference-in-difference estimations, as several periods of data before and after the treatment are required (Miller, 2023). An example of this kind of research design implemented by a TMG member is BBB’s Enterprise capital funds interim evaluation (BBB, 2021), which conducts a “pipeline design” assessment of the impact of a specific type of funds on the outside investment attracted by subscribing firms.

Qualitative methods and mixed methods

Outside of quantitative methods, several institutions conduct qualitative interviews with beneficiaries as part of their monitoring and evaluation process. For instance, DBN conducts 20 interviews with beneficiaries SMEs as part of their “spot-checks”, which complements the SME survey conducted remotely. BBB also uses qualitative interviews in its evaluations.
Qualitative interviews are mainly used to illustrate the results of the quantitative impact. Verbatims are frequently interspersed in the main text, as well as photographs.

The goal of qualitative methods is not to establish causal impacts. The sample is too narrow and too selected to allow for generalization of the results, and the answers suffer from self-serving and desirability bias. Nevertheless, qualitative interviews have their place in the process of impact assessment, notably to help uncover the mechanisms through which the program has an impact. Statistical methods and counterfactual analysis are sometimes unable to explain why some effects took place. In this case, qualitative interviews can be a good additive as they allow institutions to ask directly to beneficiaries what the mechanism is.

Another case where qualitative interviews can be useful is in complement to the process evaluation to get feedback on the design and parameters of a specific program. The questionnaire administered for the process evaluation consists of closed questions; open-ended questions and semi-structured interviews with beneficiaries may be more successful in highlighting roadblocks in the process. One example of this is DBN, which provides loans to the private education sector. The reimbursement dates were not in sync with the academic year, which constituted a problem with the delivery partner. Qualitative interviews helped identify this problem and led to the creation of a specific product for the education sector with increased delays for repayment.

A long time lag before the results are available

A general point about the ex-post assessment based on outcome data is the long time lag before data is available. At the minimum 3 years, but most frequently 4 to 5 years is common between the publication of the impact assessment and the start of the program. The reasons for such delay are depicted in Figure 3. An incompressible delay of one to two years is necessary to let the program deploy itself. Some effects, especially in the real economy, need time to materialize; conducting an evaluation too soon after the start of the program would risk finding null results simply because the effects have not yet materialized. Second, nationally representative data on firms are usually released with a one- to two-year delay by the national statistical institutions. Finally, the analysis itself takes time: about one year from the pre-analysis to the first results.

This time lag between the start of the program and the impact assessment is unavoidable but threatens the relevance of the evaluation process. A common criticism by stakeholders is that the results of the counterfactual analysis are outdated: the context may have changed, or the design of the program may have been modified. There is thus a trade-off between the robustness of the impact assessment and its practical usefulness for policymaking.

To remedy this problem, some institutions (Bpifrance and BDC for instance) re-evaluate the programs at regular intervals (5 years approximately) to ascertain that the empirical results
Another solution is to supplement the impact analysis on final outcomes with an evaluation of intermediate outcomes that are available more quickly after the start of the program. BBB, through its interim reports, has institutionalized this practice.

**Figure 2: Stylized timeline of an impact assessment**

<table>
<thead>
<tr>
<th>Year</th>
<th>Launch of the program</th>
<th>Effects appear</th>
<th>Data on year 2 is available</th>
<th>Impact assessment is ready</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.3. Counterfactual analysis through propensity score matching

**Counterfactual analysis is the most used method for ex-post purposes**

Counterfactual analysis is the most used method for assessing the impact of public bank interventions. Out of 12 institutions interviewed, 5 are conducting some form of counterfactual analysis of their interventions, with two more exploring the possibility of implementing this method in the future.

The institutions using this kind of impact assessment are generally satisfied with it. No one mentioned replacing it or conducting a major overhaul. At most, it will be supplemented by more refined methods.

Counterfactual analysis is the method of choice for banks seeking to assess the additionality of their action or to address concerns about the good use of public funds and public “value for money.” The output of counterfactual analysis can also be used as a building block for cost-efficiency or cost-benefit analysis.

