IMPACT EVALUABILITY TOOLKIT

THE ABDUL LATIF JAMEEL POVERTY ACTION LAB (J-PAL)

CLEAR SOUTH ASIA

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

IFMR
INSTITUTE FOR FINANCIAL MANAGEMENT AND RESEARCH

clear
Regional Centers for
Learning on Evaluation and Results

Suggested Citation: Jetha, Qayam, Kanan, Harini, and Escueta, Maya. 2017. "Impact Evaluability Toolkit." J-PAL South Asia and CLEAR South Asia.

Design: Amanda Kohn, Elizabeth Bond

FRONT COVER: AUDE GUERRUCCI | ENABLING MICROENTERPRISE DEVELOPMENT IN SUB-SAHARAN AFRICA THROUGH THE PROVISION OF FINANCIAL SERVICES

TABLE OF CONTENTS

Table Of Figures

Table Of Boxes

ABOUT THIS TOOL

A fundamental question for development practitioners to ask is: did our program work? That is, did our program have an effect on a specific outcome, and if so, what is the magnitude of the effect? Answering this question credibly requires rigorous impact evaluation (IE) studies that identify the changes in outcomes that occur as a direct result of the program. Evidence generated by such studies enables both learning and accountability, and can increase the capacity for evidence-informed decision making. Ultimately, by expanding the state of knowledge on what development solutions work, what don't, and why, policy makers have greater scope to combat poverty through the design of effective social programs.

However, impact evaluations are a costly and complex tool that cannot answer all questions asked by every stakeholder. As such, being able to evaluate the right questions, at the right time, and in a rigorous manner is essential. The Impact Evaluability Toolkit (IET, hereafter) offers a comprehensive framework for formulating an impact evaluation strategy. The objective of the IET is to guide users on the Selection Process, which is defined as the steps required to shortlist and ultimately select the evaluations that the user decides to commission for an evaluation. The Selection Process involves (1) setting an impact evaluation agenda, and (2) identifying the technical requirements necessary for an impact evaluation to make a credible, causal claim. Such information is relevant to a diverse group of actors including governments, NGOs, academics, donors, and any other organization that endeavors to carry out purposeful, strategic, and informative impact evaluations.

Section 1 of the IE Tool, "What to Evaluate", discusses the process of characterizing the usefulness of an evaluation, defines criteria for when to do an evaluation, and sets out steps to build a theory of change. Section 2, "How to Evaluate", presents a technical overview of impact evaluation design to help users determine the feasibility of evaluations, and also includes budgetary considerations to factor into decision making as well as steps on how to manage and commission an impact evaluation. The Appendix provides useful resources on generating an Impact Evaluability Activity Assessment, crafting a theory of change, finding the right methodology to use in an evaluation, and drafting terms of reference for evaluating agencies.

ANJALI GUPTA | FIRST DAY OF MATH GAMES BASELINE IN THE KONDLI VILLAGE, NEW DELHI. FIELD MONITOR/MANAGER, DIRECTING THE SURVEY TEAM

INTRODUCTION

I. WHAT IS IMPACT EVALUATION?

When we talk about the word "impact" in the context of impact evaluation, what exactly do we mean? The word has many different uses and meanings. Even when used in the context of impact evaluation, various organizations will use the term differently.[1]

One common definition for the evidence produced by an impact evaluation, shared by J-PAL/CLEAR South Asia and the Independent Evaluation Group at the World Bank, is as follows:

The causal effect of the program or policy on an outcome of interest determined by comparing the outcomes of interest (short-, medium-, or long-term) with what would have happened without the program—a counterfactual.[2]

There are two key concepts within this definition:

1. **Establishing causality:** In the context of impact evaluation, this means isolating the singular effect of the program or policy, independent of any other intervening factors, and being able to estimate the size of this effect accurately.

2. **The counterfactual:** Represents the state of the world that program participants would have experienced in the absence of the program. Unfortunately, this state of the world is purely theoretical; we have no way to observe it. The challenge therefore, is to carefully select a *comparison group*, that is, a group that can be argued to credibly approximate the *counterfactual* - what would have happened without the program. The comparison group must be similar to the counterfactual along all characteristics that might arguably effect the outcome of interest (randomized experiments provide perhaps the most widely accepted way of creating such groups).

II. WHY EVALUATE?

There are a number of reasons why an organization or entity might want to evaluate, but the primary reasons that are commonly discussed in the evaluation and development community are that evaluations enable two processes: ***learning*** and ***accountability***.

a. **Learning** means using the findings of an evaluation to feed back into program design or implementation. Thinking a bit more deeply about how learning can function and what it really means; learning involves not just "pragmatic problem-solving, but also reflection on the process by which this happens."[3] This process involves not just a critical reflection, but also identifying and testing underlying assumptions, analyzing multiple lines of evidence, and relating this to expectations and consequences. For learning to truly be integrated into organizational behavior, a feedback system must be in place for the findings of an evaluation to be integrated back into program design and implementation.

b. **Accountability** means using the findings of an evaluation to answer key questions about what was done, why it was done, and how effectively. This information can be used to report upwards, to donors or superiors, or downwards, to beneficiaries or constituents. Accountability involves a process of developing clear expectations, using the evidence from an evaluation to draw conclusions about the degree to which expectations were met, and engaging in a dialogue with relevant actors who need to be held to account.[5]

Increasingly, impact evaluation has become a strategic tool for developing organizational priorities.[6] However, organizations must think strategically about how they wish to use the results before choosing which programs or activities are worth evaluating. With the understanding that impact evaluation provides precise answers to narrowly defined questions, let's discuss in greater depth the nature of questions that can be answered by impact evaluations.

III. WHAT TYPES OF QUESTIONS CAN IMPACT EVALUATION ANSWER?

Different types of evaluations answer very different sets of questions. Evaluations that answer the question, "how is our program or policy doing?", are broadly defined as **program evaluations**. More formally, a program evaluation is the process of assessing the design, implementation, and results of programs and policies considering five criteria: *need, relevance, efficacy, effectiveness,* and *efficiency*. These five criteria provide a useful typology for differentiating the questions that can be answered by program evaluation methods[7].

---

[1] Alternative impact evaluation definitions suggest that a counterfactual is not strictly necessary to claim impact. However, in this toolkit we restrict our definition to quantitative impact evaluation that produces a credible and rigorously defined estimate of the counterfactual.

[2] World Bank Group, "Impact Evaluations: Relevance and Effectiveness," and J-PAL website: www.povertyactionlab.org.

[3] Guijt, *Capacity Development in Practice, Chapter 21: Accountability and Learning,* 282.

[4] Ibid.

[5] Guijt, *Capacity Development in Practice,* 283.

[6] See World Bank Group, "Impact Evaluations: Relevance and Effectiveness," and Goldstein, "DFID's Approach to Impact Evaluation."

[7] The typology of program evaluation methods presented is based loosely on the framework set forth in, "Evaluation: A Systematic Approach" by Rossi, Lipsey, and Freeman

[8] Paul Gertler et al., *Impact Evaluation in Practice.*

[9] Blomquist, "*Impact evaluation of social programs: A policy perspective*", September 2003.

PROGRAM EVALUATION CRITERIA AND METHODS

PROGRAM EVALUATION CRITERIA	PROGRAM EVALUATION METHOD
Need *"What is the diagnosis of the social problem the iron supplement program intends to address?"*	**Needs Assessment:** A needs assessment is a systematic approach to identifying the nature and scope of a social problem, defining the population affected by the problem, and determining the service needed to solve the problem.
Relevance *"To what extent does providing the iron supplement program meet the needs of intended beneficiaries?"*	**Program Theory Assessment:** A program theory assessment (a) models the theory behind the program, presenting a plausible and feasible plan for improving the target social condition, and (b) assesses how well the program's theory meets the targeted needs of the population.
Efficacy *"Is the iron supplement program implemented as intended to the appropriate recipients?"*	**Process Evaluation:** A process evaluation analyses the effectiveness of program operations, implementation, and service delivery. They help identify whether services are delivered as intended, how well service delivery.
Effectiveness *"What is the casual effect of the iron supplement program on anemia rates?"*	**Impact Evaluation:** An impact evaluation determine the causal effect of the program or policy on an outcome of interest determined by comparing the outcomes of interest (short-, medium-, or long-term) with what would have happened without the program—a counterfactual.
Efficiency *"How much does the iron supplement program lower anemia rates relative to the program's cost?"*	**Cost Effectiveness Analysis:** A cost-effectiveness analysis takes the impact of a program and divides that by the cost of the program. Usually this is a comparative exercise, taking multiple programs and comparing them using the same unit.

Impact evaluation is a program evaluation method that focus exclusively on answering cause-and-effect questions. Usually, cause-and-effect questions are framed as: *What is the impact (or causal effect) of a program on an outcome of interest?*[8] For the purposes of defining our research questions, and for thinking about impact evaluation strategically, we will think of an impact evaluation as a methodology that identifies the changes in an outcome that are *directly attributable to the program itself.* To get a sense of the kinds of questions that can be answered through impact evaluation, here are some examples:

Examples of questions an impact evaluation *can* answer:

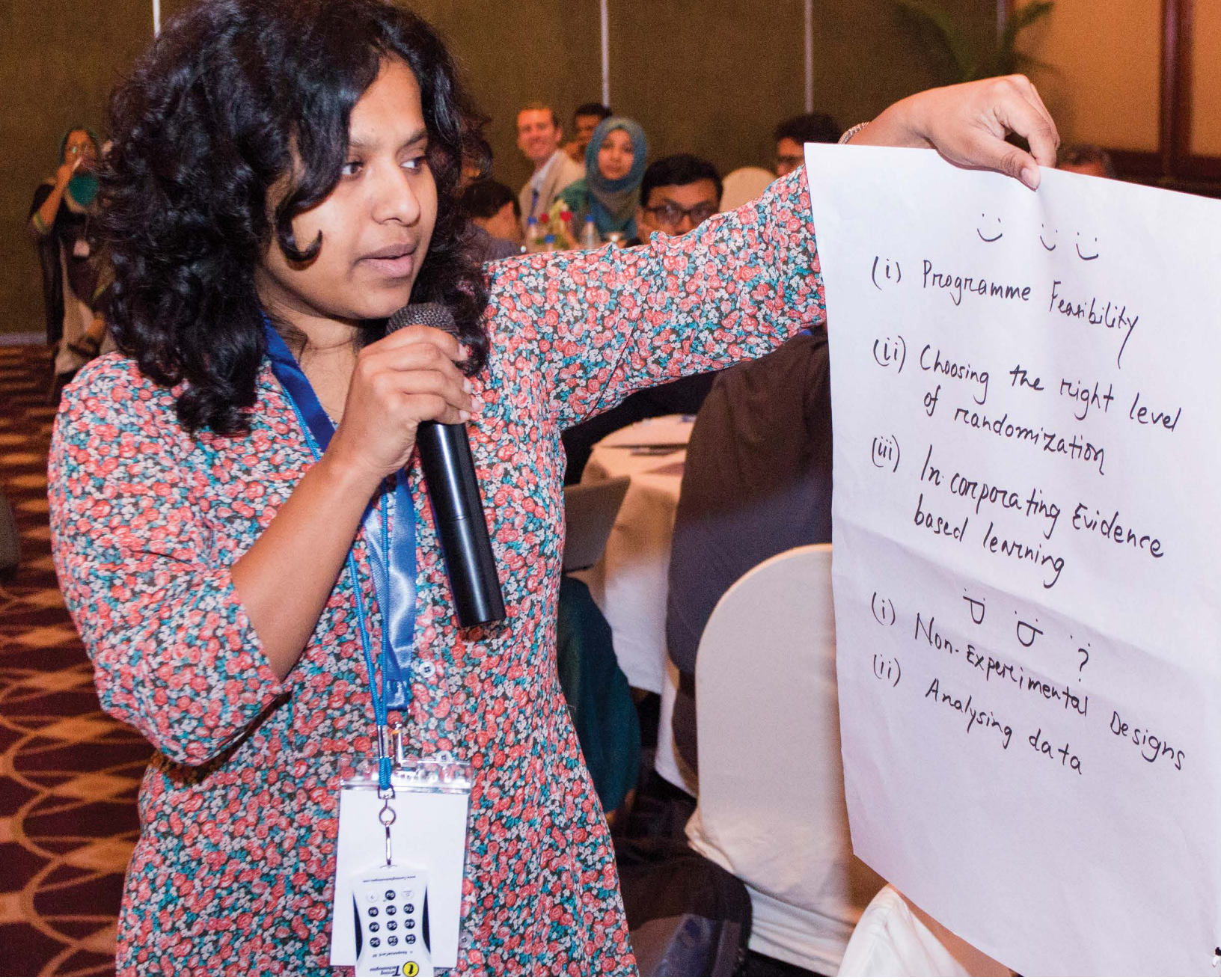
1. *What is the causal effect of student scholarships on school attendance?*
2. *What is the causal effect of introducing high quality ACTs (anti-malarial treatments) on the subsequent demand for such drugs?*
3. *What is the causal effect of conditional cash transfers on household consumption?*

There are, nevertheless, many questions of interest that cannot be addressed through an impact evaluation and would be better suited to other types of evaluation.

Examples of questions an impact evaluation *cannot* answer:

1. *What was the impact of the 2010 Minimum Wage Bill in Hong Kong on the living standards of workers?*
   a. Impact evaluation cannot answer questions for which it is difficult to establish a credible counterfactual. To identify the answer to this question through an impact evaluation, one would have to find a proxy for the state of the world where the Minimum Wage Bill was not legislated in Hong Kong, but yet all other environmental and institutional factors remain the same. Without variation in adoption or implementation of the Bill within Hong Kong, this is an impossible task.
2. *What would the impact of the Mexican cash transfer program Progresa have been on primary learning outcomes if the program had been universal instead of targeted?*
   a. Impact evaluations focus on assessing existing programs as they are currently implemented. They are unable to answer hypothetical "what-if" questions.[9]

Armed with an understanding of what an impact evaluation is, why one would choose to conduct an impact evaluation, and the sorts of questions one can reasonably expect to answer through an impact evaluation, we can now proceed to the IE Tool and begin to answer substantive questions related to, **what to evaluate using an impact evaluation** and **how to credibly conduct an impact evaluation**.



J-PAL/IPA | SOUTH ASIA REGIONAL IMPACT EVALUATION WORKSHOP

THE IMPACT EVALUABILITY TOOL

Settling on a final selection of programs to be evaluated using an impact evaluation is a lengthy process involving high stakes; only a limited number of evaluations can be undertaken and so careful thought must be given to selecting prospective evaluations. To aid in decision-making, the Impact Evaluability Tool provides considerations for selecting impact evaluations that are both strategic and rigorous. In the "**What to Evaluate**" section, practical guidelines for defining an impact evaluation agenda are presented, while the subsequent "**How to Evaluate**" section lays out the necessary vocabulary and processes for determining the technical feasibility of evaluations.

Both the What to Evaluate and How to Evaluate sections inform the **Selection Process**, which is defined as the steps an organization follows to move from the universe of possible evaluations to the subset of evaluations that an organization commissions for an evaluation. The Selection Process involves an initial shortlisting of evaluations based on an organization's impact evaluation agenda, and a final selection of projects based on technical aspects of the evaluation and budget. The Selection Process is foreshadowed in Box 1 and Figure 1 below.

To assist in the final selection of evaluations, the "**Impact Evaluation Activity Assessment**" form, provides a useful template for compiling information on impact evaluation opportunities that have cleared the initial shortlisting. The document is divided into five parts: (1) criteria satisfied by the evaluation; (2) the usefulness of conducting the evaluation; (3) preliminary research questions; (4) proposed evaluation method and methodological limitations; (5) estimated budget and funding available. These five parts, in addition to the program's Theory of Change, can be used to determine a final selection of evaluations out of the initial shortlist. Systematic use of the Impact Evaluation Activity Assessment can create transparency in the Selection Process. See **Appendix A** for the Impact Evaluation Activity Assessment as well as detailed instructions on how to correctly fill in and use this form.

