Guide to the Assessment of Socio-Environmental Impact

for Use in Impact-Oriented Projects and Investments

Insper METRICIS

Núcleo de Medição para Investimentos de Impacto Socioambiental

About Insper Metricis

Insper Metricis is a research center focused on studies of organizational strategies and management practices for projects with the potential to generate high socio-environmental impact. Special emphasis is given to the development of tools for planning, executing and evaluating impact projects performed by companies, nonprofit organizations and governments. With management and assessment procedures on the potential contribution of socio-environmental projects, it is possible not only to expand the reach of impact-oriented investments, but also to develop new ways to fund and assess such projects. Furthermore, the lessons learned from these experiences enable the continuous documentation and dissemination of best practices through academic research, policy reports, case studies and management guides related to high-impact projects.

About Insper

Insper is an independent, nonprofit institution dedicated to education and research in the fields of Business Administration, Economics, Law and Engineering. Its mission is to be a leading center and to explore the complementarities of these fields of knowledge. Insper's teaching activities offer programs for various stages of a career: Undergraduate (Business Administration, Economics, Law, Computer Science and Engineering), Graduate and Research Degree (Certificates, MBAs, Law, Professional Masters and PhD) and Executive Education (Open-Enrollment programs and Custom programs for companies). In knowledge generation, Insper works through endowed chairs and research centers, which bring together researchers to work on studies and projects focusing on public policy, finance, and management. Insper also has centers to foster entrepreneurship (CEMP) and education (Education Center). Insper's academic quality is certified by AACSB, AMBA and Anamba. Insper Metricis is part of the Center for Public Policy and Administration.

São Paulo
Fifth edition, May of 2022

General guide focused on monioring and verification of additionality

The first version of this document was written by Sergio G. Lazzarini, Leandro S. Pongeluppe, Pui Shen Yoong, and Nobuiuki Costa Ito. This new revised edition was elaborated by Sergio G. Lazzarini, José Geraldo Setter Filho, Carolina Pedrosa Gomes de Melo, Jorge Norio Rezende Ikawa, Octavio Augusto Darcie de Barros and Carolina Pellegrino Castejon, and received additional input from Lígia Vasconcellos. Previous editions benefited from the contributions of Amanda Arabage, Sandro Cabral, Sergio Firpo, Pedro Machado de Godoy, Guilherme Lichand, Carolina Pedrosa Gomes de Melo, Marina Ribeiro, José Geraldo Setter Filho, Mariana Suplicy, Carlos Kazunari Takahashi, and Rafael Vivolo. and, as well as suggestions by Fernando Carnaúba, Ben Carpenter, Amanda Feldman, Naercio Menezes Filho, Ricardo Paes de Barros, Luis Fernando Guedes Pinto, Brian Trelstad, and Maurício Voivodic. Initial discussions to develop these guidelines evolved from the work of a discussion group including Raquel Costa, Célia Cruz, Tatiana Fonseca, Frederik Kuonen, Simon Locher, Angélica Rontondaro, Franco Veludo, and Rafaella Ziegert, and with the financial support of the Latin America Economy Impact Innovations Fund (Rockefeller Foundation, Avina, and Omidyar) in the proposal coordinated by ICE (Célia Cruz and Maria Amélia Sampaio).

Introduction to the guidelines

Several private, public, and nonprofit organizations have tried to assess whether their projects generate consistent and sizeable improvements in the lives of their target populations. In several cases, the sponsors of those projects as well as their beneficiaries would like to know if the proposed interventions generate benefits as a way to reinforce successful initiatives and discontinue projects lacking evidence of positive impact.

There is also a growing concern that organizations rushing to demonstrate their alignment with socio-environmental practices (as in the recent "ESG" trend) might misreport their actual impact and present superficial analyses to advertise social benefits that can be misleading (a problem that is known as "greenwashing" or "impact washing").

Even in cases where organizations select and follow appropriate metrics of socio-environmental performance, confusion often emerges with respect to whether the intervention effectively caused the claimed improvements in the target population. In this respect, it is also important to distinguish between the **monitoring** and the **impact evaluation** of a project.

Project managers usually collect data tracking outcomes on the target population and monitor those outcomes over time. They may establish targets for those metrics and then adopt corrective measures when actual outcomes are below the established targets. For instance, in a project to increase the wellbeing of poor rural communities, project managers can follow metrics of income, education, health, and several others.

However, to effectively assess whether the project *caused* changes in those metrics, managers would need to evaluate the impact of the project by answering the following question: "What would have happened to the targeted communities or individuals had they *not* been exposed to the project?"

This question is important because there may have been simultaneous, external improvements affecting the target population that are independent of the project. For instance, in an initiative to provide public schools with technological learning tools, the project manager may incorrectly conclude that the tools have improved student learning, when, in reality, the improvement may have been caused by changes in the school system's pedagogical practices as a whole.

The so-called verification of **additionality** in social projects attempts to avoid precisely this kind of erroneous judgment. In this approach, impact is defined by the difference between the outcome observed for the individual who participated in the project and the outcome that would have been observed for the same individual, in the same time period, had the intervention not been implemented – the so-called **counterfactual**.

Therefore, ideally, we would need to compare the *same* person or group receiving a certain intervention. In some cases, this is possible. For instance, the effect of wearing glasses on the reading capacity of visually impaired people can be instantaneously gauged, based on a simple comparison of how much the person can see with and without glasses.

In most cases, however, the intervention outcomes occur over an extended period of time and can be affected by several factors that are not directly related to the focal project. In such cases, the question "What would have happened to the targeted communities or individuals had they *not been* exposed to the project?" can be answered based on information from people who, simultaneously with the project beneficiaries, did *not* receive the intervention. Nonetheless, one should consider an evaluation strategy that allows for controlling for confounding factors that may affect both the treatment assignment and the outcomes of interest.

In most cases, the intervention outcomes occur over an extended period of time and can be affected by several factors that are not directly related to the focal project.

For example, in health studies, the **treated group** (usually chosen at random) receives a drug or medical treatment, while the **control group** does not receive the treatment. This allows for measuring how the medical intervention affected the population by accounting for natural changes that might have otherwise occurred. Because of random assignment, the control group functions as a counterfactual at the group level. When a randomized experiment cannot be implemented, other measurement strategies based on additionality can be applied to mimic the counterfactual as well as possible.

Figure 1 exemplifies this type of measurement approach based on additionality. Imagine that a company wants to invest in the development of several poor communities and to assess the impact of those investments. As such, the company defines communities who will be affected by the project (the treated group) and communities who will not receive the interventions (the control group). Before starting the project, those two groups exhibit distinct income levels, but they have similar growth patterns with respect to this outcome metric.

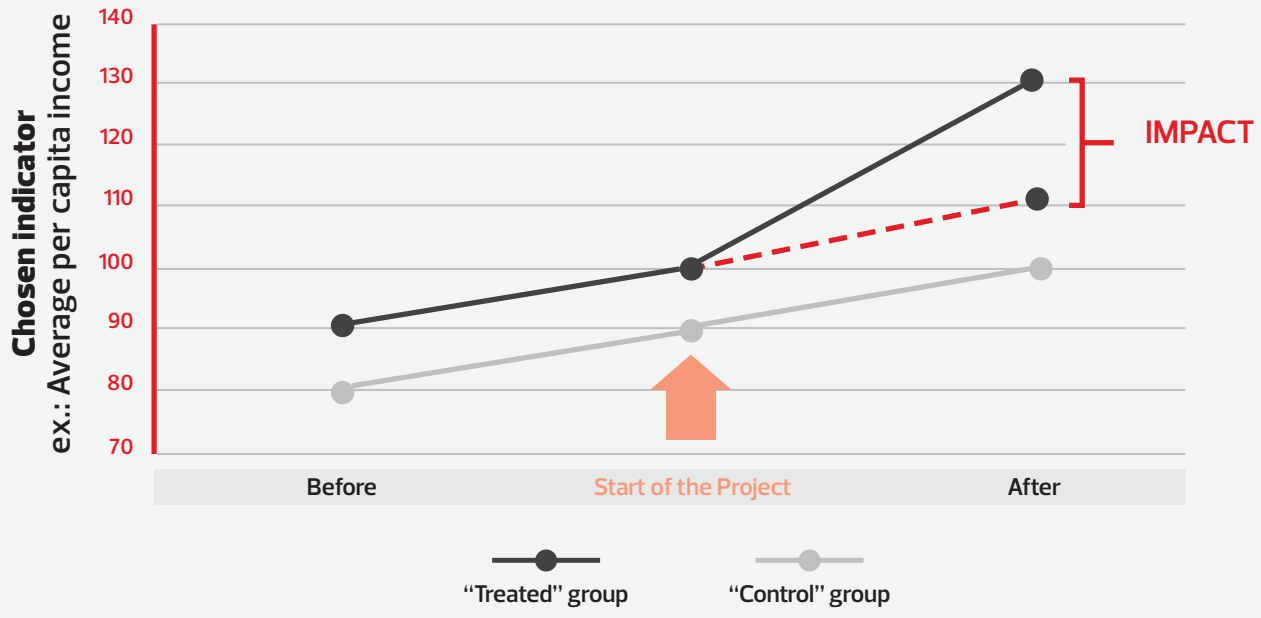
At the project's start, the company measures the income level of individuals in the control group and the treated group. At this time, the average income of the control group is $90, while the average income of the treated group (which will receive the investment) is $100.