**Box 4: What is counterfactual analysis?**

The observation that SMEs helped by a development bank has better outcomes after the intervention of the bank is not enough to conclude to a positive impact of the intervention; some of this improvement may be due to other causes, such as the general improvement of the economy. A simple comparison of treated and untreated firms is also not sufficient: the observation that treated firms do better may reflect a selection bias (the assisted firms were going to perform better anyway) rather than a true impact of the intervention. Hence the necessity to construct a counterfactual (“what would have happened to the client firms, had they not received assistance of the bank”).
In general, it is not possible to directly observe the counterfactual outcomes; it must be reconstructed through the careful application of econometric techniques of on micro-data. The output of this type of analysis is a parameter known as the “average effect of the treatment on the treated” (ATT), from which an “average treatment effect” (ATE) may be deduced. This parameter has then to be combined with other assumptions (notably on the cost of public funds and the valuation of benefits), to lead to cost effectiveness and cost-benefit analysis.

Counterfactual analysis is the appropriate method when the main concern regarding the data is one about additionality of the public intervention or appropriate use of public funds, although for the latter questions assumptions are required.

Reference

A good balance between the ease of implementation and robustness of the results

Counterfactual analysis delivers robust results that seem congruent with the experience of practitioners. The method is transparent, and its general principles are communicable to the decision-maker. This method has also the advantage of supplying a common vocabulary to collaborate with academia and specialist consultancies. Thus, it lies at a “sweet spot” between divergent requirements: to be scientifically credible and to be communicable.

Counterfactual analysis is well established in academic and policy circles, which implies that technical resources to help practitioners with the implementation are forthcoming. Among TMG members, this method is sometimes implemented in-house and sometimes outsourced to specialists. It is sophisticated enough to be presented in academic conferences and sometimes published in peer-reviewed publications, yet the results may be communicated to non-economists.

The method also fits naturally into the data collection and analysis process. Once the infrastructure for comparing internal data to nationally representative statistics is in place, it follows almost immediately to try to adjust based on covariates between one’s own portfolio and the general universe of firms, which in turn leads to the estimation of average treatment effects.

Propensity score matching is the preferred method for constructing the counterfactual

There are several possible methods to construct a counterfactual, which entail various degrees of scientific credibility. There is no one-size-fits-all method: some advanced techniques are only applicable in specific conditions. In TMG members, the preferred method is propensity score matching (PSM), which is used by all institutions conducting counterfactual analysis, alone or in conjunction with other methods.
One advantage of PSM is its ability to be applied in the general population and in a wide variety of contexts. Some methods that have a higher level of scientific credibility, such as instrumental variable estimation or regression discontinuity, may only be applicable if some conditions regarding the program are met[1], and deliver results that concern only a subsample of the target population. By contrast, PSM matching can be applied to a wide variety of different contexts, as long as the same variables are available for treated units and the comparison group, and its results are readily generalizable to the general population.

Propensity score matching has a middling ranking in the hierarchy of scientific methods. The Maryland Scale of Scientific Methods (Maryland SMS) gives it a level 2 out of 5 (Figure 4). However, it is possible to reach a higher level of credibility when the method is used in conjunction with a before-after comparison. This technique, known as “difference-in-difference propensity score matching” (PSM-DiD for short) allows controlling for one important source of bias time-invariant unobserved heterogeneity. The Maryland SMS gives it a level of 3, and several institutions have been able to implement PSM-DiD over the years. The drawback is a higher data requirement than simple PSM.

Box 5: What is Propensity Score Matching?

The principle of propensity score matching

Propensity score matching is a statistical technique by which the counterfactual is constructed as the firm that is the most similar (in a statistical sense) to each treated unit. Formally, the probability to be treated is estimated on the entire sample (treated and non-treated) based on observable characteristics by help of a binary outcome model (logit or probit). A “propensity score” is then attributed to each observation, defined as the probability for one observation to be treated. The treatment effect for a particular observation is the difference in the outcome between one treated firm and the untreated firm (or firms) that has the closest propensity score.

Propensity score matching is a more flexible way to assign a counterfactual than linear regression. The flexibility of PSM has its drawbacks: it relies crucially on the “common support” assumption to be valid: there have to be similar units in the treated and in the untreated groups, which is not always the case. This is an additional requirement compared to linear regression.

Difference-in-difference PSM and unobserved heterogeneity

The matching of treated firms to the comparison group is done based on observable characteristics, but this may not be enough to warrant a causal interpretation of the difference in outcomes between the treatment and comparison group. The firms receiving assistance from the banks might differ from the others in aspects that are not available for
the observer to see: motivation of the founder, technical know-how, growth perspectives... these unobserved factors might be what is driving the difference between treated and untreated units; without controlling for them, the difference in outcomes between treated and untreated might be falsely attributed to the effect of the bank's intervention. The process of differencing (taking the difference between units before and after the treatment) allows for the “purging” of these unobserved factors that do not change with time. This requires information on outcome variables for the treatment and the comparison group before and after the intervention.