BOX 1: THE SELECTION PROCESS

Section 1 – What to Evaluate: Setting an Impact Evaluation Agenda

Step 1: Identify the usefulness (use-value) of evaluations

Step 2: Define a set of shortlisting criteria that reflect an organization's evaluation priorities

Use Steps 1 – 2 to make an initial shortlist of evaluations.

Step 3: Draft a Theory of Change (ToC)

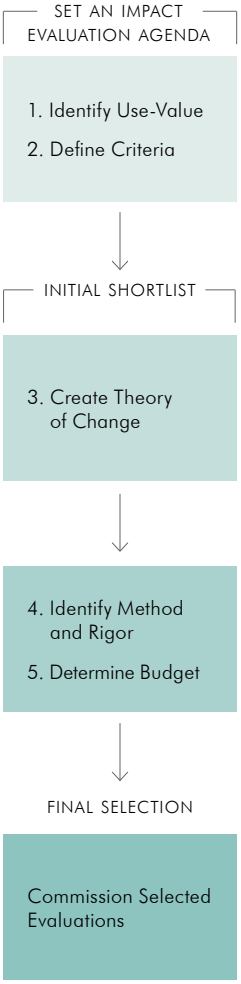
Section 2 – How to Evaluate: Identifying the Technical Feasibility of a Rigorous Impact Evaluation

Step 4: Identify the most appropriate impact evaluation method

Step 5: Create a detailed budget

Use Steps 4 and 5 to arrive at a final evaluation shortlist. Commission selected evaluations.

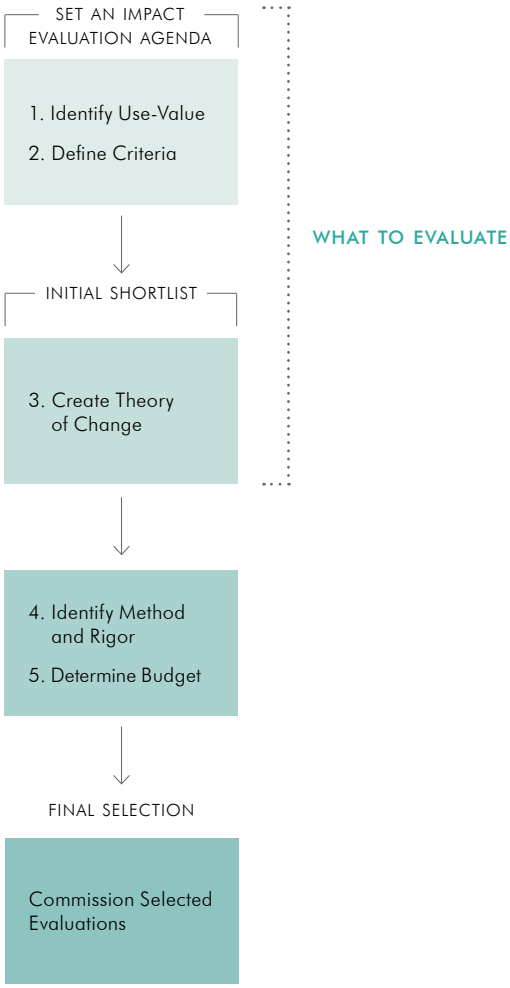
FIGURE 1. THE SELECTION PROCESS



1. WHAT TO EVALUATE

The What to Evaluate section lays out steps to set an organization's impact evaluation agenda. By assessing the extent to which prospective evaluations align with the evaluation agenda, an organization can create an initial shortlist of evaluations to consider. Following an initial shortlisting, the next step is to identify preliminary research questions by putting together a Theory of Change (ToC) for each program that a shortlisted evaluation proposes to evaluate. The first part of the What to Evaluate section introduces the impact evaluation agenda, while the last two sections provide guidance on narrowing the initial shortlist and formulating a Theory of Change.

FIGURE 2. STEPS IN THE SELECTION PROCESS COVERED IN THE WHAT TO EVALUATE SECTION



1. SETTING AN IMPACT EVALUATION AGENDA

BOX 2: STEPS IN SETTING AN IMPACT EVALUATION AGENDA

1. Identify who will use the results of the evaluation and which decisions the impact evaluation will inform.
2. Based on organizational priorities, set criteria for deciding which programs warrant an impact evaluation. Criteria include *program feasibility, degree of innovation, ease of scale and/or transfer, relevance to broader organization mandate, ability to inform a global debate, program size,* and *suitability of project timelines.*
3. Shortlist programs for an impact evaluation by ranking programs and their prospective evaluations against the usefulness of the evaluation and the set criteria.

IDENTIFYING USE

When considering whether to answer a research question throughan impact evaluation, an organization should first and foremost consider what decisions or actions will be informed by the information generated through the evaluation. To maximize the usefulness of a prospective evaluation, an organization must have a clear idea of how they will use the information from their evaluations. The term, **use-value**, is used to describe the extent that lessons from an evaluation will meet the needs of an organization. A few questions that may help identify use-value are:

- Will the information from an impact evaluation strengthen the program or policy being tested?
- Will the information from an impact evaluation identify activities for scale-up or transfer?
- Will the information from an impact evaluation inform future funding decisions?

Impact evaluation is a resource intensive tool and should be used wisely with careful consideration as to how the results will be used to inform future decision making. An organization must also think strategically about how to answer the particular research question of interest precisely and credibly. For example, if the information from an impact evaluation will be used to demonstrate a promising development innovation as tested and proven, a more rigorous evaluation that can demonstrate the causal impact of the activity with few assumptions will be most appropriate.

In some cases, it can be helpful to know that relationships or correlations exist between the activity and intended outcomes. In this case, non-experimental, or less credible, methods can be used to provide results that although inconclusive, may be suggestive and sufficient for decision-making. These less rigorous evaluations,

which will come with caveats and limitations, may be "good enough" in these contexts. Likewise, a ***process evaluation*** may be useful if the primary research questions focus less on outcomes and more on understanding how well a program is implemented.

DEFINING CRITERIA FOR DECIDING ON AN IMPACT EVALUATION

This part of the What to Evaluate section proposes various criteria to consider when setting an impact evaluation research agenda. These criteria reflect current thinking on strategic evaluation strategies.[10] It should be noted that these criteria are neither exhaustive nor static. Also, it is important to keep in mind that the criteria presented below are **un-weighted**. A criterion is inherently neither more nor less important than another. Instead, it is up to the organization to determine the weighting and relevance of the criteria presented below. Prioritizing criteria should be based on the types of research questions an organization would like to answer through impact evaluation. Under each criterion, some questions surrounding trade-offs have been provided to motivate the internal discussion. These can be used as a starting point to discuss the criterion in more depth and can also be used to refine the broad criteria.

BOX 3: SUGGESTED CRITERIA TO USE FOR DECIDING ON THE LIST OF ACTIVITIES TO ASSESS USING AN IMPACT EVALUATION

1. Program feasibility
2. Innovations
3. Scale and/or transfer
4. Relevance to an organization's broader development objective/focus
5. Ability to inform a global debate
6. Large projects
7. Program and evaluation timelines

1. Program feasibility

Ideally, an activity or program should be evaluated only when its feasibility has been proven. This means that the program has either been implemented on the ground for an extended period of time and has been shown to work as intended or it has been thoroughly vetted through the use of a needs assessment and/or a pilot. Having a thorough understanding of the feasibility of a program requires knowledge along several dimensions. If the program or activity has issues along any of these following dimensions, it may not be amenable to an impact evaluation.

i. Is the program or activity actually needed or demanded for by the intended beneficiaries?

ii. Is the innovation appropriate for the local setting? Does it have any inadvertent negative consequences or externalities?

iii. Does the program or activity have the necessary resources and do program staff have the skillsets required for proper implementation?

iv. Does the program have buy-in from all relevant stakeholders?

v. Has the program been proven effective in a controlled environment or under ideal circumstances?

vi. Will the results of an impact evaluation (either: positive, negative, mixed, or null) be palatable?

An example of a large scale evaluation of an unproven program comes from a randomized evaluation carried out in Rajasthan to test whether a system for monitoring nurses could improve nurse attendance in rural health sub-centers . The monitoring data, collected using time and date stamps, was transmitted to the district health administration and used to levy a punitive pay incentive system based on nurse attendance. The first result of the evaluation was positive; as long as the system of incentives were properly in place, they led to a dramatic improvement in attendance. However, as time went on, the nurses and the health administration colluded to undermine the incentive system by deliberately breaking the time and date stamp machines. The program stopped having a positive effect on nurse attendance. This example is a cautionary tale for organization's considering an evaluation of an unproven, untested program. Prior to an evaluation, an organization should feel confident that their program is not susceptible to such cases of failed implementation and unintended consequences.

2. Innovations

Does the evaluation identify, test, or scale innovations that have the potential to make large developmental impact but have not been rigorously tested? Innovations lead to substantial improvements in solving challenges by producing development outcomes more *effectively*, more *cheaply*, more *sustainably*, by reaching more *beneficiaries*, in a *shorter period of time*."[12] Given their importance, evaluations of innovations may rank high among an organization's priorities. However, given scarce resources, certain "kinds" of innovations can be prioritized. The following are types of innovations that can be evaluated:

i. Does the organization have a preference between evaluating innovations which are being "field tested" for the first time **(Stage 1 Innovations)** or evaluating the impact of innovations that have already shown to have an impact in Stage 1 **(Stage 2 Innovations)**?
   a. Choosing Stage 2 over 1 could provide evidence of innovations that are closer to being ready for scale.
   b. Choosing Stage 1 over 2 may be advisable when

information from IE can be used to strengthen the innovation before it is retested at a larger scale.

As an example, in 2001 J-PAL affiliate researchers conducted an evaluation to test an educational intervention known as Teaching at the Right Level (TaRL). TaRL is a pedagogical innovation that involves evaluating children using a simple assessment tool and then grouping students by ability level rather than by age or grade. The first evaluation of this approach tested a program in Gujarat, India that hired young women from local communities as tutors called Balsakhis, who would teach underperforming children in grades 2, 3, and 4 for two hours during the school day[13].

The Balsakhi program was found to be highly effective in increasing children's academic achievement and provided evidence supporting the teaching to the level of the child concept. The next step was to refine the implementation of TaRL to create a scalable, cost-effective model. To this end, J-PAL evaluated a second iteration of the TaRL model, the Learning Camps Evaluation. In the Learning Camp program, trained staff provided intensive short bursts of instruction in math and Hindi for 8-10 days at a time for up to 2 months. Students in grades 3-5 were grouped according to learning level and taught using level-appropriate materials tailored for each group[14].

The evaluations of the Balsakhi and Learning Camps programs illustrate the distinction between testing a Stage 1 and a Stage 2 innovation. The Balsakhi evaluation functioned as a proof of the TaRL concept and was a necessary Stage 1 innovation to allow for refining the TaRL model through the Learning Camps, Stage 2 Innovation. The Indian NGO, Pratham, is currently scaling the Learning Camps model in several states across India.

ii. Does the organization have a preference between evaluating the impact of innovations that all aim at achieving the same objective (say, improving the reading ability of children aged 6 to 10), or evaluating the impact of activities that have different objectives (say, early grade reading, preschool readiness, secondary enrollment, etc.)?
    a. Evaluating a variety of activities with a similar objective would allow comparisons between impact achieved and cost-effectiveness of these activities.
    b. Evaluating activities with different objectives would help build a diverse portfolio of projects that can be pushed for scale or transfer.

iii. Does the organization have a preference for evaluating the impact of certain "kinds" of innovations? For example, innovations that are strongly dependent on technology over those which do not use technology (or use it minimally)?

iv. Does the organization have a preference for evaluating the impact of innovative products over evaluations that target innovative service delivery mechanisms?

3. Scale and/or transfer

If an important objective of an organization is to support the diffusion and broad adoption of programs, then priority may be placed on evaluating activities which have a potential to be scaled and/or transferred. Conducting an impact evaluation of such activities allows an organization to learn more about (a) whether the activity has an impact, (b) the size of the impact, (c) which participants it affects the most, and (d) the cost-effectiveness of the activity. A *process evaluation*, in addition to the impact evaluation, will also provide information on the institutional and contextual factors essential to implementing the activity. Process evaluation information, in addition to a sound understanding of the theory underpinning a program, is extremely useful when assessing whether an activity can be scaled or transferred to regions somewhat dissimilar to the environment where the activity was first implemented and the impact was evaluated. As mentioned earlier, given scarce resources for impact evaluation, prioritizing impact evaluations of certain types of scale/transfer activities may be an efficient use of resources. To do so, the following should be considered:

i. Does the organization have a preference over evaluating the impact of an activity that has a potential to scale within a government set-up over those that can be scaled by NGOs?
    a. Depending on the scale-up strategy, having hard evidence on an activity that is conducive to be scaled through a particular system could be used in outreach activities.

ii. Does the organization have a preference for evaluating the impact of an activity whose "scope" is large over an activity whose "scope" is more limited (e.g. in which specific subsets of the vulnerable population may be affected or the delivery of the program may be too dependent on certain NGO specific characteristics)?
    a. Choosing to evaluate the impact of an activity that can impact a large subset of a vulnerable population while not being extremely contextual may increase an organization's ability to identify transferable programs.

---

[10] Paul Gertler et al., *Impact Evaluation in Practice*. Criteria to choose projects to be evaluated include projects that are innovative, replicable, strategically relevant, untested or influential. Bloomquist (2003) advises evaluation of projects which are strategically relevant to public policy, projects whose design can be influenced by evaluation results, and projects/policies whose evaluation contribute to improving the existing state of knowledge of that particular area.

[11] Banerjee, Duflo, and Glennerster, "Putting Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Healthcare System", 2007

[12] USAID, "India Country Development Cooperation Strategy", 9.

[13] Banerjee, Cole, Duflo, and Linden, "Remedying Education: Evidence from two randomized experiments in India", 2007

[14] Banerjee, Banerji, Duflo, Glennerster, and Khemani, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India", 2012

4. Relevance to an organization's broader development objectives/focus

An organization may also want to consider evaluating the impact of projects that are relevant to its broader objectives/focus. When considering this, the following questions should be answered:

i. What is the broader development objective/focus in an organization's development strategy?

ii. Are there any specific objectives that the organization would like to inform using their impact evaluations?

5. Ability to inform a global debate

At any given moment, certain topics in development are more salient than others. If interest is expressed in evaluating the impact of activities that will inform global debates, evaluations that provide critical and timely evidence on a prominent or polarizing issue should be given preference. Salient impact evaluations have the added benefit of being eligible for greater funding opportunities.

For example, despite consensus on the importance of subsidizing preventative health products, there has been a long-running debate on what proportion of the cost to subsidize. To add evidence to this debate, Pascaline Dupas and Jessica Cohen conducted an impact evaluation on a program that delivered insecticidal bed nets in Kenya. The researchers found that charging even small prices considerably decreased demand of bed nets and that women who paid for nets were no more likely to use them as those who received nets for free.[15] This evidence played a role in motivating organizations to reconsider their policies of charging for preventative health products and instead opt for an abolition of user fees.

6. Large projects

"Large" projects should also be considered for impact evaluations. Large projects typically have huge outlays or reach and have a long lifespan. Impact evaluations of such projects are important to establish whether they delivered the results intended and are therefore worth continuing. Although learning and accountability findings tend to have higher payoffs for larger programs, it may be more difficult to integrate such findings back into program design. This is especially true for programs that are so large and important that they become "sacred cows" for an organization. If termination of the program is unlikely to be politically feasible, perhaps one should reconsider expending resources on an evaluation. It is also important to realize that larger programs usually involve more stakeholders and higher visibility, which makes evaluations more prone to political resistance and other logistical challenges.