After making these measurements, the company then begins the project with the community. After one year, it measures the average income of the treated and control groups again. An inexperienced analyst might conclude that the impact of the project is $30, since the average income of the treated group increased from $100 to $130. However, using the verification of additionality, the treated group's evolution is compared to what would probably have happened without the project. This counterfactual scenario, in this case, is indicated by what did happen in the control group.

Due to other factors that may have affected the community (such as a natural evolution in living standards or government cash transfers), the group that did not receive the investment also experienced an income increase of $10 per individual. Measuring the program's real impact is possible if we do not limit the analysis to changes over time in the treated group, but also account for other natural trends that might occur (as evidenced by the control group).

In this case, it is clear that the program increased the average per capita income of individuals in the treated community by $20. This additional gain is simply the difference between the verified change in the treated community ($30) and the verified change in the control community ($10), which serves as an estimate of the counterfactual scenario (what would have happened to the target population without the project).

Figure 1 — Impact Assessment through Verification of Additionality



Following this discussion, the present Guide provides a structured set of steps to develop measurement plans with two main objectives:

1. Propose highly relevant metrics for assessing potential improvements that are valued by the project's target population.

2. Assess what would have happened to the target population without the investment (usually through comparisons with similar control groups not included in the project).

While objective (1) is generally used for monitoring purposes, objective (2) goes a step beyond and seeks to evaluate the impact that the project caused.

In this context, the *Guide to Assessment of Socio-Environmental Impact for Use in Impact Investments and Social Enterprises* provides a practical tool for impact investors, nonprofit organizations, companies, and governments to monitor and evaluate outcomes that they potentially generate through their sponsored projects.

The Guide proposes a sequence of steps, beginning from the definition of the project objectives and its target population, followed by a sequence of tools to select appropriate metrics and find alternative ways to assess the impact caused by the project (its additionality).

The Guide is a useful tool for the practice of impact management. For instance, following the approach proposed by the Impact Management Project (IMP)[1], the Guide provides a focused discussion on the dimension "**contribution**", that is, the extent to which outcomes were distinct from what would have happened without the project. However, aligned with the general IMP framework, the Guide also discusses **who** the potential beneficiaries are (definition of the target population), **what** the most relevant outcomes are (definition of the theory of change and selection of metrics), and **how much** the project can improve the wellbeing of the target population (based on results from previous studies and the expected outcomes of the planned interventions).

The Guide is a useful tool for the practice of impact management.

[1] See: Impact Management Project. (2022a). *Mainstreaming the practice of impact management*. Retrieved from https://impactmanagementproject.com/ (accessed on March 9, 2022).

Preparing a measurement plan

1. DESCRIPTION OF THE PROJECT AND ITS BROADER OBJECTIVES
2. DEFINING THE TARGET POPULATION
3. BENCHMARKING
4. DEFINING THE THEORY OF CHANGE
5. DEFINITION OF METRICS
6. DEFINITION OF THE METHOD FOR VERIFYING ADDITIONALITY
7. SAMPLING PLAN
8. MEASUREMENT TIMELINE

Preparing a measurement plan

Preparing a measurement plan is an essential step in properly setting up, operating, implementing and controlling measurement. To implement robust and accurate data measurements, the following steps should be considered. Steps 1 to 5 guide the creation of monitoring scorecards comprised of appropriate metrics of socio-environment performance. Steps 6 to 8, in turn, involve the evaluation of the project's impact (i.e., its additionality).

1. DESCRIPTION OF THE PROJECT AND ITS BROADER OBJECTIVES

The strategy to define how to estimate the impact of an intervention begins with a preliminary description of the organization and the planned interventions. At this step, the organization's purpose and the general objectives of its proposed projects should be described and discussed.

In impact management, it has been customary to link the project's overall objectives to the United Nations' **Sustainable Development Goals (SDGs)**, which are increasingly being adopted as key fundamental issues that must be addressed to improve the wellbeing of populations worldwide.[2] It is possible that the same project may be related to several SDGs. For instance, consider a project to increase the access of poor communities to clean water and sanitation infrastructure. This project is closely related to SDG 6: "Ensure availability and sustainable management of water and sanitation for all." However, at the same time, the project can positively affect the target population's health and promote sustainable local infrastructure, thus creating additional links to SDGs 3 and 11, respectively.

Often, even when the proposed interventions are very well defined by their managers, the objectives are poorly discussed and detailed. It is also common to find projects involving myriad objectives and interventions whose execution requires capabilities that the organization lacks. Therefore, deepening the discussion of project objectives can also help managers understand which of these objectives are directly influenced by the program, and which, although desirable and related to the implemented actions, are beyond the program's scope.

2. DEFINING THE TARGET POPULATION

A key step in defining a measurement plan is to determine the **target population** of the project. In order to objectively demarcate the target population, it is necessary to start by identifying all relevant actors that can be affected by the project or can influence its results (the so-called **stakeholders**). These actors can include project sponsors, service providers, support organizations and in particular the final beneficiaries of the project. At this step, it is crucial to assess if their priorities and objectives are in line with the project's planned actions.

[2] See: UNITED NATIONS. Sustainable development goals. Retrieved from http://www.un.org/sustainabledevelopment/sustainable-development- goals/ (accessed on March 9, 2022).

As organizations have limited resources, targeting is also an opportunity to focus on who really needs help, or to emphasize subgroups for which the intervention can be most effective.

For example, a project may seek to improve the health conditions of vulnerable populations by increasing their access to medicine and effective methods for treating chronic diseases at a lower cost. A critical question then is: Who are those vulnerable populations and which individuals should be given priority, especially in conditions involving scarce resources?

Factors that can be considered when defining the target population include: [3]

- **Geography:** the project's regional boundaries.
- **Demographic characteristics:** age, sex, gender, and others.
- **Socioeconomic characteristics:** income, schooling, occupation, and others.
- **Special needs:** physical and cognitive disabilities and other relevant vulnerabilities.
- **Susceptibility to environmental factors:** adverse climate conditions, risk of disasters, and so forth.
- **Initial conditions:** the current status of the focused group in terms of key variables of interest.

Box 1 illustrates how to define the target population considering some of the dimensions above.

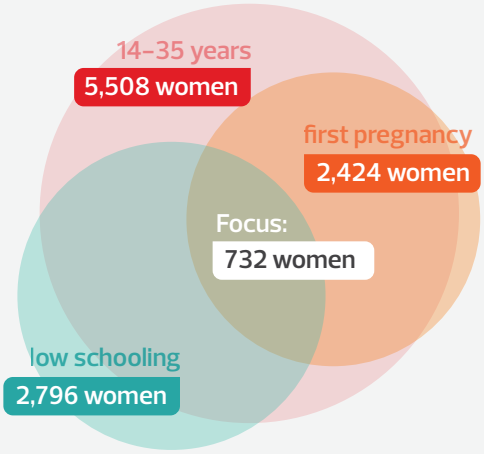
BOX 1. Example of definition of the target population in the context of a prenatal program for pregnant women.

A government program in partnership with a nonprofit organization seeks to promote prenatal consultations with the objective of increasing maternal and newborn health indicators. The government defined a set of municipalities where 6,024 pregnant women could be reached.

Previous studies have indicated that the efficacy of these programs increases in the case of vulnerable women aged between 14 and 35 years (*demographic characteristic*), with up to 7 years of schooling (*socioeconomic characteristic*), and in their first pregnancy (*initial condition*).