A method that makes use of high-quality data that may not be available to all

To perform counterfactual analysis, it is necessary to have detailed information on economic and financial variables outcomes for the clients as well as for a pool of comparable firms. The availability of such data is not always under the control of the institution.

Contrary to a simple “before-after,” counterfactual analysis typically requires data on nonparticipant firms: ideally, a firm census or at least a representative sample. The same information on outcomes as well as control variables should be available for the treated firms and for the comparison group. Ideally, this information should be collected in the same way for all firms regardless of treatment status, so as to minimize the risk of bias due to differences in the data collection process.

In practice, the TMG members rely on data collected by the national statistical authorities: firm census or nationally representative surveys. The only institution conducting its own representative survey of SMEs is Finnvera. This survey is done in collaboration with another institution in order to share the cost. This situation is only an interim solution, pending access to the public registry of company balance sheets.

The necessity to use nationally representative firm data is an important barrier to the conduct of counterfactual analysis. For some countries in our sample, such data may simply not exist. In other cases, the data exists but is not accessible to external researchers, or only under very stringent conditions. These restrictions to access are justified by the need to preserve statistical confidentiality, but they represent a barrier to the conduct of impact assessment.

Sometimes, there may be practical workarounds, such as in Canada where the statistical authority (Statistics Canada) is commissioned to conduct the analysis on behalf of BDC. Such solutions may not be generalizable to every institution in our sample. Thus, access to good quality data is likely to remain an impediment to the conduct of counterfactual analysis.
A common challenge: identifying a relevant comparison group

Several participants mentioned a practical difficulty in the construction of a comparison group: it is not easy to ascertain that firms in the comparison group do not receive any form of public support. In most of the countries under consideration, there are multiple actors providing support to SMEs. To assess the impact of the program by one institution, it is necessary to be able to identify which firms received other types of support (e.g., subsidies, tax credits, state help). If this information is not available, there is a risk of getting biased results from counterfactual analysis. If the units in the control group receive some support as well (“contamination” of the control group), the effect estimated by counterfactual analysis might be biased downwards. On the other hand, if firms in the treatment group receive support from other sources, the estimated effect will be biased upwards. The issues compound when companies in the treatment and in the control group both receive other forms of support.

Moreover, constructing an adequate control group is not always feasible. Finnvera also mentioned it as an important challenge as its economy is relatively small without a lot of comparable businesses.

Among the interviewed institutions, BBB, BDC, Bpifrance and Finnvera mentioned some variation around this problem, although the specifics are country-dependent. For Finnvera, it is anticipated that the problem will be solved in the future through access to a government registry of state help; such a registry is not yet available for other countries in the sample.

Cost considerations

The process of getting, cleaning and readying the data is time-intensive (especially when portfolio data has to be matched to external data). Selecting the relevant covariates and the precise implementation of the statistical method relies a lot on expert knowledge that requires extensive involvement with specialist scientific literature. All these tasks require important time commitment and specialized know-how: all the institutions interviewed employ university-trained economists, at the master’s level at least. This implies that the cost of an impact assessment made using counterfactual analysis is probably higher than any other sort of method. The effort may be judged not to be worth it for some institutions.
3.4. Going further than PSM: refinements and complements

Strict causality and additionality require additional steps

The rationale for doing counterfactual analysis is the goal of being able to prove the additionality or the banks’ actions. Unfortunately, simple PSM is no longer regarded as sufficient to prove the causality of a policy intervention. The main issue is that analogous to linear regression, PSM relies on the assumption that every single factor governing the choice to participate in the program and related to the outcome is both observable and accessible to the analyst. There are no unobservable factors correlated with the outcome that influence selection in the program. This is a strong assumption, that is typically not testable: it has to be accepted as such. The selection on observable assumption explains why the use of PSM has faded in the academic literature (Currie, Kleven, and Zwiers 2020), and its middling place in the Maryland scientific methods scale.