An example of the challenges of evaluating large projects comes from a J-PAL evaluation of a State wide program in Andhra Pradesh to use biometric identification smartcards as a way to transfer government benefits to the poor[16]. The program officially began in 2006, but due to several logistical challenges, the Government restarted the program in 2010 to eight districts where biometrics smartcards had yet to be rolled out. These eight districts had a combined rural population of almost 19 million people. Despite slow program roll-out due to the size and complexity of the project, an evaluation found that the biometric smartcards led to reduced payment time to beneficiaries and lowered rates of leakages.

7. Program and evaluation timelines

Program and evaluation timelines have an implication on the timeliness of results, the flexibility of program adaptation, as well as the methodology of the evaluation.

i. **Timeliness of Evaluation Results:** Certain programs can be expected to impact outcomes only after a significant amount of time, sometimes long after the program ends. Evaluations of such programs will require a long time horizon in order to provide meaningful results. An organization should decide whether they would prefer evaluating programs that yield an impact in the short-term over these longer-term programs.

   Early childhood development (ECD) programs are an example of a type of intervention that tries to affect long term outcomes. On rare occasions researchers are able to rigorously evaluate the long term effects of ECD programs. One example of this occurred as a follow up evaluation of a program introduced in Kingston, Jamaica in 1986[17]. The original program randomly assigned a cohort of 127 stunted children into either a psychosocial stimulation intervention, a nutrition intervention, both interventions, or none. Twenty years later, researchers re-interviewed 105 of the original 127 study participants and found that individuals who received the psychosocial intervention earned, on average, 25 percent more income than stunted children who did not receive the stimulation. While most organizations would scoff at the idea of waiting ten to twenty years for evaluation results, a long term evaluation might be more amenable for others.

ii. **Flexibility of Program Adaptation:** Impact evaluations can be performed on programs that are at different stages in their life-cycle. Evaluations of new or nascent programs have the benefit of easier integration into the program's implementation. However, new programs are often subject to frequent changes, and adherence to an evaluation method may limit implementers' ability to evolve or tinker the program as needed. An organization should consider this trade-off and look to prioritize evaluations that are of sufficiently established programs.

iii. **Methodology:** Rigorous impact evaluations establish causality between a program and an outcome by incorporating an accurate estimate for the counterfactual (what would have happened without the program).[18] Ultimately, the rigor of an evaluation depends largely on how well the counterfactual is

defined or created through a comparison group. The most statistically unbiased method of creating a counterfactual is a randomized evaluation that randomly assigns participants to a group exposed to the intervention and a group that is not. Such random assignment can only take place **before** the program is implemented.

While ex post evaluations of programs have the potential to avoid selection bias,[19] and can be appropriate for answering certain types of questions, it is usually far more difficult than for randomized evaluations (and sometimes impossible to prove that they are unbiased.[20] Rigorous evaluations incorporating counterfactuals and randomization are therefore extremely time dependent. If an organization prioritizes rigorous evaluations that generate a high degree of confidence in the accuracy of the impact estimate, then activities that have yet to be implemented or are about to be scaled should be given precedence so that a randomized evaluation can be feasible.

Combinations of Criteria: Careful consideration must be made on whether certain combinations of the main criteria should be listed and prioritized. For example, an organization may want to indicate that activities that meet all of the following criteria should be included in the list for consideration: (1) are feasible, (2) are innovative, (3) relate to using technology in the public health care system, (4) can be scaled within the government setup and (5) haven't been implemented yet. Weighting of the criteria to match organizational preferences can also be undertaken to emphasize certain criteria over others.

II. SETTING AN INITIAL SHORTLIST

After prioritizing the above criteria, an organization has enough information to do an initial shortlisting of evaluations. The shortlisting procedure should be done systematically by documenting each prospective evaluations' use-value and extent to which it satisfies the prioritized criteria. Ideally, an organization should decide on a ranking strategy, either quantitative or qualitative, develop a matrix to score evaluations, and determine the cut-off for further consideration. Collaboration at this stage is key. Setting organizational priorities and an impact evaluation agenda cannot be performed by one individual and neither can the shortlisting process. It is important to have diverse perspectives from various staff, such as program managers, senior staff, and monitoring and evaluation leads.

The initial shortlisting process is intended to be a cursory weeding out of evaluations that are least strategic. The process is not intended to give an organization a clear understanding of the evaluations to commission. To arrive at that final selection, each candidate evaluation must be compared across more detailed technical information such as the rigor of the chosen evaluation method, and the financial cost of the evaluation. Conducting an initial shortlist now can save an organization from unnecessarily expending time and effort determining this information for low priority evaluations.

Before moving to the technical side of the selection process, a theory of change should be developed for each program that has cleared the initial shortlist, if one hasn't been prepared already.

III. DRAFTING A THEORY OF CHANGE

A theory of change is a way of unpacking the black box between an intervention and an outcome by describing the theory behind how an intervention delivers the intended results. This heuristic sets out a causal logic flow of how and why a program or policy will reach its intended outcomes. A theory of change includes the following:

- A listing of the outcomes to be achieved (e.g. higher primary school attendance for girls)
- What program will induce those outcomes (e.g. providing free bicycles to all girls of primary school age)
- What pathways theoretically lead to those outcomes (e.g. lowered opportunity cost of school travel)
- What assumptions are associated with each link of the causal chain (e.g. bicycles wont brake down)

A theory of change is known as the blueprint for a program's design, and therefore, the blueprint for any evaluation of that program, since it not only details what outcomes the program will achieve and how, but also guides what data an evaluation will collect. To test whether each step in the causal chain between program and outcome was achieved, an evaluation should collect information related to the inputs, outputs, outcomes, long-term goals, and assumptions of the program. In this manner, if an evaluation failed to detect a positive impact of a program on our main outcome, one can use the information collected to see where along the theory of change the program failed. Collecting the right information will allow you to conclusively say whether a program failed due to an implementation error (inputs did not get turned into outputs), or an error of program theory (outputs did not result in outcomes).

Following the initial shortlist of evaluations to consider, organizations should draft theories of change for each program that is subject to

---

[15] Cohen, J., and Dupas, P., "Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment", 2010.

[16] Muralidharan, K., Niehaus, P., Sukhtankar, S. "Building State Capacity: Evidence from Biometric Smartcards in India", 2016.

[17] Gertler et al., "Labor market returns to an early childhood stimulation intervention in Jamaica", 2014.

[18] The technical side of the toolkit provides a definition of this term.

[19] See technical side of toolkit for more details.

[20] Murnane and Willett, *Methods Matter: Improving Causal Inference in Educational and Social Science Research.*

a shortlisted evaluation. Organizations should use each program's theory of change to develop a set of preliminary research questions for the potential impact evaluation. Questions should reflect the interests of the stakeholders who will use the information, and may have broader questions about whether an activity of innovation is working on average, or rather if it is working for only a particular subgroup. These questions can be further defined by an evaluator, once one is brought on board. An example of a simplified theory of change is illustrated in **Appendix A**.

2. HOW TO EVALUATE

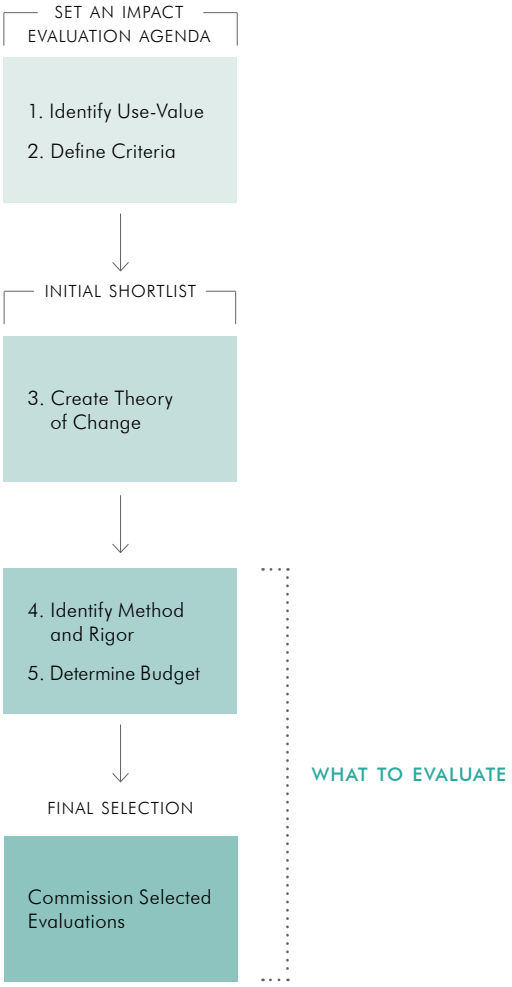
Setting an impact evaluation agenda and identifying an evaluation shortlist is a necessary initial step in determining which evaluations to conduct. The next step is to identify the technical feasibility of shortlisted impact evaluations – can the evaluation produce a credible estimate of the program's impact? Answering this question depends on the rigor of the impact evaluation and the impact evaluation method used.

Rigor of the Evaluation: The most credible evaluations are those which can establish that the intervention[21] directly resulted in the outcome, rather than establish the degree and direction of a relationship (correlation) between the intervention and the outcomes. Ultimately, the rigor of the evaluation depends largely on how the comparison group mimics the counterfactual--that is, how well a group of non-participants approximates the unobservable state of the world of what would have happened to the program recipients had they not received the program.

Selecting the Appropriate Impact Evaluation Method: There are a variety of methods that can be used to establish a causal impact of a program on an outcome of interest. The methods differ by the ways they create a counterfactual group. As we will see, the most rigorous method employs randomization to select who gets the program and who doesn't. Other methods make qualifying assumptions that only if satisfied, creates a counterfactual group that is as good as randomly selected. Unfortunately, it us often impossible to test or otherwise know with any level of confidence whether these necessary assumptions are indeed satisfied.

The goal of the How to Evaluate section is to help users navigate practical considerations regarding both the rigor of the evaluation and selecting the appropriate impact evaluation method. The section is organized to closely follow the second half of the Selection Process (see Figure 3). The first part of the How to Evaluate section presents a guide to impact evaluation and evaluation design. The **Table of Methodologies** and the **Impact Evaluation Method Tree** introduce the various impact evaluation methodologies, while **Appendix C** goes over each method in more detail, including information on the method's major assumptions and practical limitations. Once an impact evaluation method has been selected, the final step is to put together a detailed budget, make the final selection of evaluations, and commission selected evaluations. The final three parts of this section goes over these steps in turn.

FIGURE 3. STEPS IN THE SELECTION PROCESS COVERED IN THE HOW TO EVALUATE SECTION



1. GUIDE TO IMPACT EVALUATION

BOX 4: LEARNING OBJECTIVES OF THE GUIDE TO IMPACT EVALUATION

- To distinguish how the terms "impact" and "causality" are defined in the context of impact evaluation.
- To understand selection bias, why it is a problem, and how RCTs are a particularly useful tool to avoid selection bias.
- To recognize why inferring causality from observational data/non-experimental methods can be misleading.
- To select the most credible impact evaluation method given a set of practical constraints.

ABCs of Impact Evaluation: Before examining individual impact evaluation methodologies, we will first define basic terms and concepts that are necessary for discussing impact evaluation. As we move through these definitions, we will use the term "program" to broadly cover any activity, innovation, policy or a component of a larger project. Additionally, we will use "participant" to refer to any unit receiving a program, which may (depending on the program) be an individual, a household or an entire community, school, block, or other entity.

BOX 5: KEY DEFINITIONS

Causality: Isolating the effect of the program and the program alone, independent of any other intervening factors, on an outcome or outcomes of interest.

Impact: Impact is defined as the comparison between (1) an outcome of interest measured at an appropriate time after a program has been introduced, and (2) the outcome at that same point in time had the program not been introduced.

Counterfactual: The counterfactual represents the state of the world that program participants would have experienced in the absence of the program (i.e., had they not participated in the program). Note that this is, by construction, a state of the world that is never observable since a person is only observed in one of the two states.

Sample Frame: A term used in the data collection aspect of an evaluation, which refers to the group within the target population that can be accessed for data collection. Sampling frames are usually constructed through an available list or map which includes all accessible units for data collection. The sample for study is selected from the sample frame, ideally in a manner that makes the sample representative of the sampling frame (e.g. through random selection).

Comparison Group: A group that is used for comparison in an impact evaluation, which stands as a proxy for the counterfactual.

Selection bias: A problem that occurs when program participants are compared to non-participants to measure impact, but differ from nonparticipants in ways that cannot be observed or measured, and these differences can affect both the decision to participate (or selection for participation) and the outcome of interest.

[21] Intervention here can mean program or activity.

Causality

Isolating the effect of the program and the program alone, independent of any other intervening factors, on an outcome or outcomes of interest. Although we often use terms such as cause and effect on a day to day basis, when establishing causality in the context of impact evaluation, one must be careful. Claiming causality involves empirically establishing to what extent a program, and that program alone, drove changes in a particular outcome of interest. We use impact evaluation to rule out the possibility that any other factors, other than the program we are evaluating, are the reason for these changes.[22]

Impact

Impact is defined as the comparison between (1) an outcome of interest measured at an appropriate time after a program has been introduced, and (2) the same outcome at the same point in time had the program not been introduced.[23] Impact can be positive, negative, mixed, null, or undetectable.

Causality

Isolating the effect of the program and the program alone, independent of any other intervening factors, on an outcome or outcomes of interest. Although we often use terms such as cause and effect on a day to day basis, when establishing causality in the context of impact evaluation, one must be careful. Claiming causality involves empirically establishing to what extent a program, and that program alone, drove changes in a particular outcome of interest. We use impact evaluation to rule out the possibility that any other factors, other than the program we are evaluating, are the reason for these changes.[22]

Impact

Impact is defined as the comparison between (1) an outcome of interest measured at an appropriate time after a program has been introduced, and (2) the same outcome at the same point in time had the program not been introduced.[23] Impact can be positive, negative, mixed, null, or undetectable.

For example, let's say we were interested in the impact of iron supplements on anemia prevalence in adolescent girls. A program was run in two districts, Jhajjar and Sirsa, in the state of Haryana from June 2011 – June 2012 that provided adolescent girls ages 14–18 with iron supplements. The ideal, yet impossible, way to measure the impact of the iron supplement program would be to compare the hemoglobin level in June 2012 for a girl who received the program, relative to her hemoglobin level (in June 2012) had she not been part of the program. This would be expressed as:

$$\text{Impact}_i = H_{P2012,i} - H_{NP2012,i}$$

where $H_{P2012,i}$ is the hemoglobin level of an adolescent girl i if she was part of the program (measured in June 2012, after the program) and $H_{NP2012,i}$ is the hemoglobin level of the same girl in

2012, had she not received the program. Both $H_{P2012,i}$ and $H_{NP2012,i}$ are referred to as potential outcomes; however, the fundamental feature of impact evaluation is that for each person in the study we only observe one of these two potential outcomes.

For a girl who is in the program, we observe what her hemoglobin levels are in June 2012 after she has been part of the program ($H_{P2012,i}$), but clearly we then cannot observe her hemoglobin level in 2012 without the program. That is to say, for a program participant we cannot observe $H_{NP2012,i}$. The subscript i in the equation above refers to an individual, in this case an adolescent girl who was part of the program. To make this clearer, say individual i's name was Zara. To determine the impact of the iron supplement program for Zara, you would have to subtract the level of Zara's hemoglobin concentration in her blood sometime after she receives the program, say June 2012, with Zara's hemoglobin concentration at the same point in time, again in June 2012, had she not received the iron supplement program. If you were able to compare Zara's hemoglobin level in June 2012 with the program to Zara's hemoglobin level in June 2012 without the program, you would know that any difference in hemoglobin levels would have to be due to the program itself. No characteristics, other than the program could explain this difference. Of course, the problem is that it is impossible to observe Zara both with and without the iron supplement program.