The Venn diagram below shows the number of women with each of these characteristics and the total number of cases at their intersection: 732 women. These pregnant women could be considered as a priority group in the proposed interventions, even though other segments could also be included in the program if there are sufficient resources.

Total number of pregnant women in the municipalities of the program: 6,024



[3] For more details, see: Impact Management Project. (2022b). *Who.* Retrieved from https://impactmanagementproject.com/impact-management/impact-management-norms/who/ (accessed on July 3, 2019).

3. BENCHMARKING

The expansion of impact-oriented projects depends on continuous learning based on the mistakes and successes of similar projects, as well as on drawing on prior measurements generated by previous research. Therefore, *benchmarking* is an essential step to identify what has been done or studied before in the domain of the project's target activities and populations.

Useful information sources include academic studies, publications by international organizations, reports from other sponsors who invested in the same activity, and several other sources. There are also various internet platforms that consolidate research and evaluation studies in several areas, with detailed information on the potential interventions and their effectiveness.[4] In impact assessment studies, it is becoming increasingly common to examine the results of *meta-analyses*: studies that seek to compile the results of various previous assessments focused on a given topic.

For example, imagine a manager interested in helping unemployed people to find job opportunities (thus, connecting with SDG 8). A meta-analysis published in 2008 analyzed 207 studies of active labor market programs, including training and job intermediation, as well as subsidies for companies.[5] The study concludes that the results observed after one year of intervention are higher than programs focused on the short run (up to one year). Comparing the types of intervention, programs with an emphasis on training seem to be more effective than programs focused only on job searching.

The benchmarking step also enables the quantification of the expected effect size of the project. For instance, in the previously cited meta-analysis, the authors conclude that training programs tend to increase the likelihood of employment by up to 8.7 percentage points in the long run. This effect is higher than what is observed after one year (1.6 percentage point), which is in itself an indication that it is desirable to assess the program's impact over a longer period of time. If the analysis of previous programs suggests that the effects are negligible, then project managers can also try to change or complement the proposed interventions.

It is also possible that the effect of the project will be higher or lower depending on heterogeneous traits of the population, as discussed in Step 2. Still using the example of job market interventions, project managers may try to focus on low-income young adults who recently graduated from high schools. The benchmarking should then search for studies that specifically focus on this target population.

[4] For instance, the library of evaluation studies compiled by J-PAL (J-PAL. (2022). *Abdul Latif Jameel Poverty Action Lab*. Retrieved from https://www.povertyactionlab.org/); the education studies toolkit developed by the Education Endowment Foundation (Education Endowment Foundation. (2022). *Teaching and Learning Toolkit*. Retrieved from https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit) and the health studies reviewed by Cochrane (Cochrane. (2022). *Our Evidence*. Retrieved from https://www.cochrane.org/evidence) (all accessed on March 9, 2022).

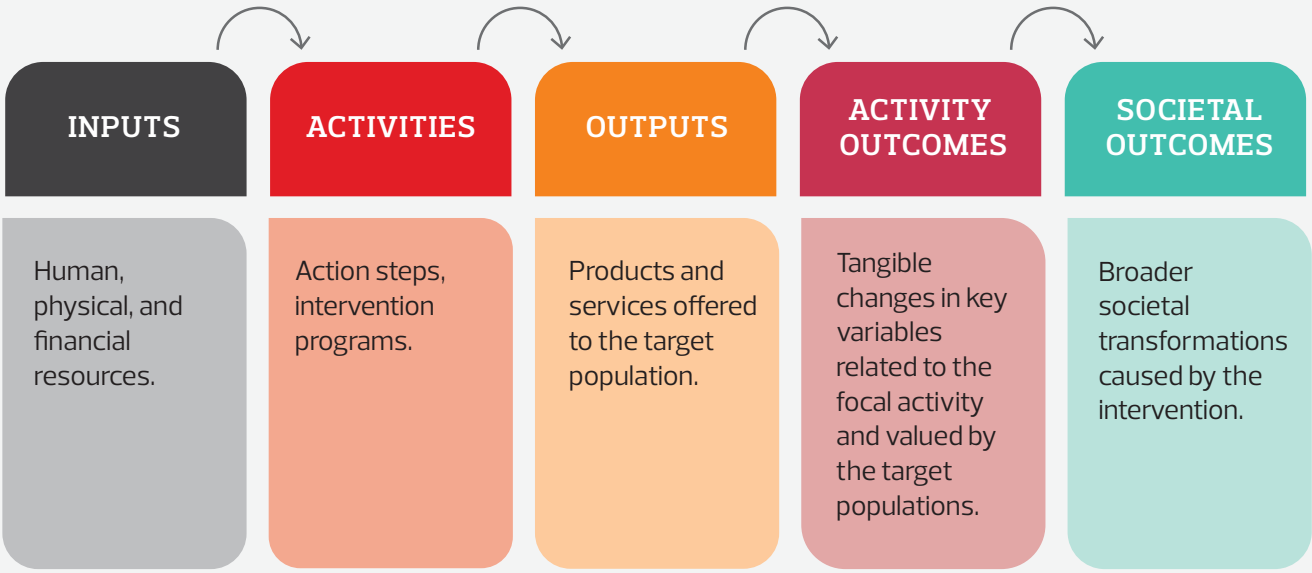
[5] See: Card, D., Kluve, J., & Weber, A. (2018). What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. *Journal of the European Economic Association*, 16(3), 894–931. https://doi.org/10.1093/jeea/jvx028

4. DEFINING THE THEORY OF CHANGE

Before delving into the issue of how to measure the project's impact, it is necessary to define what to measure. At this stage, project managers should develop their **theory of change:** a clear and logical way of connecting the proposed interventions to the expected socio-environmental outcomes.

Figure 2 shows the five stages of a theory of change, whose mapping is essential for any measurement plan: **inputs** and **activities**, which generate certain immediate **outputs** offered to the target population, which, in turn, trigger relevant **outcomes**, linked either with the project's main activities or with broader societal transformations. This analysis can greatly benefit from the previous steps of the measurement plan. For instance, benchmarking helps identify what potential inputs and activities can be more effective in promoting these results.

Figure 2 – **Theory of Change Applied to Impact-Oriented Investments**



Source: Adapted from McLaughlin & Jordan (2004)[6]

Project managers usually report outputs rather than outcomes. To illustrate this, consider a company whose main activity is to provide technological tools that improve the academic performance of public-school students, using various inputs such as online platforms and support teams coordinating the interventions. The outputs of this activity are the number of students who accessed those tools or the number of interactive sessions with students. Notice, however, that these outputs do not necessarily indicate if students really have improved as a result of the intervention. We thus need to examine outcomes, i.e., relevant changes that improved the lives of the target populations or the broader society.

[6] See: Mclaughlin, J. A., & jordan, G. B. (2015). Using logic models. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 62–87). Jossey-Bass. Also see: Hehenberger, L., Harling, A.-M., & Scholten, P. (2013). *A Practical Guide to Measuring and Managing Impact*. Retrieved from https://www.oltreventure.com/wp-content/uploads/2015/05/EVPA_A_Practical_Guide_to_Measuring_and_-Managing_Impact_final.pdf.

In this case, an expected activity outcome is improvement in students' grades in a particular standardized test. The societal outcome of the project, in turn, would be the future wellbeing and even future income gain that students will be able to reap as a result of their added skills. In this process, it is important to describe and justify the **causal mechanism** that will influence the desired change. For instance, the technological tools could include games and videos potentially increasing students' motivation to learn and promoting their ability to understand and absorb new knowledge.

The theory of change also allows for the creation of a scorecard of relevant metrics to be monitored based on data from the project's beneficiaries. In the above example, the company can follow the total number of hours spent by the support team (inputs), the actions that they perform with these students (activities), the number of students who accessed the learning tool (outputs) and their assessed learning (outcomes), among other indicators.