TMG economists are all aware of this limitation. BBB follows an “eclectic” approach, using practical workarounds to assess additionality. First, BBB systematically collects data on self-reported additionality: clients are invited to report whether their investment would have taken place without BBB intervention. Second, BBB performs a “sensitivity check” by arbitrarily adding plus or minus 20% to the size of the impact, allowing for overestimation of the positive impact of one intervention.
This limitation in establishing causality does not seem to represent a fatal flaw for TMG members. Among the institutions that used it, no one mentioned replacing it in the foreseeable future or even making significant modifications to its implementation. One important argument for the use of PSM is the sentiment that it allows to “go as far as possible” with existing data. All interviewees from institutions using PSM echoed the feeling, giving spontaneously some variation over the same argument: “We do as much as we can with the data that we've given.”

On occasions, some TMG have been able to implement more advanced techniques that come closer to establishing causal impact. Bpifrance has been able to leverage a source of (exogenous) discontinuity in the rules of distribution to construct an instrument for its guarantee program. BNDES has been able to implement a more advanced version of the difference-and-difference estimator, leveraging variations in the timing of the treatment among units (beyond simply before after), known as the Callaway & Sant’anna estimator. BNDES has also been able to implement a shift-share IV design. These methods would reach level 4 on the Maryland scale. However, these efforts remain the exception and are harder to use systematically, as they rely crucially on features of the program design that are somewhat ad hoc and may not be present in every case.

The Conspicuous Absence of RCTs

The academic literature considers randomized control trials (henceforth RCTs) to be the “gold standard” for impact evaluation. It allows the analyst to control for some of the most frequent biases that affect observational studies: selection bias, observed and unobserved confounders, and reverse causality.

RCTs are not currently used by TMG members, although some members have the ambition to explore the possibility in the future. Two TMG members have had experience with RCT in the past: DBN and BNDES. In both cases, the results were not considered satisfactory, hence the decision not to renew the experience.

The main reason mentioned for this is the perception that the conduct of an RCT is not compatible with the mode of operation of a development bank. In some cases, such as Finland, there are legal barriers to conducting randomized trials. In other cases, the practice of excluding some potential clients from the benefit of a program would be perceived as undercutting the commercial operation of the bank. The operational challenge for most members is to reach as many firms as possible; randomizing access to the program would seem to contradict that goal. This is all the more the case that in contrast to a typical RCT (health, education and social assistance programs in developing countries), financing or equity programs by public development banks are not massively oversubscribed, hence the difficulty of conducting an oversubscription design. Alternative methods of randomization seem not to have been explored by TMG members.
**Box 6: Alternative Methods of Randomization**

Randomizing access to a program is the standard procedure for an RCT, but it is not the only one. If simple randomization is not feasible, alternative methods can be considered:

- **Oversubscription methods:** in this approach, more participants are recruited than can be accommodated in the program. Then, a random selection process is used to determine which participants will receive access to the program, while the remaining individuals serve as a control group.

- **Randomization of the order or phase-in design:** randomization of the order or phase-in design is employed when it is not feasible or ethical to deny access to the program to all participants at the same time. In this method, the rollout of the program occurs in stages, and the order in which individuals or groups gain access to the intervention is determined randomly. The later treated as used as controls for the earlier treatment.

- **Within-group randomization:** within-group randomization is a technique used when it is more practical to randomize individuals within specific groups, clusters, or geographical areas, rather than randomizing individuals independently. This approach is often chosen to avoid contamination between individuals within the same group, especially in community-based interventions.

- **Encouragement design:** in this approach, participants are not directly randomized to the intervention or control group. Instead, the intervention consists of a randomized encouragement to self-select into the treatment. This encouragement can be used as an instrumental variable to evaluate the effect of the treatment.

**Reference**

*Adapted from Duflo, Glennerster & Kremer (2007).*
Conclusion

Overall, NPBI s have a lot in common when it comes to impact measurement. While they use a wide range of indicators, there is some common ground around GDP and employment creation indicators. Additionally, they face similar challenges and adapt their methodologies to their environment.

This report reinforces the importance for more collaboration between NPBI s. Two areas for additional research/collaboration can be identified:

There is a need for more international benchmarking. Often, governments look at what is being done elsewhere before creating a new program. Better benchmarking could help NPBI s compare the impact of different programs across different countries.

There is also a need for more work on environmental indicators. With climate change, most NPBI s are getting more involved in developing programs to help lower GHG emissions. However, measuring impact on overall carbon emissions remains challenging.


THE MONTREAL GROUP

A global forum for micro, small and medium-sized enterprise development

www.themontrealgroup.org
secretariat@themontrealgroup.org