The question then is how we can learn about $Impact_i$ when one of the terms on the right is not observed. How can we learn the impact of the hemoglobin program for Zara, when we can't observe how she would have fared without the program? This is the fundamental challenge of assessing causal impact. Impact evaluation methods attempt to answer this question by providing an estimate for the unobserved counterfactual.

Counterfactual

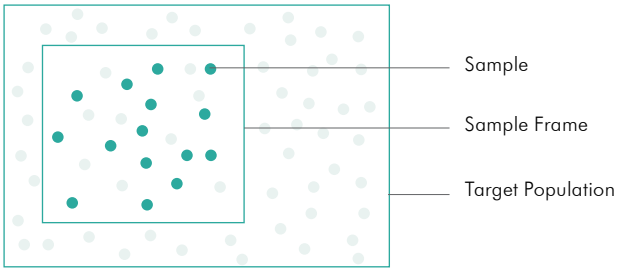
The counterfactual represents the state of the world that program participants would have experienced in the absence of the program (i.e. had they not participated in the program).[24] The true counterfactual is never observed, so we must create a proxy using a comparison group (see definition of *comparison group* below). Recall, in the iron supplement example, that the counterfactual measure is the hemoglobin level for adolescent girl program participants if they had not been exposed to the iron supplement program.

Sample Frame

A term used in the data collection aspect of an evaluation, which refers to the group within the target population which we have defined as the sample of interest in the target population. Sampling frames are usually constructed through an available list or map which includes all accessible units for data collection (e.g. households, students, patients, employees).[25] If those who can be sampled are different from the target population of interest, the results may be specific only to that group, and not apply to the whole target. Often times, practical constraints of available lists are a part of what drives this bias.

For example, in the case of the iron supplement program, let us say the target population was all adolescent girls 14–18 years of age in Jhajjar and Sirsa, and the program used government school enrolment lists to construct the sample frame (and from this sample frame drew a sample of girls aged 14–18). Note that girls aged 14–18 who did not attend government schools would be left out of the sampling frame. Thus, girls who either went to private school or had dropped out of school would not be included, and the sample frame (and any sample we drew from it) would not be representative of the target population of interest.

FIGURE 4. TARGET POPULATION, SAMPLE FRAME, AND SAMPLE



Comparison Group

A group used for comparison in an impact evaluation by standing as a best approximation for the counterfactual. The best comparison group is one that perfectly mimics the counterfactual. A perfect comparison group is impossible to achieve, i.e. we cannot find a non-participant in the comparison group that is exactly identical to each individual in the treatment group.

However, a properly constructed comparison group can be as similar as possible to the counterfactual not at the individual level, but on **average** and general composition at the group level. In practice then, a "good" comparison group should be, on average, similar to the recipient group across a set of characteristics that are related to the outcome of interest (usually this is checked using statistical tests on differences between average levels of characteristics between comparison and recipient groups before the program is implemented).

---

[22] Paul Gertler et al., *Impact Evaluation in Practice*, 34.

[23] J-PAL Executive Education Course, "Why Randomize?"

[24] J-PAL Executive Education Course, "Why Randomize?"

[25] Bamberger, Rugh, and Mabry, *Real World Evaluation*, 252

[26] World Bank "Evaluation Designs."

[27] Baker, Judy, *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*, & J-PAL Executive Education Course. *Case 2: Learn to Read Evaluations.*

We use the average value of the outcome in the comparison group as our proxy for the counterfactual. That is to say, we compute the average hemoglobin level in our comparison group and *assert* that this is the average hemoglobin level the treatment group would have experienced in the absence of the program. Note that this is an assertion that is only convincing to the extent that our comparison group is comparable to the treatment group in relevant ways.

Differences between comparison and treatment groups on characteristics that do not affect whether an individual receives the program nor affect the outcome of interest are non-problematic. For instance, in our iron supplement example, it would be inconsequential that a comparison group of non-participants were on average more likely to prefer the color blue than participants. It would, however, be a problem if say the control group were more likely to wash their hands than the treatment group. This is because individuals who are more likely to wash their hands are less likely to have parasitic infections such as hookworm, and hookworm causes a decrease in hemoglobin levels. A theory of change can be a useful guide in determining consequential characteristics that must be the same on average between comparison and program recipient groups. As theory can't detect all relationships between complex phenomena, it is always more reassuring to have groups be similar even on characteristics that are construed as inconsequential.

BOX 6: WHAT MAKES A GOOD COMPARISON GROUP?

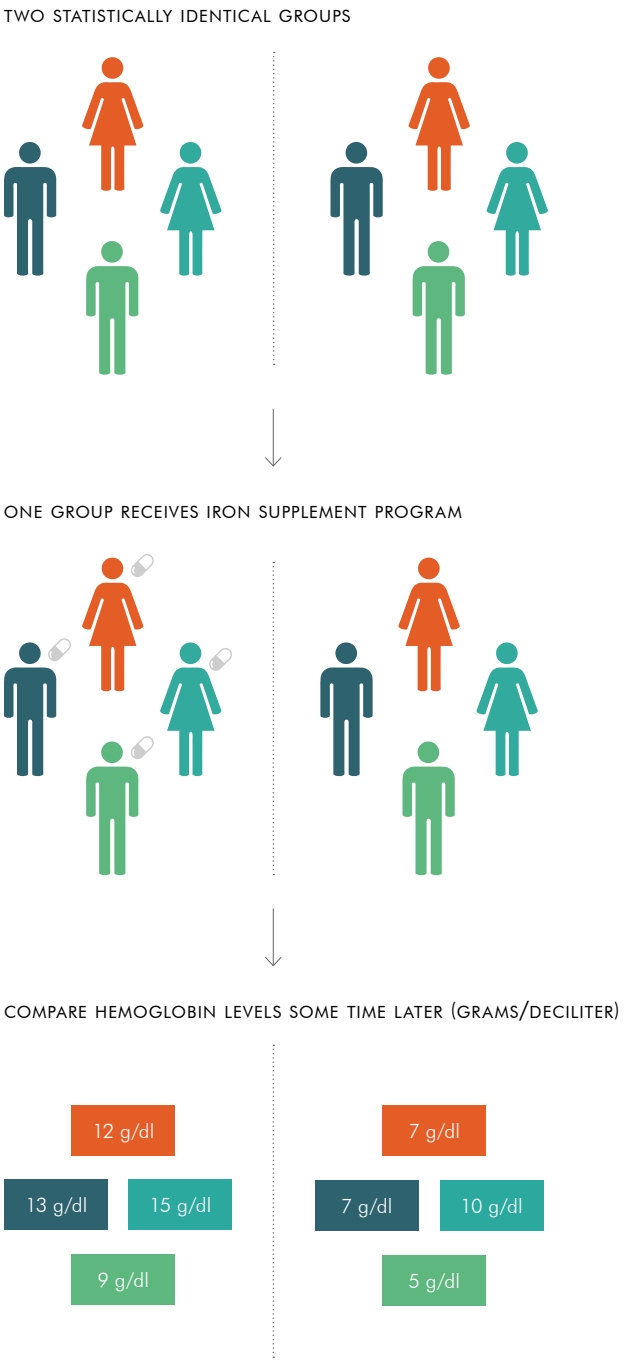
- The reason participants received the program is random (or close to random), and not due to voluntary selection or certain characteristics that make participants different from non-participants.
- Two groups (comparison and program) are statistically identical before the program is implemented.
- The two groups are experiencing the same environmental factors unrelated to the program during the time that the program is administered.

For example, if we identified girls 14–18 in the same districts who went to private schools, and therefore didn't receive the program, as the comparison group, this would not be appropriate. Girls who went to private school are likely to be different from those who attend government schools for a number of reasons and these differences may be directly related to the outcome we are interested in (hemoglobin levels).

Irrespective of how comparison groups are formed, we can never be certain that *unobservable characteristics* (factors that cannot easily be measured or quantified, such as motivation, ethical values, etc.) are distributed equally across both groups. However, a reasonable assumption is that the more similar both groups are across observable characteristics, the more likely unobservable characteristics are also equally distributed. Unobservable differences are a major concern when choosing a comparison group because these differences may also be related to other important factors that need to be similar in order for the groups to be comparable.

Since we are unable to measure unobservable characteristics, our next best approach is to construct two groups and test whether they are statistically identical across observable characteristics. If the two groups are identical on select observable characteristics, this lends evidence that the two groups are similar, and therefore comparable. Ideally, the two groups should also be exposed to the same environmental factors (socioeconomic, political, policy-related, and natural conditions) over the time the program is administered (with the exception of the program itself). If both of these criteria (i.e. statistically identical before the program and having the same environmental factors over time) are achieved, then this is the closest we can get to a comparison group that perfectly mimics the counterfactual. Hence, any differences observed between the comparison group and the group receiving the program, after the program has been implemented, can be attributed to the program itself, and not any other intervening factor or pre-existing characteristics that might be the true driver of these differences.

FIGURE 5. IMPACT OF THE IRON-SUPPLEMENT PROGRAM



Selection bias

A problem that occurs when program participants differ from nonparticipants in ways that cannot be observed or measured, and these differences affect both the decision to participate (and/or selection for participation) and the outcome of interest.[26] Selection bias leads to inaccurate impact estimates because the two groups that are being compared are no longer similar on all characteristics except for their exposure to the program. When selection bias is an issue, the comparison group no longer credibly mimics the counterfactual.[27]

If girls attending private schools were used as a comparison group for the iron supplement program, selection bias would be an issue because girls who received the program would differ from those who did not on a very important characteristic (attending government versus private schools), a factor that is also likely related to health outcomes (for instance if girls in private schools had very different diets from girls in public schools and these diets affected hemoglobin levels).

II. EVALUATION DESIGN

BOX 7: TYPOLOGY OF IMPACT EVALUATIONS

Experiment: A study in which a program is deliberately introduced to observe its effects.

Randomized experiment: An experiment in which units are assigned to receive the program by a random process such as the toss of a coin or a table of random numbers.

Quasi-experiment: An experiment in which units are not assigned to the program using deliberate randomization but instead a process that is "almost" random so that treatment assignment is "as if" randomly assigned.

Natural Experiment: A natural experiment occurs when program recipients and non-recipients are determined by nature or by other factors outside the control of researchers, yet this external process governing treatment assignment can be argued to be random.

Observational study: Usually synonymous with a non-experimental or correlation study: a study that simply observes the size and the direction of a relationship among variables.

There are two types of study designs that are typically used for an impact evaluation: **experimental** and **non-experimental**. The primary distinction between these two categories is the rule that assigns whether a group (or individual) receives the program. If assignment is manipulated so that an evaluator selects who receives the program and who does not, then the method is considered experimental. If program assignment is not manipulated in this way, then the study is non-experimental.

The predominant type of experimental study is the randomized experiment, where an evaluator uses the rule of random assignment to select who receives the program and who doesn't. There are many types of non-experimental studies. At one end, closest to a randomized experiment, are "quasi" or "natural" experiments where assignment is "as-if" randomly assigned even though no conscious randomization was done (for instance, an earthquake happens to affect some "participants" but not others, and the research assumes that the earthquake was assigned as-if randomly). In a quasi-experiment, the researcher must be able to

argue that treatment assignment was "as-if" randomly assigned in a manner uncorrelated with individual characteristics.[28]

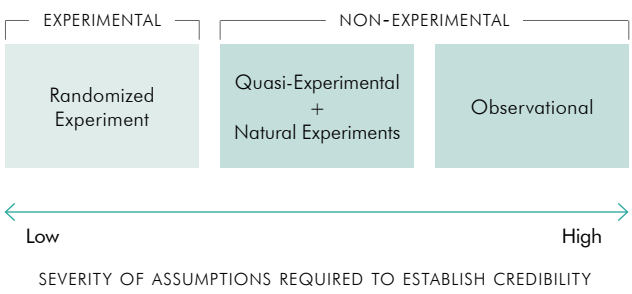
Then, there are studies where treatment assignment is not randomly or as-if randomly assigned, but instead chosen by some process that remains opaque to the researcher. The central distinction between these designs is that with experimental designs (and to some extent with quasi-experimental designs) the comparison group frees us from the primary concern of selection bias, whereas with most observational studies the comparison group does not share this same feature. This distinction has important implications for the credibility of the corresponding impact estimates.

Sometimes, observational data might be used to help establish a relationship between receiving a program and the outcome of interest, but does not offer enough evidence that the program is the singular cause of the changes in outcomes that we are observing. These designs, often called a "correlation study," merely offer information on the general size and direction of relationships among factors, but cannot be used to credibly establish causality, and therefore are not considered impact evaluations.

Among the community of academics and evaluators, many agree that the most compelling evidence for claiming that a program caused a change in a certain outcome of interest typically comes from randomized evaluations. "When randomization is not possible, quasi-experimental techniques are used to create counterfactuals that aim for statistical equivalence with the treatment group."

Among the community of academics and evaluators, many agree that the most compelling evidence for claiming that a program caused a change in a certain outcome of interest typically comes from randomized evaluations.[29] "When randomization is not possible, quasi-experimental techniques are used to create counterfactuals that aim for statistical equivalence with the treatment group."[30]

FIGURE 6. QUANTITATIVE EVALUATION METHODS



QUANTITATIVE EVALUATION METHODS

There are many different quantitative evaluation methods that might be used by an evaluator to estimate the impact of a program on an outcome of interest (see the **Table of Methodologies**, the **Impact Evaluation Method Tree**, and **Appendix C**, which gives a detailed description of each method). The diagram above gives a broad picture of the spectrum of evaluation designs that might be used, and the stringency of assumptions required for a causal claim to be credible. It is important to recognize that every method comes with a set of assumptions, but each method aims to construct a credible comparison group by coming as close to estimating the counterfactual as is possible. There are many ways of constructing comparison groups, some more credible than others. Depending on how a method is used in the specific context of program implementation, different methods will fall in various places along the above spectrum. Here are three examples of how a comparison group might be constructed in the case of the iron supplement program:

- Researchers identified girls 14–18 in the same districts who went to private schools and therefore didn't receive the program. Girls who went to private school are likely to be different from those who attend government schools for a number of reasons and these differences may be directly related to the outcome we are interested in (hemoglobin levels). For instance, girls in private schools may on average have better (or worse) diets, which in turn affects their hemoglobin levels. If we use them as a comparison group, the differences in hemoglobin levels between the two groups would in part be driven by the iron pill program, but also by the differences in diets between the two groups.

- Researchers tested hemoglobin levels for program participants before the program started (early in 2011) and then compared the average baseline hemoglobin level to the average hemoglobin level from the 2012 test. This is sometimes called a "before-after" or a "pre-post" design. While it may be convincing in some situations, the primary problem with the design is that many things may have changed in the intervening period that could affect hemoglobin levels and our simple comparison confounds the effect of the program and these other changes. For instance, the state may have initiated a subsidized food scheme (to keep with the nutrition example) or a malaria reduction program after the program started in 2011. Both of these could plausibly affect hemoglobin levels. Simply comparing 2012 to 2011 levels would mix up the effects of the pill supplementation program with that of the subsidized food (or malaria reduction program).

- Researchers sample a set of girls public schools from the sampling frame above and randomly assign the program to girls in one-half of the schools in the sample, while girls from the remaining other half of schools constitute the comparison group and do not receive any supplements (in this experimental context, the group is often called the control group).

Of the three comparison groups derived above, the third example using random assignment requires the least number of assumptions

to meet the criteria of a credibly drawn comparison group (though it is the hardest to put into practice).