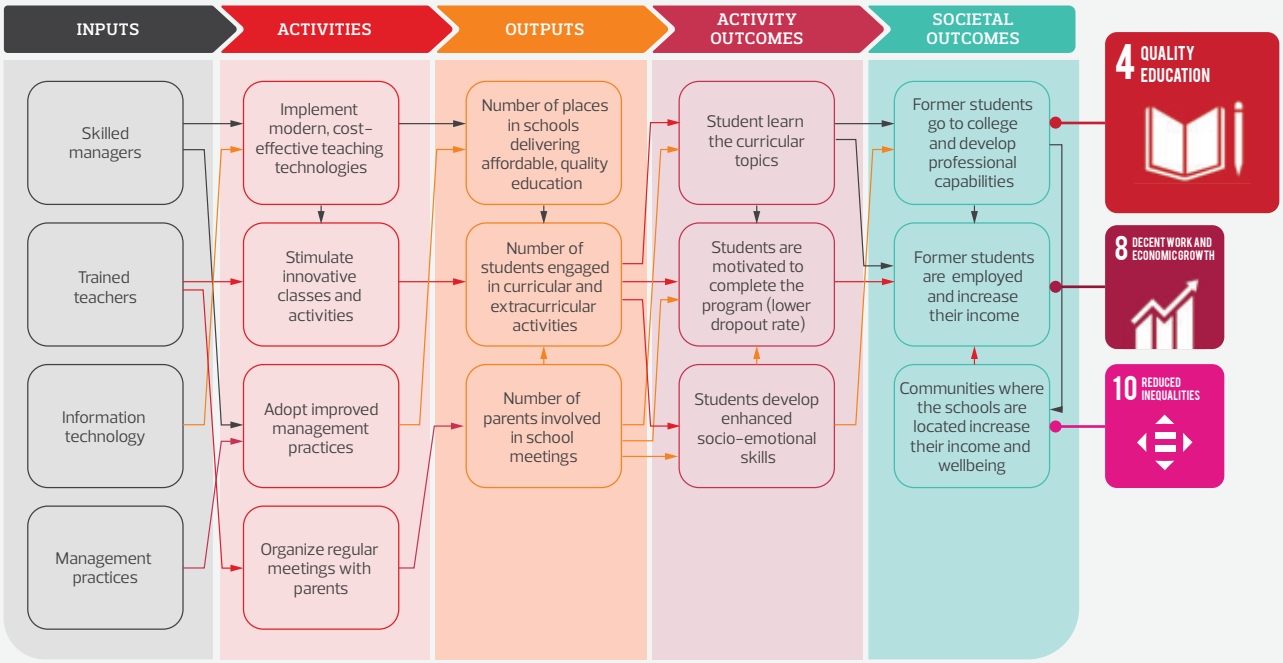
In some representations of the theory of change, the last step in the analysis is sometimes termed "impact." However, from an additionality perspective, the theory of change is not, in itself, a method for assessing impact. The project's impact, as discussed before and detailed in Stage 5, involves changes that were brought on by the project, rather than other external factors affecting the target population.

Box 2 presents another example of theory of change in the educational field that includes a more complex set of inputs and interventions and more detailed specifications of the cause-and-effect links among the building blocks of the theory. In the example below, an impact investor has acquired a group of private schools targeting low-income families and has pursued a set of pedagogical and managerial practices to improve learning and reduce dropouts in secondary schools.

Observe, in particular, how the components of the theory of change are written. Activities identify actions that managers carry out (notice the verbs "implement," "adopt," and others). Outputs involve, in general, changes in numbers and quantities. Outcomes, in turn, identify how the target population improves as a function of the proposed interventions ("students learn the curricular topics").

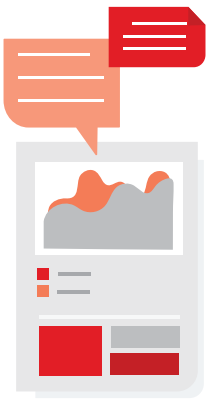
In some representations of the theory of change, the last step in the analysis is sometimes termed "impact." However, from an additionality perspective, the theory of change is not, in itself, a method for assessing impact.

BOX 2. Example of theory of change for a chain of secondary schools targeting low-income students.



The theory of change can be connected with Step 1 of the measurement plan by indicating how the proposed societal outcomes are linked with the project's objectives and SDGs. In **Box 2**, the icon representing SDG 4 ("quality education") is bigger than the others because it represents the core objective of the project, related to improving educational outcomes in secondary schools. Yet, by increasing students' access to high quality education, the project should also help them find good jobs in the future and potentially reduce income inequalities in the target communities.

When defining the theory of change, it is important to take into account all possible outcomes, both positive and negative. Neglecting negative outcomes increases the risk of generating unintended results. For instance, a project to increase income in rural communities may stimulate new agricultural activities that boost farmers' income, but at the same time cause negative spillovers in the natural environment. Certain activities may also risk violating basic human rights, for instance, in cases of gender discrimination or precarious working conditions. A suggestion is to consider potential interactions between the general objectives of the project (as discussed in Step 1), in order to provide a broader view of all possible results that may be influenced by the intervention.



5. DEFINITION OF METRICS

After completing the previous steps, it is now possible to choose the metrics for the project. Anchored in the project's theory of change, managers can develop a panel of metrics for monitoring purposes. Inputs, activities, and outputs directly inform the creation of an *operational* panel tracking the implementation of the project. In the example in **Box 2**, managers can, for instance, track the presence of skilled teachers (input), the implementation of support technologies (activity), and the number of students engaged in the activities of the school (output).

Although metrics can be chosen for all stages of the theory of change, including for the purposes of monitoring activities and outputs, monitoring and evaluating fundamental improvements in the target population requires an emphasis on outcome metrics. Thus, outcomes can be conceived as *key performance indicators* (KPIs) in the monitoring panel. Box 2 has three KPIs linked with outcomes: student learning of curricular topics, completion of the program, and development of socio-emotional skills.

For the sake of focus and simplicity, we recommend a few highly relevant outcome metrics for assessing the project's impact. These metrics should be closely linked to the outcomes outlined in the theory of change (Step 4).

Good metrics have the following attributes: [7]

- They are highly **relevant** in terms of impact generation. In other words, they must represent outcomes that are highly desirable for the target populations. If the metric follows the theory of change described in Step 4 and if the theory correctly identifies key desirable outcomes, then the metrics should be relevant.
- They must be reasonably **affected** by the action of the project managers. One should avoid, for instance, metrics that are too aggregated and broad and whose result is affected by a myriad of factors beyond those under the direct managerial control. In general, metrics linked with activity outcomes tend to be more actionable by project managers than metrics related to broader, societal transformations.

- They must be measured and verified with sufficient **precision**. In general, metrics based on objective information ("hard data") are preferable to subjective indicators with high potential for measurement error. Furthermore, the data should not be susceptible to "gaming" (manipulation) by managers, investors and other parties interested in reporting positive results. In this sense, data collected by third parties (such as independent institutes) and widely used by actors in the sector of the project tend to be preferable with respect to the precision criterion.
- They must be measured at **low cost**, including the cost of gathering data from populations used as a comparison (i.e., subjects in the control group). Measurement costs tend to drop with the use of publicly available (secondary) data, compared to customized (primary) data gathered specifically for the project.

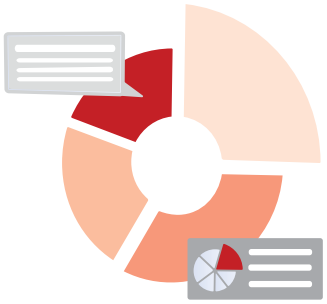
[7] These characteristics are particularly relevant since impact evaluation can also be used to set management goals; see: ROBERTS, J. (2010). Designing incentives in organizations. *Journal of Institutional Economics*, 6(1), 125–132. https://doi.org/10.1017/s1744137409990221. Characteristics of good metrics are also discussed in: Kusek, J. Z., & Rist, R. (2004). Ten Steps to a Results-Based Monitoring and Evaluation System. In *Ten Steps to a Results-Based Monitoring and Evaluation System*. Washington, DC: World Bank. https://doi.org/10.1596/0-8213-5823-5.

As a suggestion, when considering alternative metrics, each one should be assessed based on the above criteria, selecting the one best aligned with the most attributes. Unfortunately, in most cases, it is not possible to meet all of those criteria. In those cases, managers should then select metrics that best reflect the project's objectives, as well as metrics broadly used and validated.

Considering the attributes discussed above, **Box 3** illustrates an application of the **metrics menu** tool: a way to compare alternative metrics, visually identifying their advantages and limitations. For each attribute, metrics receive ratings based on the rubrics described in **Box 4**, on a 1–4 scale.

We recommend that all metrics should be compared based on their individual attributes. It is clear in the example in **Box 3** that all metrics have advantages and disadvantages. However, a way to make a general comparison of metrics is to multiply their attribute-specific scores, assuming that a preferable metric scores well in most attributes.