Process Evaluation

If a credible impact estimate is not possible, but one still wants to understand more about how a program is working, then process evaluations may also an option. Process evaluation[31] tends to focus on questions that address how well a particular program is being implemented. If resources, time or other practical constraints are a barrier to a good impact evaluation, process evaluations are recommended.[32]

Mixed Methods

The table below provides information on quantitative evaluation methodologies. However, using qualitative methods in complement to quantitative methods may allow for richer interpretation of the results of the analysis. For example, suppose a randomized evaluation conducted for the iron supplement program concluded that the program had no impact, meaning there was no difference in hemoglobin levels between program recipients and non-recipients. Stakeholders will be interested in knowing why no impact was detected. To answer this question, one needs much more information than just hemoglobin measurements. What if the girls collected theiron supplements from the schools but shared them with other family members? A focus group discussion on use of these supplements may be able to throw light on such leakages.

---

[28] Concretely, consider a job-training program that is oversubscribed and officials decide to allocate individuals to the program by conducting a lottery. In a study looking at the effects of this program, one can credibly compare lottery winners to lottery losers since these two groups were created in a manner very close to what a researcher running a randomized experiment would do. A classic, historical, example of a quasi-experiment is Snow's (Freedman 1991) studies that demonstrated the water-borne nature of cholera. The key assumption in his analysis was that households in London chose their water suppliers without regard to whether these suppliers drew their water above or below a particular point on the Thames so that one could in effect treat the households as if their water suppliers had been provided randomly.

[29] Murnane and Willett, Methods Matter: *Improving Causal Inference in Educational and Social Science Research*, 30.

[30] World Bank Group, "Impact Evaluations: Relevance and Effectiveness."

[31] The USAID Evaluation policy notes that, "Performance evaluation focuses on descriptive and normative questions: what a particular project or program has achieved (either at an intermediate point in execution or at the conclusion of an implementation period); how it is being implemented; how it is perceived and valued; whether expected results are occurring; and other questions that are pertinent to program design, management and operational decision making." It also states that "Performance monitoring of changes in performance indicators reveals whether desired results are occurring and whether implementation is on track." The process evaluation therefore incorporates many elements of performance evaluation.

[32] Community Interventions for Health website: http://www.oxha.org/cih_manual/index.php/process-evaluation

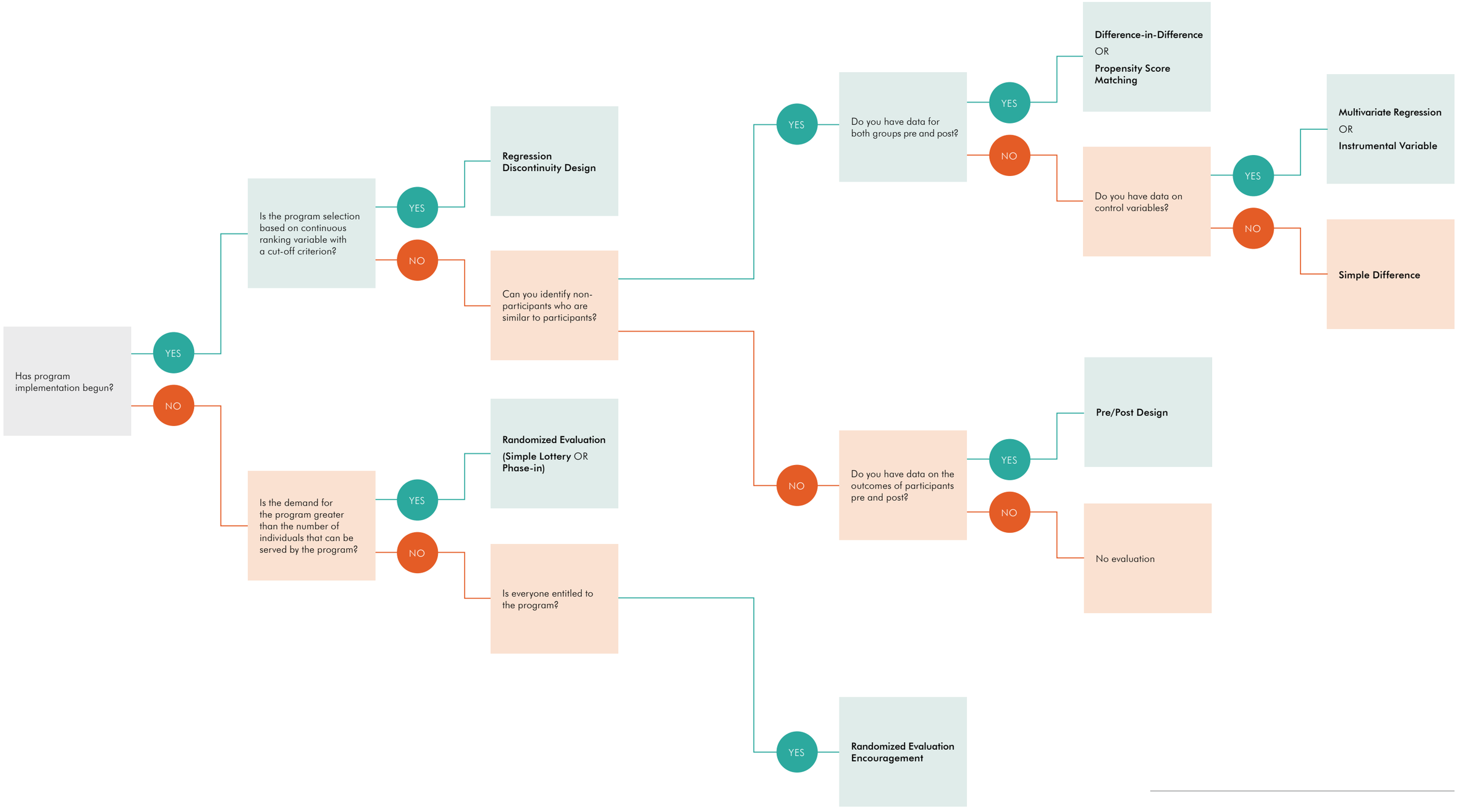
TABLE OF METHODOLOGIES[33]

METHODOLOGY	DESCRIPTION	WHO IS IN THE COMPARISON GROUP?	REQUIRED ASSUMPTIONS	REQUIRED DATA? WHAT TO COLLECT?	WHEN TO COLLECT?	WHO TO COLLECT?	EXAMPLE
Pre-Post	Individuals who received the program are compared before and after the program	Individuals who participated in the program themselves	The program was the only factor influencing any changes in the measured outcome over time.	Outcome variables	Baseline Endline	Participants	The average hemoglobin level of adolescent girls receiving the program in 2012 minus the average hemoglobin level of adolescent girls receiving the program in 2011 (before the program was run).
Simple Difference	Individuals who received the program are compared with individuals who did not receive the program, after the program is completed.	Individuals who didn't participate in the program	1. Nonparticipants are identical to participants except for program participation 2. Nonparticipants were equally likely to enter program before it started	Outcome variables	Endline	Participants Comparison group	The average hemoglobin level of adolescent girls receiving the program in 2012 minus the average hemoglobin level of adolescent girls who didn't receive the program in 2012.
Difference-in-difference	Measure changes over time of program participants relative to the changes overtime of comparison group (nonparticipants)	Individuals who didn't participate in the program	If the program didn't exist, the two groups would have had identical trajectories over this period.	Outcome variables	Baseline Endline	Participants Comparison group	The difference in the change over time in average hemoglobin levels of adolescent girls receiving the program and adolescent girls in a neighboring district who didn't receive the program.
Multivariate Regression	Individuals who received the program are compared with those who did not, and other factors that might explain differences in the outcomes are accounted for using statistical adjustments (referred to as "controlling")	Individuals who didn't participate in the program (controlling for other factors)	The factors that were excluded do not bias results because they are: a. uncorrelated with the outcome b. do not differ between participants and non-participants	Outcome variables Control variables	Endline*	Participants Comparison group	The average hemoglobin level of adolescent girls receiving the program in 2012 minus the average hemoglobin level of adolescent girls who didn't receive the program in 2012, controlling for additional factors such as socio-economic status, type of school attended, etc.
Statistical Matching	Individuals who received treatment are compared with a comparison group that is constructed by finding non-participating individuals who have similar characteristics to the treated individuals	1. *Exact matching:* For each participant matched with at least one non-participant who is identical on selected characteristics related to the outcome of interest 2. *Propensity score matching:* Participants matched with non-participants who have a mix of characteristics which predict that they would be as likely to participate as participants	The factors that were excluded do not bias results because they are: a. uncorrelated with the outcome b. do not differ between participants and non-participants	Outcome variables Control variables	Endline*	Participants Comparison group	Girls receiving the program are matched with girls who didn't receive the program (perhaps in a neighboring district) based either on individual characteristics related to health outcomes, or based on their likelihood to participate in the program. Hemoglobin levels are compared between matched individuals.
Regression Discontinuity Design	Individuals are ranked based on specific, measureable criteria. There is some cut-off that determines whether an individual is eligible to participate. Participants are then compared to those who missed the cut-off (comparison group)	Individuals who are close to the cut-off, but fall on the "wrong" side of that cut-off, and therefore do not get the program	1. The differences between individuals directly below and directly above the cut-off score (besides program participation) will not bias results 2. The cut-off criteria are strictly adhered to	Outcome variables Measure on criteria Control variables	Endline*	Participants Comparison group	Girls receiving the program who just made eligibility (turned 14 by June 2012) are compared with girls who didn't make the cut-off (i.e., turned 14 in July or August of 2012).
Instrumental Variables	Participation can be predicted by an incidental (almost random) factor, or "instrumental" variable, that is uncorrelated with the outcome other than the fact that it predicts participation (and participation affects the outcome)	Individuals who, because of this close to random factor, are predicted not to participate and (possibly as a result) did not participate	1. Instrument is at least partially related to whether or not someone participates in a program of interest 2. Instrument is completely unrelated to everything else that might drive changes in the outcome of interest	Outcome variables Instrumental variable Control variables	Endline*	Participants Comparison group	A lottery for adolescent girls to receive free iron supplements can be used as an instrument for receiving iron supplements.
Randomized Evaluation	Individuals are randomly assigned to participants and control group. Experimental method for measuring a causal relationship between two variables	Individuals who are randomly assigned to the control groups	Randomization worked: That is, the two groups are statistically identical (on observed and unobserved factors).	Outcome variables Control variables	Baseline Endline	Participants Comparison group	Girls 14–18 in Jhajjar and Sirsa are randomly assigned the chance to participate in a program distributing free iron supplements. Average hemoglobin levels of girls receiving the program are compared with those who didn't receive the program after the program was run.

*Baseline data for the method is not strictly necessary to collect, but if baseline data is collected it can be used as "control" variables making the method more robust.

[33] Adapted from J-PAL website: https://www.povertyactionlab.org/sites/default/files/documents/Randomization%20Methods%20PDF.pdf

THE IMPACT EVALUATION METHOD TREE[34]



[34] Adapted from Hempel and Faila, "Measuring Success of Youth Livelihood Intervention: A Practical Guide to M&E."

III. BUDGETARY CONSIDERATIONS

The costs of an impact evaluation will vary widely depending on the evaluation method, location, capabilities of the evaluation team, length of time in the field, level of survey inputs and infrastructure, as well as a whole host of additional factors associated with the specifics of the evaluation itself. The context-specific nature of impact evaluations makes it impossible to give exact cost estimates. Instead, the goal of the following section is to provide a useful starting point for thinking about all the various inputs involved in an impact evaluation budget. Organizations will face the challenge of (a) adapting the general cost considerations presented here to their particular context, and (b) monetizing the inputs into a bottom line cost figure. It is recommended that an organization consult with an expert or an evaluating agency to ensure that cost estimates accurately depict the costing realities of an evaluation.

Calculating an inclusive and exhaustive budget is a difficult procedure given the many and diverse inputs needed to carry out an evaluation. When an organization is in the process of shortlisting its impact evaluation agenda, it may not be strictly necessary to calculate an intensive costing of the prospective evaluation. Instead, a more tentative "back of the envelope" calculation may be sufficient to get a rough sense of the comparative costs of different evaluations. Once an organization has decided upon an evaluation, a complete budget will need to be drawn up in collaboration with the evaluating agency. At this point any errors or omissions made may have serious implications on the quality of the survey work.

One important point to keep in mind when starting a cost estimate exercise is the importance of separating out the cost of an intervention from the cost of the evaluation. The cost of a program or intervention is not the same as the cost of an evaluation.

Two key outlays are essential when estimating a budget for an impact evaluation:

- **Data collection costs:** Includes survey operation, training, data quality check, and data entry costs
- **Personnel and overhead costs:** Includes personnel costs of the team of research associates, managers, and other personnel to oversee operations and associated overheads

Typically, data collection costs account for 60 – 65 percent of a total evaluation budget, while the rest consists of personnel costs and overheads. Data collection costs are driven largely by: (1) the size and dispersion of the sample, (2) the number of survey rounds, (3) the length of the questionnaire, (4) the types of survey tools used, (5) the specifics of the field work, and (6) the data management strategy. **Box 8** presents a list of questions to think about when conceptualizing the data collection cost outlay of an impact evaluation. Outlays associated with personnel and overhead costs include (1) personnel and (2) equipment. A list of questions relevant to personnel and overhead costs are included in **Box 9**.

In addition to data collection costs and personnel and overhead costs, another less sizable component to include in a budget is a contingency fund. Given the uncertainties faced in budget setting, having a pot of funds set aside to use when unexpected expenses arise is extremely important. The size of this fund should usually be around 5-10 percent of the total cost of the impact evaluation.

Funding an Impact Evaluation

Often, even if there is an evaluation budget built into individual projects, likely the earmarked amount may not be enough to conduct a sufficiently rigorous evaluation. Pooling allocated evaluation budgets across all activities into a common fund is generally a best practice. A budget constraint for an evaluation should be determined by the size of the fund and the number of total impact evaluations to be conducted over a given period.

Additional financing opportunities outside of direct program budgets can come from many other sources including project loans, research grants, or donor funds. Common sources of donor funding include governments, development banks, multilateral aid organizations, foundations, and impact evaluation organizations.

BOX 8: IDENTIFYING COST COMPONENTS OF DATA COLLECTION

The Sample:

- How large is the sample?
- How spatially scattered is the sample?
- Is the sample in rural and remote areas?
- How many strata and clusters are there?
- Will a census or initial household listing be needed to create the sample?

The Questionnaire:

- Is the questionnaire paper based or will data collection be done digitally?
- What is the length of the questionnaire?
- How long is the average interview?
- How many modules is the questionnaire?
- How clear and coherent is the formatting?
- How will respondents give consent to be interviewed?
- Does the questionnaire ask open or closed ended questions?
- Who is the main respondent of the questionnaire?

The Types of Survey Tools:

- Will there be any qualitative data collection?
- Will there be any cost compensation or token of appreciation for respondents of focus group discussions, informant interviews, etc.?
- How sophisticated are the measurement tools (e.g. anemia tests, student exams, implicit association tests, etc.)?

The Data Management Strategy:

- What is the system of evaluation monitoring employed?
- How many survey back-checks will be conducted?
- Will data entry be doubled?
- How many back checks will there be in data entry?
- How long will it take to clean the data?

The Fieldwork:

- How long is fieldwork projected to last?
- How many rounds of surveying will be conducted (e.g. baseline, midline, and endline)?
- How difficult will it be to follow up respondents at a later survey round?
- What is the estimated non-response rate from respondents?
- Will survey instruments be piloted?

BOX 9: IDENTIFYING COST COMPONENTS OF PERSONNEL AND OVERHEADS

Personnel Costs:

- How many enumerators will be needed?
- How long will it take to train enumerators?
- What will be the wage rate of enumerators, project managers, data managers, field managers, data entry operators, drivers, secretaries, translators, accountants, and any other personnel involved in the evaluation?
- How will field staff travel?
- What will be the travel allowance for personnel involved in the evaluation?
- Where will staff be accommodated while in the field?
- Will the principal investigators or senior researchers be travelling to the field?

Equipment Costs:

- Will scales, measuring tapes, measuring boards, and other survey equipment be needed?
- Will digital data collection tablets or computers need to be purchased?
- Will cars and fuel be purchased?
- Are maintenance costs projected for any equipment?
- How much printing and photocopying will be needed?
- Will any communication devices be needed for field staff?