For instance, the second metric related to tourism-related activities has an overall score of 4 x 3 x 4 x 4 = 192. The third metric gauging user satisfaction, in turn, has an overall score of 4 x 4 x 1 x 1 = 16. The former metric is therefore generally more attractive than the latter. Yet this does not necessarily mean that user satisfaction should be ignored. If a metric is linked with an outcome from the theory of change, it is generally relevant and should be considered in the monitoring panel. Managers can possibly consider alternative ways to increase the attractiveness of the metric (for instance, reducing the costs of collection) or finding other ways to measure the same outcome.



If a metric is linked with an outcome from the theory of change, it is generally relevant and should be considered in the monitoring panel.

BOX 3. Metrics menu: Example of a comparative analysis involving potential metrics applied to a tourist park.

The following example illustrates the choice of potential metrics for a park in a forest conservation area and with tourist attractions. The government is interested in establishing contractual targets of impact to guide the work of a private company selected to manage the park. There is a particular concern to develop the municipalities around the park, as well as assuring that the company will pursue environmental preservation. In the analysis below, each metric is rated based on the rubrics presented in **Box 4**; the final score is the multiplication of the individual ratings for each attribute of the metric. Income-related metrics are relevant, precise and low-cost, however, the metrics gauging income gains from tourism-related activities is more easily affected by the private company, which can hire local people to work in the park or stimulate entrepreneurs to offer complementary services (such as hotels and restaurants). The metric of user satisfaction, although relevant and actionable, is costly and less precise – it is based on users' perceptions and subject to gaming (for instance, managers can influence data gathering to occur at moments of low visitation, thus reducing the incidence of complaints). The metric gauging the percentage of areas following environmental regulations, if informed by the managing company, may require the government to hire external technical auditing, which tends to increase costs.

OUTCOME (from the theory of change)	METRIC	RELEVANT?	ACTIONABLE?	PRECISE?	LOW COST?
Neighboring communities attain higher income and wellbeing	Increase in the per capita income of the park's municipalities and its borders.	● Local development is a key objective of project.	○ The local economy can be affected by factors other than the park's activities.	● There are reliable and publicly available statistics on local income.	● In the country where the project is located, statistics are publicly available and easily accessible.
The local community increases its income from tourism-related activities	Increase in the per capita income of services such as hotels, restaurants and others in the municipalities affected by the park.	● Local development is a key objective of the project.	◉ The park might boost the revenues of hotels and restaurants, for instance, even though they also depend on other contextual conditions.	● There are reliable and publicly available statistics on local income.	● Statistics are publicly available and broken down by sector of the economy.
Park users become more satisfied	Average satisfaction score of users, based on surveys conducted by hired institutes.	● Higher satisfaction tends to reflect higher quality of services in the park.	● An effort to promote better services can directly affect user satisfaction.	○ Satisfaction indicators are more subjective and may be prone to gaming.	○ To avoid manipulation, costly data collection by a third party will be required.
The park meets conservation regulations	Percentage of the park's areas in compliance with environmental regulations.	● Because the park is located in a conservation area, environmental conservation is critical.	● The park managers can directly control the status of the conservation areas.	◉ There are clear standards of conservation, but a concern is that they will be monitored by the park managers.	◉ Although managers can monitor the status of conservation areas, external auditing may be required.

BOX 4. Rating the metrics

The following rubrics allow for the detailed analysis of each proposed metrics according to the four attributes of relevance, actionability, precision, and cost. Each attribute receives a score on a scale from 1 to 4.

	1	2	3	4
Relevant?	The metric is weakly linked with some outcome specified in the theory of change.	The metric is linked with some outcome specified in the theory of change, but it is not clear how its variation could be valued by the target population.	The metric is linked with some outcome specified in the theory of change and its variation will be potentially valued by the target population.	The metric is linked with some outcome specified in the theory of change and its variation will be strongly valued by the target population.
Actionable?	Contextual conditions out of the control of managers involved in the activities specified in the theory of change substantially limit the effect of such activities on the proposed metric.	Although the metric can be potentially affected by the activities specified in the theory of change, this effect depends on myriad conditions beyond the control of project managers.	There are logical arguments supporting the conclusion that the metric can be affected by the interventions specified in the theory of change.	There is previous evidence that the metric can be strongly affected by interventions specified in the theory of change.
Precise?	The metric weakly reflects the focal outcome and/or can be manipulated and distorted to reflect a performance outcome that is greater than what was effectively attained in the project.	The metric moderately reflects the focal outcome and/or involves data sources with a considerable risk of manipulation or whose interpretation is questionable.	The metric adequately reflects the focal outcome and involves data sources that are not broadly recognized but have low risk of manipulation.	The metric adequately reflects the focal outcome and involves data sources that are broadly recognized and seen as legitimate indicators of the focal outcome.
Low cost?	The metric requires specific and costly measurements (for instance, local surveys).	The required data requires additional costs due to specific measurements and/or the need for external validation.	The required data can be collected at low cost using existing data sources, but access depends on authorization or extra effort by third parties.	The required data can be collected at low cost via easily accessible public sources.

6. DEFINITION OF THE METHOD FOR VERIFYING ADDITIONALITY

After choosing the metrics to track operational performance indicators and especially key outcomes, it is important to define how they will be used to verify the additionality of the project, i.e., what would have probably happened to the target population without the intervention. This stage involves the definition of the treated group (individuals and communities benefited by the project) and the control group (individuals and communities that did not receive the intervention) indicating what would have happened to the treated group without the project.

Because it might be difficult to have available data for both groups, treated and control, we recommend an initial emphasis on a few metrics discussed in the previous step that most likely meet the requirements of relevance, precision, actionability, and low cost. In the example in **Box 4**, the second outcome metric (income from tourism-related activities) is a potential candidate for the next steps involving impact evaluation based on additionality, even though the other metrics can also be followed to monitor the project's evolution.

The method of verifying additionality varies depending on the **measurement tier** chosen for the project, as detailed in the next section. Higher measurement tiers increase the confidence that the changes detected in the target population were caused by the project; however, these higher tiers also tend to increase analytical complexity and cost. At this stage, especially if higher measurement tiers are adopted, we recommend the support of professionals to help identify the most appropriate techniques as well as their potential limitations.

After choosing the method to verify additionality, and based on the selected metrics, it is then important to define precisely how to evaluate the impact of the project. In the example in **Figure 1**, impact is measured as the difference in the variation in income between the treated and control communities, considering the two pre-defined periods. This measurement allows us to gauge the total impact **size**, defined as the average gain per individual times the total number of individuals included in the treated group.

This step also involves a definition of the required **time horizon** to measure the impact. Several impact projects mandate an extended time horizon due to inherent characteristics of the activity and the need to implement a complex set of complementary actions. Steps 3 and 4, involving benchmarking and definition of the theory of change, can also allow for an assessment of previous studies examining not only the size of the expected impact, but also how long it takes for the interventions to improve the desirable outcomes.

After choosing the method to verify additionality, and based on the selected metrics, it is then important to define precisely how to evaluate the impact of the project.

7. SAMPLING PLAN

Defining the sample size (the number of people or groups that will participate in the project) is another crucial step in the measurement plan. This step, which is particularly important if higher measurement tiers are adopted, uses techniques to measure with precision the causal effect of the intervention on the target population. With an adequate sample size, the difference between the outcomes measured before and after the project and the difference between the treated and control groups can be measured with greater statistical precision.

Essentially, a larger sample reduces the risk that the analyses will not detect a positive impact of the project, supposing that a positive effect actually exists. In general, the recommended sample size depends on the expected size of the impact; all else being constant, the lower the expected impact, the larger the sample size required for statistical purposes. However, the costs of implementing the treatment and the collection of the relevant data must also be taken into account. Because the computation of optimal sample sizes requires previous knowledge of statistics, it is crucial to draw from the extant technical literature and rely on experts in impact evaluation. [8]

8. MEASUREMENT TIMELINE

At this stage, it is important to define the period of data collection, clearly indicating, if applicable, measurements that will occur before and after the intervention (see **Figure 1**). Project managers can also define intermediate measurements as a way to follow the project's evolution. As discussed before, project managers should be aware that, in most cases, the verification of impact requires a reasonable timeframe after the expected interventions. For example, in education-related projects, where student learning requires continuous and long-term effort, the time horizon tends to be long, especially if there is an interest in measuring the completion of an educational cycle or its impact on their future earnings.

The decision about the ideal temporal window to follow and measure the project's outcomes can be informed by previous studies that have evaluated the impact of similar interventions over time (as discussed in Stage 3, on benchmarking).

[8] A technical discussion on how to define sample sizes can be found in: Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61 Using Randomization in Development Economics Research: A Toolkit. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Elsevier. https://doi.org/https://doi.org/10.1016/S1573-4471(07)04061-2. A practical tool for computing minimal sample sizes can be found here: Raudenbush, S. W., et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Computer software]. Retrieved from https://sites.google.com/site/optimaldesignsoftware/home (accessed on March 20, 2022).

Measurement Tiers

Measurement tiers essentially define alternative ways to correct for external factors arising from pre-existing differences between the treated and control groups. For instance, suppose that a sponsor of educational projects wants to assess the effects of a new teaching method and the sponsor then offers this method to various schools. It is possible that the schools that will adhere to the program are precisely those schools whose managers are more motivated to pursue improvements in learning. If this variable affecting choice ("managerial motivation") is not observed by the evaluator, then they will erroneously conclude that the improvements detected in the schools that voluntarily adhered to the program came from the interventions of the program itself, rather than the natural propensity of the chosen schools to improve their teaching.

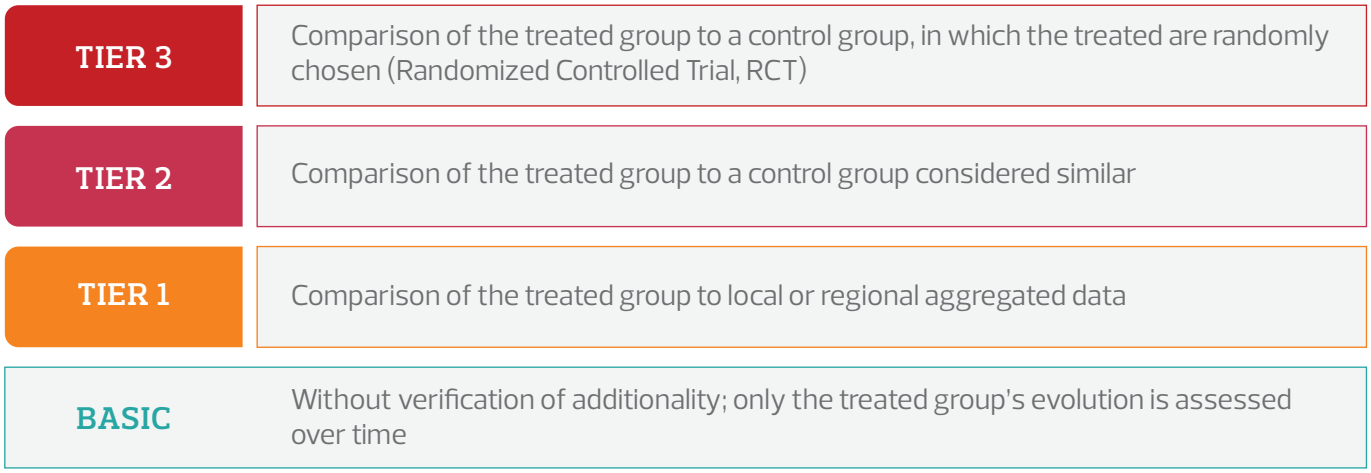
To mitigate this error, ideally the treated and control groups should be chosen at random. With a sufficient number of cases (random draws), the treated and control groups will tend to be similar in terms of attributes that the impact evaluator can directly measure (such as demographic characteristics) as well as attributes that are more difficult to directly observe (in the example above, managerial motivation). Such similarity is crucial because, as discussed before, the purpose of the control group is to indicate what would have happened to the treated group without the intervention. In other words, the control group should mirror a "parallel reality" where the individuals of the treated group were not subject to the proposed project.

However, randomizing individuals or groups is a particularly complex process, and in many cases even unfeasible. Regardless of whether the sponsors are funds, individuals, companies or governments, the decision to invest or not in a particular project often follows a pre-specified plan. Therefore, in most cases, sponsors and project managers have to adopt methods that attempt to correct for the pre-existing differences between the treated and control groups.

Whatever the verification method, it is important to strive for total **transparency** regarding the method's likely limitations. We propose alternative **measurement tiers** below that vary in accordance with the desired robustness of the assessment, especially in terms of the rigor with which the *causal* effect of the project is assessed.

As seen in **Figure 3**, we begin with a type of measurement referred to as *basic*, which does not involve assessment of additionality, in order to show how each additional level allows for more rigor in the estimation of the project's impact. This approach is usually employed in monitoring panels that simply want to track changes in the treated group (i.e., those directly affected by the project). Because project managers usually adopt more than one metric, it is possible that the various selected indicators will be evaluated following distinct measurement tiers. Thus, it is likely that most outcomes will be evaluated using the basic approach, whereas some selected outcomes will be evaluated using the more robust measurement tiers.

Figure 3 – Measurement Tiers for Verification of Additionality.



BASIC – MEASUREMENT OF HOW THE TREATED GROUP EVOLVED

This approach *does not* involve verification of additionality and *should not* be considered a measurement tier, given that it does not make use of counterfactual analysis (i.e., what probably would have happened to the target population with the project). However, for a broad range of metrics, it is the most feasible method, given that it is easier to collect data from individuals and groups subject to the intervention. In this case, we simply observe how certain metrics varied over time, before and after the starting date of the project. This practice is also widely used to monitor a scorecard of project-specific performance metrics for the treated populations, as discussed in Step 4. Despite its simplicity and practicality, managers should avoid automatically inferring that any positive variation in the measured outcomes was caused by the project. The target population itself may already be improving even before the interventions began. In addition, the project may have benefited from external changes that positively affected the target population. Therefore, managers adopting the basic approach should treat the measurements simply as indicators of changes in the treated group, which might or might not have been caused by the project's proposed interventions.

TIER 1 – MEASUREMENT COMPARING THE TREATED GROUP TO LOCAL OR REGIONAL AGGREGATED DATA

At this tier, we use aggregated data to compare the outcomes of the treated group and what would have happened to that group without the intervention. For comparison purposes, it is recommended to use aggregated data that are already available for the region or location where the investment is being made. It is important to obtain data about what happened to the target population versus variations in the aggregated metrics before and after the intervention (**Box 5**).

In this tier, managers must also take into account the size of the intervention. If the project affects a large number of individuals in a given region, the aggregate data will be significantly affected by the intervention, and a comparison will not be appropriate. A possibility in this case is to compare the treated group to individuals in other similar locations nearby.

BOX 5. Example of verification of additionality at Tier 1

A food processing company decides to acquire agricultural products from small family farmers in certain low-income municipalities, while at the same time providing these farmers with technical and managerial support. The definition of the target communities is deliberately chosen by the project manager. Using the farmers' income as an outcome metric, the assessment of impact is done by comparing the income of families supported by the project to the aggregated income of families in rural areas in the municipality or locality where the investment was made, before and after the intervention. Additionality at Tier 1 is verified if the variation in the income of the supported families is higher than the variation in the aggregated income of families in the same region.

TIER 2 – MEASUREMENT USING A CONTROL GROUP WITH SIMILAR CHARACTERISTICS TO THE TREATED GROUP

At this level, although randomization has not been carried out, we seek to create a control group with individuals or communities comparable to the treated cases. For this reason, verification of additionality at Tier 2 involves techniques that are referred to as **non-experimental**. Similar to Tier 1, whoever will benefit from the project is already predetermined. Moreover, in the evaluation process, techniques are implemented to construct a control group of individuals who are similar to those in the treated group. In other words, instead of using aggregated data from a particular location, we seek to track groups or individuals who are similar to the treated individuals, but that did not receive the intervention.

A common way to assess impact at Tier 2 is to employ the method of **differences-in-differences**, exemplified in **Figure** 1. Following this method, the impact of the project is measured as the difference between the evolution of the outcome metrics of the treated group and the evolution of the outcome metrics of the control group. The evolution of the control group essentially serves as a counterfactual estimate: what would have happened to the target population without the intervention.