IV. THE FINAL SELECTION

Before arriving at the final evaluation selection, an organization should have completed Impact Evaluability Activity Assessment forms (**Appendix A**) for each evaluation that has cleared the initial shortlist. These forms are the principal resource for informing the final selection. In a collaborative setting, organization staff should go through assessment forms, and make qualitative or quantitative rankings of the respective shortlisted evaluations. Sometimes, if there are few competitive evaluations, the evaluation budget is high, and the listing of evaluations to commission is evident, the selection process can be exceedingly easy, requiring little structure and less detailed rankings. Often though, the reverse is true – budgets are severely constrained and there are many potential evaluations. In this case, a structured, transparent, and comprehensive ranking process should be done.

The evaluation method and corresponding rigor of the evaluation will be a main focus. Discussions should revolve around the extent an evaluation can produce a credible estimate of program impact. How rigorous does the impact evaluation need to be in order to say something meaningful? Does the proposed method provide the necessary rigor? What are the key assumptions of the chosen impact evaluation method (see **Appendix C**)? Are the assumptions likely to be satisfied under the program's context? An organization does not want to be in the unfortunate situation of commissioning an evaluation and spending money on a study only to realize that the evaluation design is flawed, relies on too many tenuous assumptions, or requires a greater sample size than predicted. Having an evaluator or a more technical partner present can assist in this conversation.

Budget will, of course, be another important factor. Some subjective decision will need to be made on whether the use-value of an evaluation will justify its expense. Can the question of interest be answered rigorously enough to make the expense worthwhile? Is there a potential pool of donor funds that can used for a particular evaluation?

After the final selection of evaluations have been made, an organization may decide to bring on an evaluating agency, if one hasn't already been called.

V. MANAGING AND COMMISSIONING THE EVALUATION

Creating a Terms of Reference (ToR)

The final step in the process is creating the Terms of Reference (ToR) for the evaluator. There are various guidelines and protocols surrounding the application of the appropriate impact evaluation methodology, management of evaluations, data collection, and analysis that enhance the credibility of the impact evaluations. To help ensure the quality of the evaluation, a series of recommendations have been included on how to address some of the principal aspects of a ToR (see **Appendix D**). Appendix D also provides suggestions on how to strengthen evaluations by incorporating certain requirements for an evaluator in their ToR. The objective in providing this information is to make certain that organizations have the capacity to better manage their evaluations and evaluation stakeholders.

PAUL SMITH | A MEMBER OF THE PROSPERO MICRO-FINANCE TEAM EXPLAINS MICRO-CREDIT TO A FAMILY IN A BARANOA NEIGHBORHOOD, COLOMBIA.

APPENDICES

APPENDIX A: IMPACT EVALUABILITY ACTIVITY ASSESSMENT

ACTIVITIES RECOMMENDED FOR EVALUATION

Name of Activity:

Which broader objective/focus does this activity fall under?:

PRIORITIES:

1. Which criteria does the activity satisfy and how so?

2. How so?

3. How so?

4. How so?

5. How so?

USE VALUE:

What decisions will the evaluation inform/how will the information be used?

Who will be the primary user(s) of the evaluation?

PRELIMINARY RESEARCH QUESTIONS:

(Attach the program's theory of change)

Given its use value, what are some initial research questions?
Use the activity's Theory of Change. Are these causal questions?

METHODOLOGY:

What is the project and evaluation timeline?

Project Timeline: For example, has the project been implemented?
Evaluation Timeline: When would the evaluation start?

Which are the methods recommended for use? Why and what are the pros and cons?

Method:

Why?

Potential problems?

BUDGET

Estimated Budget:

Budget provided within the activity for IE:

INSTRUCTIONS FOR COMPLETING THE IMPACT EVALUABILITY ACTIVITY ASSESSMENT

The following paragraphs provide detailed instructions on how to complete the Impact Evaluability Activity Assessment. Care should be exercised to ensure that the Assessment is completed carefully. Investing time on this tool will help shortlist only the most beneficial evaluations. Monitoring and evaluation staff should complete Assessments, in collaboration with program staff and commissioned evaluators.

Components of the Impact Evaluability Activity Assessment

The Impact Evaluability Activity Assessment consists of five parts: priorities, use value, preliminary research questions, methodology, and budget. The first two, priorities and use value, involve justifying why a certain activity satisfies the larger research and evaluation priorities and what kinds of decisions these impact evaluations can inform. The last three parts elaborate on specific research questions and the ideal methodology to be used for the evaluation, taking into account practical considerations and the estimated budget for conducting the impact evaluation. A fulsome Assessment should also include the program's Theory of Change.

Priorities

To establish priority evaluations, list out the criteria that the prospective evaluation fulfills and provide justification on how the criterion is satisfied by the activity and to what extent. Recall that the seven criteria to consider when defining priorities are: (1) Program feasibility, (2) Innovations, (3) Scale and/or transfer, (4) Relevance to broader objective/focus, (5) Informing global debates, (6) Large projects, (7) Project timelines

Suppose the priority is to evaluate innovative projects that are easily scalable within an existing Government program and which have a potential for international transfer. The following questions must then be answered:

- How is the activity innovative? Using information provided in the activity description from existing documents, address how it is an innovation. For example, is it using a new product/service or delivery system?
- Why is the activity scalable? Using information from existing documents, identify characteristics of the activity that renders it easily scalable. For example, is it not very resource intensive? Is it a simple design? Is there an existing government program that the activity can piggyback on?
- Why is the activity transferable? Using information from existing documents, identify characteristics of the activity that render it easily transferable. For example, is the context within which the activity is implemented not specific to a particular location/culture/socioeconomic-political environment?

Use Value

Elaborate on the kinds of decisions that could be made based on the results of an impact evaluation of the activity, keeping in mind the uncertainty in the results of the impact evaluation (positive, negative, or null). For example:

- Will information from the impact evaluation inform the decision on whether the activity should be chosen for scaling or transfer?
- Will information from the impact evaluation inform the decision on whether the activity should receive sustained funding at different implementation stages?
- Will information from the impact evaluation inform some aspect of the larger objective/focus of the organization?

Following identification of the kinds of decisions that can be informed by the impact evaluation, the next step is to determine the primary consumers of the information generated by the evaluation.

Primary Research Questions

The activity in question may have a variety of different outcomes. The key to completing this section is defining the activity's theory of change (see **Appendix B**) and determining how to use the theory of change to identify the outcome that is most aligned to the set of priorities. From this outcome, a clearly and narrowly defined research question can be mapped.

For example, if the activity is an educational program that has an impact on student learning outcomes, builds teacher's teaching skills, and improves accountability within the school system, it is important to identify the most important outcome around which the impact evaluation will focus. This decision will more narrowly define the research question.

Methodology

Since the methodology chosen hinges heavily on the project timeline, details of the timeline and evaluation should be provided.

Subsequently, the method for impact evaluation should be selected and further justification should be provided by elaborating on the pros and cons of using this methodology. To aid in the completion of this section refer to the "**The Impact Evaluation Method Tree**" flowchart as well as the "**Table of Methodologies**". Additionally, **Appendix C** gives an in-depth listing of the various evaluation methods.

For example, suppose this is the description and status of the activity: *The activity provides remedial instruction to the bottom 20 percent of students in grades 3–5 in government primary schools. Implementation has already begun and there is no baseline data.*

- In the "why" section, explain the ways in which project timeline, availability of data, potential to collect data, and activity implementation characteristics affect the selection of methodology.

    i. Since the participants for the program are identified as satisfying a specific criterion (bottom 20 percent of the class), this can be used to create a credible comparison group comprising of individuals just above that cutoff.

    ii. Therefore, the evaluation methodology chosen is "Regression Discontinuity" and the sample used in the impact evaluation will be students who placed just above and below the 20th percentile.

- If there are any reasons that the assumptions for the chosen methodology may not be met, these must be elaborated in the "potential problems" section.

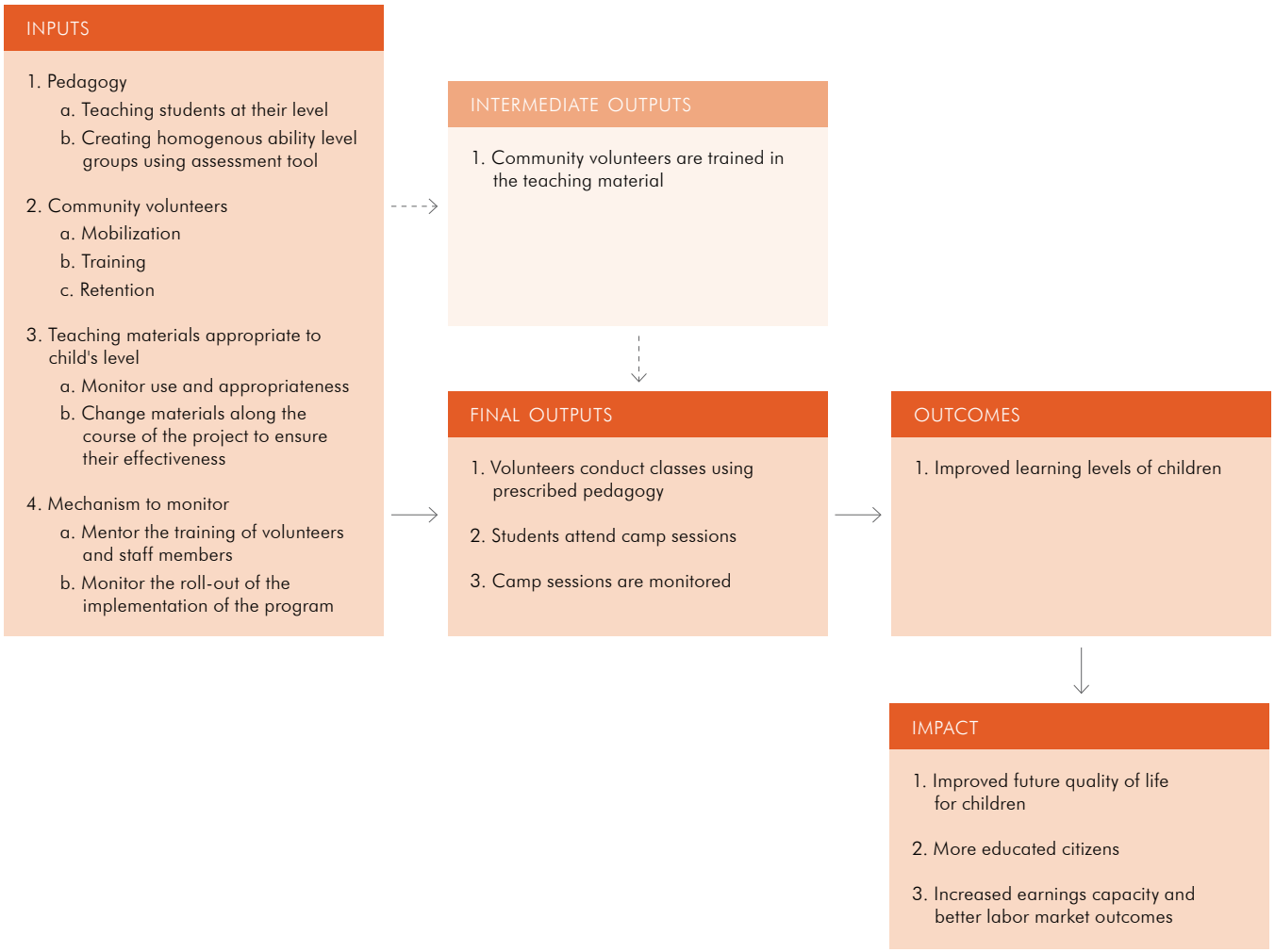
    i. Suppose there is reason to believe that the children are not chosen to receive the activity according to the decreed cutoff. In this case, using regression discontinuity will lead to misleading results and this problem must be noted.

Budget

Once the methodology has been identified, a budget estimate for the evaluation has to be calculated. To determine an accurate estimate of the cost, an organization must consider the many different "ingredients" that make up an evaluation. For an overview of these ingredients see the segment on **Budget Considerations** in the first section of the IE Tool.

APPENDIX B: EXAMPLE THEORY OF CHANGE

Learning Camps Program: This program provides 10-20 day intensive camps for children in grades 3-5 to accelerate their understanding of basic reading and arithmetic skills. Learning camps differ from normal classroom activities in the following manner: (a) camps offer a more fun and participatory environment; (b) teachers are trained community volunteers; (c) children are grouped according to their learning level as opposed to their grade level.



ASSUMPTIONS

Implementation:

- Availability of eligible and qualified volunteers
- Volunteers are incentivized to stay onto the program
- School infrastructure is sound and available
- Training is of good quality and equips volunteers with needed skills and support
- Monitoring system function properly

Theory:

- The assessment tools are reliable
- The pedagogy is the right way to address the needs of the children
- Grouping children by level does not demotivate the children

APPENDIX C: QUANTITATIVE EVALUATION METHODS

This section provides a description of the following eight evaluation methods: *pre-post, simple difference, difference-in-difference (or "double-differences"), multivariate regression, instrumental variables, propensity score matching, regression discontinuity,* and *randomized evaluations.* Each description includes a general overview of the methodology using the stylized iron-supplement example, and provides answers to the following questions:

- What are the main assumptions upon which this method's credibility in approximating the counterfactual rests?
- What are the practical constraints (data, timing, sample size, etc.) to be aware of?
- What is an example of a study that uses this method of impact evaluation?

The first question identifies the assumptions of each method that must be satisfied in order to have an accurate estimate of the program's impact. In other words, these assumptions provide a set of conditions that if fulfilled allows the comparison group to be an accurate estimate of the counterfactual. Causality can only be credibly and accurately claimed when these assumptions can be argued to have been sufficiently satisfied. The second question sets out the particular circumstances that must be in place for a method to be used. For example, some methods require identifying a comparison group and collecting data prior to program implementation. If this pre-program data are not available, then these methods will not be suitable. The final question (presented in the orange box) summarizes an actual evaluation and explains how the study fulfilled (or failed to fulfil) the necessary assumptions to claim causality.

After introducing the eight evaluation methods, a final section outlines additional considerations relevant to many of the methods presented. This includes information on sample size requirements, attrition, spillovers, and contamination. If not satisfactorily addressed, these factors can undermine the credibility of the method.

D.1. EVALUATION METHODS

PRE-POST

Sometimes, due to practical constraints, one may try to estimate impact using methods that would be classified as an observational study. Pre-post is one such method. While a pre-post study can establish a relationship between a program and an outcome of interest, the relationship is not causal. This is because the method of constructing a comparison leads to inaccurate estimates of the counterfactual.

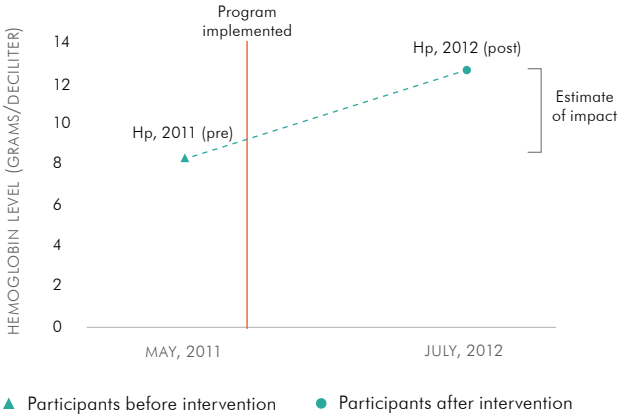
In a pre-post estimation, "impact" is measured by comparing data from the same group of program recipients both before the program is administered and after (i.e. the change in outcome indicators before and after the program). A pre-post is often used in cases where data are not available on a separate comparison group, but baseline data before the intervention are available.[35]

In the case of an iron supplement program, a pre-post would take the difference between an average outcome measure (hemoglobin concentrations) for girls who received the program in 2012 and the average outcome measure for the same group of girls in 2011 when they were not receiving the program.

$$\text{"Impact"} = (H_{p,2012}) - (H_{p,2011})$$

Where $H_{p,2012}$ is the average hemoglobin level for adolescent girls after receiving the iron supplement program in 2012, and $H_{p,2011}$ is the average hemoglobin level for the same group of girls in 2011 when they had not yet received the program.

FIGURE 7. IMPACT OF THE IRON SUPPLEMENTS ON HEMOGLOBIN LEVELS USING PRE-POST



What are the main assumptions?

In order for a pre-post method to yield an accurate estimate of a program's impact, one would need to prove that the program was the only factor that caused changes in the measured outcome of interest over the duration of the program. This is often a very difficult case to make because there are a number of other environmental factors that are likely to at least partially explain the changes we see in the measured outcome. For instance, in our iron example, some of the girls in the sample may have reached puberty between 2011 and 2012, which could affect hemoglobin levels in addition to the iron supplement program. Hence, it is very difficult to credibly argue that any observable changes can be attributed to the program of interest.

What are the potential constraints?

Pre-post studies require collecting data on program participants before and after the program is implemented. If an evaluation is planned ex post, and no data collection was conducted before program implementation, this method cannot be used.

AN EVALUATION OF THE BALTIMORE COMMUNITY LEAD EDUCATION AND REDUCTION CORPS (CLEARCORPS PROGRAM)[36]

Intervention: To combat the toxic effect of lead paint in low-income residential homes in the United States, a joint public-private initiative known as the CLEARCorps Lead Risk Reduction program began in seven cities across the United States in the late 1990s. The CLEARCorps program involved: (a) cleaning and repairing homes to make them lead safe, (b) educating residents on lead-poisoning prevention techniques, and (c) encouraging maintenance of low lead through specialized cleaning efforts.

Study Design: Using a pre/post method, a 1998 study estimated the extent to which the CLEARCorps program was successful in reducing lead exposure in treated urban housing units. To aid the investigation, CLEARCorps members collected lead dust samples from floors, window sills, and window wells, before, after, and 6 months following the intervention. Average lead dust levels declined after the education and cleaning campaign by 36, 77, and 83 per cent for floor, window sill, and window well measurements respectively. These mean differences were found to be statistically significant using paired t-tests. Six month measurements showed even more dramatic decreases in lead levels relative to baseline levels.

Threats to Validity: A major assumption of pre/post studies is that the program itself was the only factor affecting the outcome. In this case, this means that the CLEARCorps program alone caused lower lead dust levels.

Testing Assumptions: The study is unable to validate the strict assumption. As the author suggests it is unrealistic to assume that the program alone was responsible for low lead levels as other lead hazard reduction programs were occurring simultaneously in program areas. In addition, seasonal variability in lead dust levels could have caused post program measurements to appear lower than they otherwise would have been. These confounding factors would likely cause an overestimation of the program's effect.

SIMPLE DIFFERENCE

Simple difference studies, like pre-post, are observational studies that do not provide a causal estimate of the impact of a program. Simple difference evaluations measure "impact" by differencing outcomes between the group receiving the program and a group that did not receive it. When baseline data are not available, but data on a comparison group are available after a program has been implemented, this method can be used.

A simple difference may take the difference between average hemoglobin concentrations for girls who received the iron supplement program in 2012, and average hemoglobin concentrations in 2012 for girls not exposed to the treatment and who did not receive any iron-supplements.

$$\text{"Impact"} = (H_{p,2012}) - (H_{np,2012})$$

where $(H_{p,2012})$, as above, is the average hemoglobin level of adolescent girls who received the program in 2012, and $(H_{np,2012})$ is the average hemoglobin level of a group of non-recipient adolescent girls in 2012.

FIGURE 8. IMPACT OF THE IRON SUPPLEMENTS ON HEMOGLOBIN LEVELS USING SIMPLE DIFFERENCES

What are the main assumptions?

Simple difference assumes that the reasons why one group received the program and the other did not are completely unrelated to the outcomes we are observing. For the supplement example, it would be problematic if the girls who received the program were from public schools while the comparison group attended private schools. This is because girls attending private schools are likely to be different from girls attending government schools for a number of reasons that are also likely to affect health outcomes. For example, girls in private schools may come from families with higher average incomes, and may be more likely to eat nutritious foods, and therefore, in the absence of the program, have on average higher hemoglobin rates. This selection bias is a major threat to the validity of any simple difference estimate and is the reason why the method cannot be used to generate a credible impact estimate.

What are the practical constraints?

Data must be collected on an identified comparison group after the program ends.

OPINION: THE "READ INDIA" PROJECT NOT UP TO THE MARK[37]

Intervention: A 2004 survey conducted by Pratham found that 39% of children aged 7-14 in rural Uttar Pradesh could read and understand a simple story, while a further 15% could not recognize a single letter. To improve reading outcomes among primary students, Pratham designed the "Learn-to-Read" program, a sub-component of its Read India campaign. The Learn to Read program leveraged community involvement by sharing information on the status of literacy and the rights of children to education in village meetings. Pratham then trained community volunteers to teach children specially designed reading materials in after school classes.

Study Design: To answer whether Learn-to-Read "worked", an evaluation was commissioned by a private organization to test the impact of the program on students' reading levels. A team of evaluators compared program recipients with a group of children from the same village who did not attend the after school classes. After one year of reading classes, it was found that Pratham students could only recognize words whereas those who did not participate in the program were able to read full paragraphs. The study concluded that Learn-to-Read did not positively impact students' learning levels.

Threats to Validity: The results of the evaluation are contingent on the strict assumptions that the comparison group of non-participants were identical to program participants. This means that the non-participants were as likely as participants to enter the program before it had started. If there were observable or un-observable differences between the students in the participant and comparison group, then a simple-difference study would be biased.

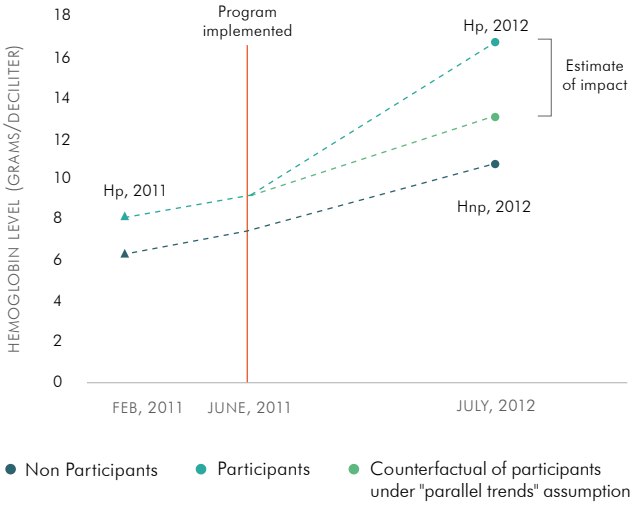
Testing Assumptions: The evaluation of the Learn-to-Read program selected comparison students that lived in the same village; therefore, it is unlikely that there are geographic related differences between the participant and non-participant groups. However, there may be many other factors that differ between the two groups, such as initial education levels, socioeconomic effects, or unobserved levels of ability. Since the study neither controlled for these other factors, nor provided an explanation of why these factors would not influence reading outcomes, the estimate of program impact is likely to be biased.

DIFFERENCE-IN-DIFFERENCE ESTIMATION

A difference-in-difference estimation compares the changes in outcomes over time between a group receiving the program and a group that did not receive the program but was exposed to the same environmental conditions.[38] This method essentially controls for factors that are constant over time (time invariant), and can account for some of those unobservable time-invariant factors, such as motivation or ability.

For the iron supplement program, researchers may identify districts neighboring the program-run districts that have similar environmental factors. A comparison group could contain girls aged 14-18 in these neighboring districts. Estimating the effect of the program on hemoglobin levels would require differencing the before-and-after hemoglobin levels for the group receiving the program and then differencing from that value the before-and-after hemoglobin outcomes for the comparison group. This difference-in-difference measure controls for factors that are constant over time (the first difference) and also captures time-variant factors (the second difference).

FIGURE 9. IMPACT OF THE IRON SUPPLEMENTS ON HEMOGLOBIN LEVELS USING DIFFERENCE-IN-DIFFERENCE



What are the main assumptions?

The primary assumption of this method is known as the *parallel trends assumption*: that the trend over time for the comparison group is the same as it would have been for the group receiving the program had they never received the program (in other words, in the absence of the program both groups would have changed in a similar manner).

What are the practical constraints?

Data for both the group receiving the program and a comparison group must be available from both before and after the program. In addition, if there is reason to believe that the "trends over time" for the comparison group and the group receiving the program are not the same (for example, if one of the groups experienced environmental changes that the other group did not experience), then this is not a credible method to use. This may be an issue if say the neighboring district is better run, or if there is an NGO present in that district that is addressing the anemia issue. This method is most credible when the only differences between the group receiving the program and the comparison group are time invariant, and therefore accounted for with baseline measures.

Ideally, the selecting criterion for who is in the comparison group versus the group receiving the program is close to random. This would be the case when a difference-in-difference estimation is combined with a natural experiment (see Box B.1.).

THE IMPACT OF THE JUDICIARY ON ENTREPRENEURSHIP: EVALUATION OF PAKISTAN'S "ACCESS TO JUSTICE PROGRAMME"[39]

Intervention: Prior to the implementation of the Access to Justice Programme in 2002, there were 1.2 million outstanding court cases in subordinate courts in Pakistan. The two year waiting time to treat court cases created a major obstacle to the formation of business, stifling entrepreneurship in the country. The Access to Justice Programme (AJP) was adopted as a way to enhance the performance and efficiency of the judiciary system by providing training for civil and criminal judges on case-flow management. The program was piloted in six districts in Labore, Peshwar, and Karachi.

Study Design: As the treated districts were selected non-randomly, the impact evaluation study uses a difference-in-difference design to estimate the effect of the reform on the performance of judges and on the level of entrepreneurship. The difference-in-difference estimation compared districts under the AJP relative to non-AJP districts prior to program implementation in 2001, and following the program in 2003. As a result of the judicial reform, judges disposed of 25% more cases and the time to process a case was reduced by one year in areas where the AJP was rolled out. This had reverberating implications on entrepreneurship – treated districts had a 10 percentage point reduction in the likelihood that a law and order situation prevented individuals from working.

Threats to validity: A common threat with difference-in-difference studies relates to the progression of treated and untreated districts over time. Districts receiving the AJP program may have a different time trend between 2001 and 2003 than non-recipient districts. This would result if the rate of court case completion in non-treated areas from 2001 to 2003 was different than the rate of court case completion for judges in treated areas had they not received the program.

Testing Assumptions: The study presents three tests to satisfy the assumption that the counterfactual group and the actual comparison group have similar trajectories across outcomes from 2001 to 2003:

a. The first test compares the trends of treated and non-treated judge's performance before the intervention began in 2002. Since there was no difference in the trend of judge's performance prior to 2003, this lends confidence that the trends during the intervention period will be the same.

b. Second, since the treated districts were selected due to being exceptionally slow in handling judicial cases, the time trend for these districts may be abnormally fast in the absence of the program since they may revert back to the mean processing time. To correct for a possibility of mean reversion, the study incorporates a specific control variable in their model.

c. Lastly, as the parallel trends assumption is equivalent to assuming that there are no omitted time-varying and district-specific effects correlated with the reform, the author accounts for factors such as police strength that may confound the reforms' effects.

BOX D1: USING NATURAL EXPERIMENTS TO APPROXIMATE RANDOM ASSIGNMENT

Sometimes researchers use natural experiments to help approximate an experimental condition. Natural experiments occur when some external condition, such as a natural disaster, an idiosyncrasy of geography, or an unexpected change in policy assigns participants randomly to a comparison group and a group receiving a particular program or policy. These natural experiments can be used in conjunction with other methods (usually with a quasi-experimental method) to produce a credible impact estimate.[40]

MULTIVARIATE REGRESSION:

To produce a more accurate impact estimate, evaluators can use a simple difference model that looks at differences in outcomes between participants and non-participants, while controlling for other factors that might be related to the outcome of interest. This method is referred to as multivariate regression. To get an estimate of the impact of the iron-supplement program, the hemoglobin level (H) could be regressed on a variable indicating whether an adolescent girl received the treatment (D), whether the girl attended a public or private school (P), income of the girl's household (I), whether the girl has reached puberty (M), and an error term capturing variables not included in the model (E):

$$H = \beta_0 + \beta_1 D + \beta_2 P + \beta_3 I + \beta_4 M + E$$

[35] Paul Gertler et al., *Impact Evaluation in Practice*, 7.

[36] Duckart, Jonathan, "*An Evaluation of the Baltimore Community Lead Education and Reduction CORPS (CLEARCorps Program)*"

[37] This is a fictitious impact evaluation used as a J-PAL case-study for purely pedagogical purposes

[38] Paul Gertler et al., *Impact Evaluation in Practice,* 96.

[39] Chemin, M. "The impact of the judiciary on entrepreneurship: Evaluation of Pakistan's Access to Justice Program", Sep. 2007.

[40] Murnane and Willett, *Methods Matter: Improving Causal Inference in Educational and Social Science Research*, 135 -136.

The estimate of the impact of the supplement program is $ß_1$, the coefficient on the treatment variable, D.

What are the main assumptions?

Multivariate regression assumes that all relevant factors have been included in the model. Any factors that have been excluded are either unrelated to the outcome or do not differ between participants and non-participants.[41]

What are the practical constraints?

Data must be collected on participants and non-participants, including not only the outcome of interest, but also all relevant factors which the researcher hopes to control for. Also, controlling for additional factors does not necessarily imply that the regression estimate is the causal impact of a program because: (1) some relevant observable factors may be left out, and (2) this method cannot account for unobservable differences that affect outcomes. For example, parental preference for investing in girls is an unobservable factor that would likely be related to health outcomes. Parents that send their daughters to private schools may value investing in their daughters more and may also be more likely to invest in health care for their daughters. This unobservable preference would likely affect prevalence of anemia among this group, but would not be accounted for using a regression.

WOMEN'S CREDIT PROGRAMS AND FAMILY PLANNING IN RURAL BANGLADESH[42]

Intervention: The provision of collateral-free credit to low-income women is believed to impact family planning decision-making as a direct result of micro-credit training activities and also indirectly through the empowerment of women. To formally test this relationship, a study examined the effect of credit programs on family planning attitudes and contraception practices of participating female borrowers. This was accomplished through an impact evaluation of five micro-credit NGOs operating in rural Bangladesh.

Study Design: Multivariate regression analysis was performed on a random sample consisting of loan recipients and non-recipients from the program areas of the five micro-credit NGOs, as well as a sample of non-recipients from non-program areas. To isolate the effect of the credit program on family planning outcomes, socioeconomic and demographic variables were included in the regression as control variables. Regression results show that micro-credit participants were significantly more likely than non-recipient women to be current contraceptive users and to report they do not desire additional children.

Threats to Validity: A major threat to the study is the potential for selection bias. Women receiving the program may be systematically different than non-recipients, and this difference may not be controlled for in the regression equation due to unobservable factors or omitted control variables. For example, credit programs may tend to recruit women who are predisposed to using contraception.

Testing Assumptions: The study tried to address this concern by ensuring broad representation of women, controlling for socioeconomic and demographic characteristics, and making sure that non-recipients were in the same vicinity as the randomly selected program areas to control for geographic specific differences. These safeguards are not sufficient to show a causal relationship between credit programs and family planning; however, results do offer evidence of correlation.