Notice that Tier 2 requires that the treated and control groups are truly comparable in terms of their underlying traits and the evolution of their outcome metrics. Thus, in the differences-in-differences method, the outcome metrics of the treated and control groups should exhibit a similar **trend** after the intervention had it not occurred. However, because one cannot test whether trends are parallel after the intervention, researchers usually test this assumption by looking at past trends. The idea is that, if past trends in the outcome of interest look parallel for the treated and untreated, one can be more confident that it would probably continue to look parallel after the treatment implementation had it not been implemented. In Figure 1, the treated and control communities evolved "in parallel" before the project started, thus suggesting that they were affected by similar factors. Divergent trends would suggest that groups were evolving in distinct ways due to various factors unrelated to the project.

Another common way of choosing individuals with similar characteristics between the control and treated groups involves a technique called **matching**. Based on observable characteristics that are relevant to influencing the participation of individuals in the program (such as age, gender, income, etc.), in this technique we try to find one or more subjects in the control group that are as similar as possible to each treated subject (see example 6.1. in **Box 6**). Matching techniques are also commonly used in combination with the differences-in-differences technique.

In cases where there is competition to participate in the project and a particular selection criterion, another possibility is to consider the **discontinuity** that appears when we observe participants close to a given threshold: those who were almost rejected (treated) and those who were almost accepted for the project but were below the selection threshold (control) (see example 6.2 in **Box 6**).

Still in other situations where only a particular case is assessed (e.g., a single organization or municipality), there is the possibility of creating a **synthetic control** from the combination of other cases that collectively approximate the treated case. This technique requires the availability of data of the treated case and other potential cases before starting the project. This is a way to create a synthetic control that can adequately replicate the prior evolution of the case that received the intervention. [9]

Instead of using aggregated data from a particular location, we seek to track groups or individuals who are similar to the treated individuals, but that did not receive the intervention.

BOX 6. Examples of verification of additionality at Tier 2

Example 6.1: In a project where the outcome to be assessed is household income (see Box 3), we compare the income levels of the treated households to those in the control group. The project manager deliberately defines who will benefit from the intervention. Households in the control group are then selected based on certain observable characteristics that are similar to those in the treated group. For example, a control group could consist of a handpicked set of households with initial levels of income and schooling and number of children that are similar to the levels observed in the treated group.

Example 6.2: An organization wants to implement a microfinance project. Managers can invite small entrepreneurs to participate in the program and then rank these entrepreneurs based on credit risk indicators (income, track record, etc.). If there is limited money to be disbursed as micro loans, only the most highly ranked will be selected, based on a cutoff criterion defining a minimum acceptable credit risk. At the discontinuity of this particular threshold, we can suppose that individuals are more or less similar to each other, since they have similar assessments of credit risk. We thus can compare the group of entrepreneurs selected just above the cutoff to those who were not selected, but were just below the selection cutoff in the ranking.

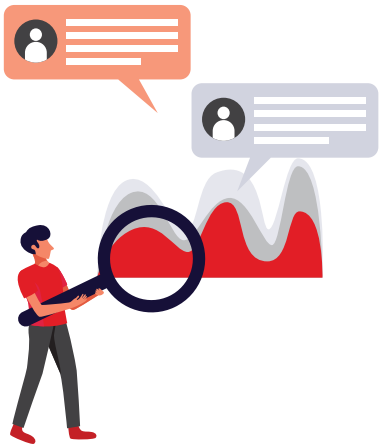
[9] As an example of matching technique see: Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing Matching Estimators for Average Treatment Effects in Stata. *The Stata Journal: Promoting Communications on Statistics and Stata*, 4(3), 290–311. https://doi.org/10.1177/1536867X0400400307. For a general discussion involving various methods, see: Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton University Press.; and: Menezes Filho, N. (Org.). (2012). *Avaliação Econômica de Projetos Sociais*. São Paulo: Dinâmica Gráfica e Editora Ltda. The synthetic control technique can be seen in: Abadie, A., Diamond, A., & Hainmueller, A. J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

TIER 3 – MEASUREMENT WITH RANDOMIZATION (RCT)

Tier 3, which involves **experimental** techniques, ensures maximum confidence in the impact estimation. At this tier, the decision of which individuals or groups will be subject to the intervention is made at random, thereby reducing potential biases in the selection of the project's beneficiaries. In cases where it is neither possible nor desirable to exclude certain groups, an alternative is to implement an experimental design where the randomly selected group is *encouraged* to adopt what is proposed by the intervention (**Box 7**).

If performed correctly, random selection will ensure there are no differences between the treated and control groups. For this reason, Tier 3 methods in general do not require the measurement of outcomes prior to the intervention, as indicated in **Figure 1**. Following adequate randomization procedures, impact evaluators can assess the comparative performance of the treated and control groups after the project has been implemented.

At this tier, it is particularly important to pay attention to the correct computations of minimal sample size, as discussed in Step 7 of the previous section. It is important to distinguish between random sampling and a randomly assigned selection of treated groups. For instance, suppose that a given education project targeted 30 schools in such a way that the choice of schools was arbitrary, without randomization. Tier 3 would not hold in this particular case even if managers randomly chose another 30 schools to serve as a control group. That is because the selection of those who received the intervention was not based on random assignment, thereby creating the risk that other unobservable factors biased the impact assessment of the treated group.



[10] To assess the effect of the intervention in this case, one option is to use the estimator known as LATE (Local Average Treatment Effect). See: Angrist, J. D., & Pischke, J. S. (2008). M*ostly harmless econometrics*. Princeton University Press.

BOX 7. Examples of verification of additionality at Tier 3

Example 7.1: A company offers a microcredit program in a location where there is no possibility of funding the community as a whole. Managers can invite potential recipients and then randomly choose who will receive credit. Alternatively, the company can lend to the best ranked entrepreneurs in terms of their credit risk, and then randomize among the set of individuals who were almost selected, that is, who were just below the minimal credit risk adopted as a cutoff criterion. Taking this set of entrepreneurs, the company can randomly pick a group of treated individuals who will receive credit; the group of non-selected individuals according to this randomization procedure will then comprise the control group.

Example 7.2: Students in a broad set of public schools are granted access to an online learning technology platform. Even though all students have access to the platform, a group of students is randomly selected to receive messages stimulating its use. Considering that only a fraction of the students who received the messages will in fact use the platform (i.e., only those students will be effectively treated), it is possible to compare the performance of those students to the performance of the control group of students who did not receive the messages. This analysis, however, requires statistical adjustments, since we should assess the performance of those who effectively used the platform, not necessarily those who received the stimulus or not.[10]

COMPARING THE MEASUREMENT TIERS

Table 1 presents a comparative assessment of the benefits and limitations of each measurement tier. In general, the assessment of impact becomes more robust as we move towards Tier 3, since there is increased confidence that the project did cause the outcomes measured in the target population. However, at the same time, the complexity of the design and analyses increase, requiring more financial and technical resources to carry out the evaluation.

Table 1 — Advantages and limitations of measurement tiers

	ADVANTAGES	LIMITATIONS
BASIC	· Data are more easily collected from the project itself and its target population. · Greater analytical simplicity (simple comparison of what happened to the target population before and after the project starts).	· Because there is no counterfactual assessment (what would have happened to the target population without the project), there is no verification of additionality.
TIER 1	· In most cases, managers have access only to aggregated data (for instance, municipal income instead of household income). · Compared to other methods for verifying additionality, the analysis is simpler (comparison of project outcomes to aggregated outcomes).	· The characteristics of the broader population can be very distinct from the target population, even in the same geographical area. · The comparison does not allow for a causal assessment of the project, since there is no use of techniques to control for potential differences across groups.
TIER 2	· In most cases, the target population is already defined, thus rendering randomization unfeasible. · By creating control groups with similar characteristics to the target population, Tier 2 methods avoid bias generated by marked differences between groups. · The various techniques at Tier 2 allow for statistical inference on the impact observed in the treated group in comparison to the control group.	· Tier 2 techniques only allow for an examination of the causal effect of the intervention if we assume that there is no critical influence of factors that are not observed and measured, which is difficult to verify in practice. · It is necessary to gather information on the characteristics of individuals included in the treated and control groups. · Tier 2 techniques require the existence of cases in the control group that are similar, in their underlying traits, to the cases in the control group. Otherwise, the comparison is unfeasible.
TIER 3	· Through randomization, it is possible to guarantee with higher confidence that the measured effect was effectively caused by the project. · In general, randomization does not require the measurement of outcomes prior to the project start; subjects can be compared after the implementation of interventions.	· Tier 3 techniques require paying special attention to sample size. Their use can be unfeasible in cases where there are few cases receiving the intervention. · Experimental designs at Tier 3 are highly subject to spillover effects, attrition, and other problems discussed in the next section of the Guidelines. · Randomization can create ethical dilemmas in cases where it is problematic to exclude individuals from a certain intervention.

MEASUREMENT PRECAUTIONS

Some extremely important precautions should be taken when performing a satisfactory impact assessment. Below, we describe some critical issues that the project manager should consider when measuring impact. [11]

EXTERNALITIES

Externalities or "spillovers" between individuals from different groups should be kept at a minimum level. Consider, for instance, that we want to assess the impact of a company's initiative to increase the overall income of a community in a given location. To do so, before the project begins, we collect data on this community and on the neighboring community, which does not benefit from the project. Externalities occur if the treated community, which receives a higher income as a result of the investment, can spend part of its resources on either buying products from or transferring income to the other community. If this occurs, the program's additional impact will be underestimated. These externalities are problematic for impact assessments, since they can distort the result by "contaminating" the control group with the intervention in the treated group.

To circumvent this problem, it is important to ensure there is no communication, information exchange or physical exchange of goods and services between the different groups. One way to do this is to make sure the treated and control groups are geographically distant or isolated, thus diminishing the risk of spillovers.

ATTRITION

Given the time between the intervention and the subsequent evaluation, it is possible to expect a mismatch between individuals who were initially observed and those monitored at the final round of measurement in both the control and treated groups. For instance, some individuals who were observed at the baseline may no longer reside in the same location during the final round of measurement.

It is worth noting that randomization does not completely prevent non-random attrition at the end. Moreover, attrition problems reinforce the need to define a sufficiently large sample size before performing the measurement. A practical example of this problem can be seen when assessing projects aimed at increasing income in rural communities. Even though the groups initially may have been selected in a random manner, the exit of individuals who did not receive the income transfer versus the continuity of individuals who received the transfer could cause non-random

It is worth noting that randomization does not completely prevent non-random attrition at the end.

[11] This section is heavily based on: Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61 Using Randomization in Development Economics Research: A Toolkit. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Elsevier. https://doi.org/10.1016/S1573-4471(07)04061-2

attrition. In this case, attrition induces biases because we lose a large and non-random amount of data on individuals in the control group.

It is always important to be transparent about the level of attrition in both the treated and control groups. In other words, it is necessary to state how many individuals in the treated and control groups were initially monitored but were not found in the subsequent rounds of measurement. It is also desirable to as compare the average characteristics of treated and control groups before and after attrition.

HAWTHORNE AND JOHN HENRY EFFECTS

Finally, another major limitation in measuring impact is the behavioral change of individuals in the treated group as well as in the control group. The so-called **Hawthorne effect** may occur when the individuals in the treated group perceive that they are under some form of intervention and consequently change their behavior accordingly.[12] In education projects, for example, teachers or students may alter their behaviors as a result of their awareness that the group is benefiting from the intervention.

The **John Henry effect**, in turn, may occur when teachers in the control group feel challenged and start to compete with their counterparts in the treated group to show that they should also be eligible to benefit from the same program.[13] Conversely, if teachers become less motivated because they are not receiving the benefits of the treated group, the program's impact will be overestimated. In both cases, the program's real effect can be distorted.

The so-called Hawthorne effect may occur when the individuals in the treated group perceive that they are under some form of intervention and consequently change their behavior accordingly

Although behavioral responses are always complex, there are ways to design the measurement to minimize the Hawthorne and John Henry effects. For example, three groups can be used: treatment, control, and **placebo**. The latter is a group that is observed or briefly interacts with project managers but does not receive the full set of interventions specified by the project. Consider, for example, a project in which a large firm wants to improve the income of certain communities by procuring their local inputs. Then the firm should consider not only a community of actual suppliers (treatment) and a comparable community of non-suppliers (control), but also monitor a placebo community of non-suppliers with which the firm has some form of relationship (e.g., the firm could send procurement managers to contact communities without actually signing procurement contracts). The evidence of impact increases when the treated group exhibits improvements above and beyond the control *and* placebo groups.

[12] The Hawthorne effect was named after an experiment in 1927 at the factory of the West Electric Company (Chicago, USA), where it was found that individuals changed their behavior according to perceived changes in the work environment.

[13] The John Henry effect is based on a tale of a worker in a location where a new tool to increase productivity was introduced, still in its testing phase. On this occasion, the person felt challenged to work harder in order to show that he could outperform the new technology.

Monetizing the impact of the project

This Guide emphasizes procedures to estimate the impact of a project based on certain objectives and outcomes outlined by the proposed theory of change. It is not the purpose of this Guide to discuss how to compute the economic return of a socio-environmental intervention, as a function of how much was spent to promote it, compared to the benefits generated to treated individuals. Nevertheless, in this final section we provide some brief suggestions of how to evaluate the impact of the project from an economic standpoint.

When there is available cost data on the interventions, managers can start by computing indicators of **cost-effectiveness**. For instance, suppose that a training program targeting 1,000 unemployed individuals managed to employ 100 additional people compared to what would have happened without the project. As seen in the previous sections, this estimate of impact (additionality) can come from comparisons between treated and control groups. If the training program cost $200,000, then its cost-effectiveness ratio would be $200,000/100 = $2,000 per each *additional* individual who found a job.

Another possibility is to monetize the evaluated outcomes. The **benefit-cost ratio** of the project indicates how much impact was generated (in monetary terms) for each monetary unit invested in the treated group. For instance, a study computed the monetary benefits of investing in superior sanitation infrastructure—including avoided health treatment due to a lower incidence of diseases and the extra time that healthier individuals can spend on their productive activities. The estimated benefit-cost ratio was 5.5: each dollar invested in the project is expected to generate 5.5 dollars in social benefits. [14]

In some cases, with these ratios one can compare the social or environmental outcomes of two or more distinct projects. For instance, a study in education showed that increasing school hours from 4 to 5 hours or reducing class sizes from 38 to 30 students had similar impacts on student grades. However, when their costs were considered, the project involving increased school hours had the highest benefit-cost ratio.[15]

There are also several ways to monetize outcomes. For instance, drawing from the example in **Box 2**, an educational project in high schools can enhance student learning and increase the likelihood that students will get into a good university, with a positive impact on their future income. These gains can be compared to the cost of the project.

Another possibility is to estimate beneficiaries' willingness to pay. The methodology involves asking the public of interest directly how much they would be willing to pay for a particular good or service. It is sometimes used to put a value on cultural projects or initiatives designed to increase satisfaction from a certain activity (such as visits to parks or museums). The "price" reported by the service users can then be compared to its cost.

For further details on how to monetize impacts, please refer to *Insper Metricis publication* *Social Impact Monetization.*

[14] See: Hutton, G. (2013). Global costs and benefits of reaching universal coverage of sanitation and drinking-water supply. *Journal of Water and Health*, 11(1), 1–12. https://doi.org/10.2166/wh.2012.105. For a more general discussion on effectiveness-cost and benefit-cost analyses, see: Boardman, A. E., Greenberg, D. H., Vinning, A. R., & Weimer, D. L. (2018). *Cost-benefit analysis: concepts and practice*. Cambridge: Cambridge University Press.

[15] See: Oliveira, J. M. de. (2010). *Custo-efetividade de políticas de redução do tamanho da classe e ampliação da jornada escolar: uma aplicação de estimadores de matching*. 31° Prêmio BNDES de economia. Retrieved from https://web.bndes.gov.br/bib/jspui/bitstream/1408/2567/1/Custo%20Efetividade%20de%20Politicas%20de%20Reducao%20do%20Tamanho%20da%20Classe%20e%20Ampliacao%20da%20Jornada%20Escolar_P_BD.pdf

Cataloging-in-Publication Data (CIP)

159g Insper Metricis

Guide to the assessment of socio-environmental impact for use in impact-oriented projects and investments: general guide focused on and verification of additionality. – 5th. ed. - São Paulo: Insper, 2022.

30 p. ill.: col.

ISBN 978-65-991497-5-7

1. Impact investing 2. Impact evaluation 3. Additionality 4. Socio-environmental 5. Project Management I. Author II. Title

CDU 504

Cataloging: Ricardo Rodrigues Ramos CRB 8/9309

Insper METRICIS

Núcleo de Medição para Investimentos de Impacto Socioambiental