MATCHING METHODS AND PROPENSITY SCORE MATCHING

Matching methods are used to construct comparison groups that are similar on a number of specified observable characteristics. There are a number of ways matching can be done. Simple Matching matches a program recipient with a non-recipient using all relevant observable characteristics. The downside of simple matching is that there may be a large number of characteristics or dimensions that may need to be matched in order to ensure similarity between groups.[43] The more dimensions there are, the less likely you are to find another unit or individual that can be matched across all relevant characteristics. One way of addressing this so called "curse of dimensionality" is by using propensity score matching. This method estimates, for each treated and non-treated observation, a propensity score, or the estimated probability of program participation given a set of observable characteristics. Program participants are then matched with individuals in the comparison group on the basis of their propensity score.

Similar to difference-in-difference, a comparison group for the iron supplementation program could be identified by finding girls in a neighboring area, but rather than choosing all girls 14-18 from government schools, individual girls are matched to participants on a number of observable characteristics such as age, socioeconomic status, years of education, scores in school, health status, etc. Hemoglobin rate differentials between matched program participants and comparison individuals are then averaged together to provide an estimate of the effect of the program.

What are the main assumptions?

This method of constructing a comparison group assumes two things: (1) that the characteristics used to predict program participation include all relevant observable characteristics, and (2) unobservable characteristics, such as motivation or family values, do not drive decisions to participate in the program. Combining methods can be a possible way to relax the assumption regarding unobservable characteristics (see Box B.2.).

BOX D2: COMBINING EVALUATION METHODS

In many cases, different methods can be used in conjunction to increase the credibility of an impact evaluation. For example, a difference-in-difference method accounts for unobservable differences that do not change over time, but is not an accurate estimator if there is reason to believe that program participants and the comparison group do not have parallel trends in the outcome of interest over time. Alternatively, the biggest threat to validity for matching methods is the presence of unobservable differences. If an evaluator can combine difference-in-difference estimation and matching methods, this would help address their respective potential threats, and would make the evaluation more rigorous and credible.

What are the practical constraints?

Data must be available for both participants and non-participants on all the pre-program factors influencing program participation, as well as on the outcomes of interest. While technically this data can be captured using a cross-sectional survey after the program has been implemented, it is much more reliable to have baseline and endline data so that the number of recall questions can be minimized.

FINANCIAL INCENTIVES FOR MATERNAL HEALTH: IMPACT OF A NATIONAL PROGRAMME IN NEPAL[44]

Intervention: In July 2005, Nepal introduced the Safe Delivery Incentive Program (SDIP), a social program designed to incentive maternal health by providing a cash benefit to mothers who gave birth in a public health facility.

Study Design: To evaluate the impact of the program on the utilization of pregnancy services at public health facilities, the study used propensity score matching to select a counterfactual group of non-participants. The chosen non-participants were matched to participants with similar characteristics based on their predicted probability of participating in the SDIP. Since the program is universal and nation-wide, non-participants included those women who were unaware of the SDIP prior to childbirth. The SDIP was found to increase institutional deliveries by four percentage points.

Threats to Validity: Two assumptions are required to ensure that non-participants unaware of the program provide an accurate estimate of the counterfactual. First, propensity scores must be based on all relevant observable characteristics. This means that after controlling for all the relevant observable characteristics, participation in the program is essentially random. An extension of this assumption is that program participation is not driven by any unobservable characteristics. The second assumption is that the treated individuals have largely the same characteristics as the non-treated. This must be the case, if treated and non-treated individuals are to be matched.

Testing Assumptions: To ensure that the non-participants unaware of the program provide an accurate estimate of the counterfactual, the authors run four tests.

a. First, they analyze the impact of the SDIP on both state and non-state health providers and check to ensure the presence of a substitution effect. Impact estimates that are biased are likely to hide evidence of a substitution effect between public and private utilization.

b. The second test uses an instrumental variable correlated with whether a mother heard about the SDIP and uncorrelated with public health facility delivery, to check for bias from unobservable characteristics.

c. The third strategy splits the treatment group into mothers who, (1) knew about the SDIP and expected to receive the cash incentive after delivery, and (2) knew about the SDIP but did not expect to receive the cash incentive. The authors then applied propensity score matching to each of the treatment groups separately, expecting no treatment effect when comparing the utilization of maternity services between the non-expectant group and the comparison group.

d. Lastly, using a histogram the authors show that pre-matched propensity scores between treatment and comparison groups are overlapping. Along with balancing tests, this ensures that treated and non-treated individuals share similar characteristics and can be accurately matched.

REGRESSION DISCONTINUITY ESTIMATION

Regression discontinuity (RD) can be applied when the eligibility to participate in a program or policy is set on the basis of a continuous measurable variable with a cutoff that determines who receives the program and who does not. In the case of the iron supplement program, if all girls below the age of 14 years were eligible for the program, then age would be a continuous measurable characteristic that could be used to assign eligibility for the program.

In an RD design, one would compare those eligible units just above the cutoff and just below, the assumption being that the cutoff is in some sense arbitrary (i.e. random) with respect to

[41] Abdul Latif Jameel Poverty Action Lab, "Methodology Table."

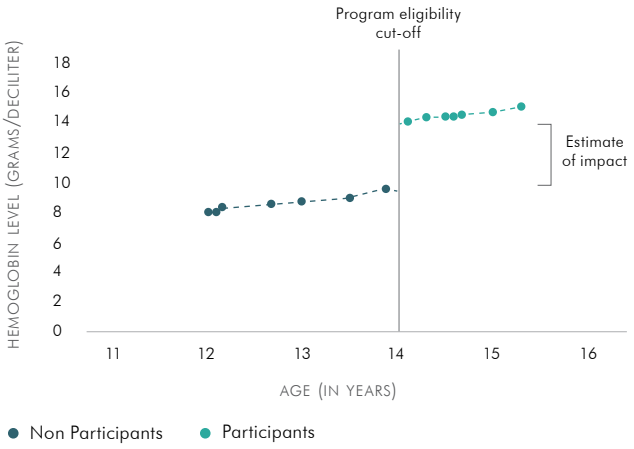
[42] Amin, Li and Ahmed, "Women's credit programs and family planning in rural Bangladesh", Dec. 1996.

[43] Handker, Koolwal, and Samad, *Handbook on Impact Evaluation*, 54.

[44] Powell-Jackson and Hanson, "Financial incentives for maternal health: Impact of a national programme in Nepal".

the outcome of interest. This implies that those just above and just below the cutoff should be very similar on average across all characteristics except whether they received the program. In the iron supplement program, an RD design that exploited the eligibility cutoff of 14 years would be appropriate only if the potential outcomes of hemoglobin levels for girls just below and just over the age of 14 would not be significantly different prior to the program. If this assumption holds, girls who turned 14 before the program could be compared with those who won't turn 14 until immediately after the program. The RD assumption would be violated, for instance, if it were the case that girls who turned 14 after the program respond much more strongly to iron supplementation than those who turned 14 before the program. While this may seem unlikely in this context, it may not be so for other studies.

FIGURE 10. IMPACT OF THE IRON SUPPLEMENTS ON HEMOGLOBIN LEVELS USING REGRESSION DISCONTINUITY



What are the main assumptions?

Another way to restate the key RD assumption is that the choice of cutoff score is not in any way related to the outcome of interest. An RD evaluation must make the case that crossing the threshold won't render those participants just above the cutoff different on a variety of factors from those who are just below the cutoff.

What are the practical constraints?

Regression discontinuity is only appropriate when two main conditions are fulfilled:

- There exists a continuous measurable characteristic which identifies eligibility. This characteristic is ordered and quantitative and we can rank the population of interest along it.
- There exists a clearly defined cutoff that determines eligibility for the program. Individuals on one side of the cutoff are ineligible for the program, and individuals on the other side are eligible.[45]

In addition, RD is a less credible design:

- When it is necessary to estimate the treatment effect beyond those within the narrow range around the eligibility cutoff. By definition, a RD design only estimates the impact for those within a narrow bandwidth around the cutoff point. For the iron supplement program, an RD design would estimate the average change in hemoglobin for adolescents who turned 14 just before and just after the program. The estimate of the program may not be generalizable for adolescent girls receiving the program between ages 14-18.
- When it is not possible to define a narrow enough bandwidth with a sufficient sample size.

IMPACT EVALUATION OF BURKINA FASO'S BRIGHT PROGRAM[46]

Intervention: In 2006, Burkina Faso implemented the Burkinabe Response to Improve Girls' Chances to Succeed (BRIGHT), a two-year program to improve school learning outcomes and mitigate the disparity in education access between boys and girls. BRIGHT constructed 132 primary schools in areas of low female enrolment and also provided a series of supplemental interventions such as: mentoring services, free daily meals, take home rations for girls who attended at least 90 per cent of their classes, school kits and textbooks, a literacy program, and a mobilization campaign to highlight and discuss barriers to girl's education.

Study Design: The evaluation probed the impact of BRIGHT on enrollment and test scores, while also investigating whether these impacts differed between girls and boys. To isolate the causal impact of the program, the study used a regression discontinuity design that exploited the cut-off score used to identify the 132 recipient villages out of the total 293 applicant villages. Short term impacts of BRIGHT were assessed after two years of the program using a household survey from a random sample of 30 households with school-age children in each of the 293 villages that applied. Results from the evaluation showed that BRIGHT had an extremely positive impact on school enrolment rates and also improved both math and French test scores by approximately 0.4 SD for both girls and boys.

Threats to Validity: There are a number of caveats to consider. First, Regression Discontinuity designs must prove that there are no unobservable differences that influence the outcome of interest between those just above and below the cut-off. In other words, this means that around the cut-off value there must be no discontinuous jumps in determinants of the outcome that cannot be controlled. Second, treatment effects are localized to villages around the cut-off and may not be generalizable to those

non-marginal villages far from the cut-off value. The third potential threat is if the implementation of the program does not follow the cut-off eligibility rule specified.

Testing Assumptions: Four separate analyses were undertaken to verify the appropriateness of using regression discontinuity.

a. Estimations were undertaken between child and household level variables and the eligibility score of the village to confirm that observable confounding factors do not significantly vary at the cut-off point. The estimations, although significantly significant, are very small, which suggests that the probability of treatment assignment was not due to other characteristics.

b. To check the assumption that the probability a village receives the treatment is distinctly higher at the cut-off point, the authors estimated the relationship between a dummy variable indicating actual receipt of a BRIGHT school and the villages' relative eligibility score. Villages above the cut-off score (the treatment villages) were 87 percentage points more likely receive a BRIGHT school than the villages below the cut-off score (comparison group). This confirms the use of the eligibility rule in deciding which villages would receive the BRIGHT schools.

c. Based on estimates of the probability that villages had a school in 2003, participant villages were not significantly more likely than comparison villages to have a school prior to 2005. This indicates that the treatment and control were comparable prior to the program.

d. To test whether the treatment effect can be generalized to villages with eligibility scores that are far from the cut-off, the characteristics between "marginal" and "non-marginal" villages were compared. Differences in characteristics between the villages are deemed to be sufficiently small, suggesting that the treatment effect is in fact generalizable.

INSTRUMENTAL VARIABLES ESTIMATION

An instrumental variable (IV) is a factor that is related to program participation, but is unrelated to anything else that might affect the outcome of interest. An IV essentially isolates the part of program participation that is due to random variation and looks only at the relationship between this randomly assigned part of program participation and the outcome of interest. For example, suppose a lottery was used to determine which schools in a district gave adolescent girls access to free iron supplements. The lottery can be used as an instrument for receiving iron supplements. The instrument then allows us to identify the relationship between the portion of treatment assignment that is due to random chance. This allows examination of the effect of being randomly assigned to take iron supplements on hemoglobin levels.

What are the main assumptions?

An instrumental variable is only valid if two key assumptions are fulfilled:

- The instrument must be strongly related to whether or not someone participates in a program.
- The instrument must be completely unrelated to everything else that might drive changes in the outcome of interest.[47]

What are the practical constraints?

Data must be available on the outcomes of interest, the instrument, and any other control variables that are relevant to the analysis. The biggest practical constraint is finding a valid instrument; most factors that are related to program participation will also be related to other factors.

IMPACT EVALUATION OF THE IRRIGATION MANAGEMENT REFORM IN NORTHERN CHINA[48]

Intervention: Traditionally in Northern China, water provision, maintenance, and tariff collection was collectively managed by village committees. However, changes brought on by economic growth prompted irrigation management reform in rural areas. Starting in 2002 the government began promoting the transfer of water management to Water User Associations (WUAs). Contracting also emerged as another alternative, whereby a WUA or a village committee would relinquish control of a portion of the irrigation system to a manager.

Study Design: The purpose of the study was to evaluate the impact of irrigation transfer from village committees to WUAs or contractors. Data was used from the China Water Institutions and Management Survey, which tracked villages in Northern China as they instituted management reforms over the years 2001, 2004, and 2007. Although collecting data on the same villages over time allowed the authors to control for time-constant differences within villages and their water sources, unobservable factors may still be omitted from the regression. Selection bias will occur if these unobservable factors are both correlated with the outcome of interest and a village's choice of switching to WUA or contracting. To mitigate the effects of selection bias, the study used "the average number of meetings per year upper level governments held to promote WUAs or contracting" as an instrumental variable. Results show that WUAs have been successful in improving irrigation system outcomes including: higher maintenance expenditure, more timely water delivery, higher percentage of irrigated land, and higher rates of water fees collected.

---

[45] Paul Gertler et al, *Impact Evaluation in Practice*, 82.

[46] Dan Levy et al, *Impact Evaluation of Burkina Faso's BRIGHT Program.*

[47] Khandker, Koolwal, and Samad, *Handbook on Impact Evaluation.*

[48] Huang, "Impact evaluation of the irrigation management reform in northern China", May 2014

Threats to Validity: Using an instrument to control for selection bias requires that the instrument does not correlate with the outcome of interest (maintenance expenditure, water delivery time, percentage of irrigated land, water fees collected) but does correlate with the treatment indicator. In this study, this means that the average number of meetings per year upper level governments held to promote WUAs (or contracting) strongly influences a village's choice to form a WUA (or contract out irrigation management). It also requires that the number of meetings to promote WUAs does not influence any other factor that could impact the outcomes of interest: maintenance expenditure, water delivery time, percentage of irrigated land, and water fees collected.

Testing Assumptions: To ensure that the WUA government meetings play a role in encouraging irrigation management reform, the study displays results of the model showing the estimation of the probability that villages will undertake irrigation reform. Results show that the government's efforts to promote WUAs (contracting) have a positive and significant effect on the probability of a village switching to WUAs (contracting). Checking the second assumption (instrument uncorrelated with any other determinants of the outcome of interest) is much harder to directly test. The only way to satisfy this assumption is providing a theoretical explanation why the instrument is unlikely to affect the outcome of interest. It is up to the reader to decide whether the explanation provided by the study is sufficient.

RANDOMIZED EVALUATION

The most credible method of avoiding selection bias from unobservable variables is to use randomization to create a counterfactual. Randomized evaluations use a lottery to decide who among the eligible population receives the program and who does not. Every unit that is eligible for accessing the program has an equal chance of being selected.[49] If the total number of eligible participants is sufficiently large, then the randomization process produces two groups that have a high probability of being statistically identical on all factors **EXCEPT** for its exposure to the program.[50]

Certain timings and circumstances lend themselves to randomization more than others. First, a randomized evaluation is best suited for the pilot phase of a program after the design kinks have been worked out but before resources are allocated to a full scale-up or rollout of the program across the entire population (which would eliminate the possibility of a comparison group).[51] The following circumstances lend well to using a randomized evaluation:

- When the program is oversubscribed; i.e. when the eligible population is greater than the number of program spaces available. In this case, scarce resources prevent the program from being scaled up to the entire target population, and a random lottery becomes a fair and transparent way of allocating scarce resources.
- When a program needs to be phased in gradually to the entire population. In this case, those who haven't received the program yet can serve as a comparison group.[52]

What are the main assumptions?

Provided the randomization created statistically similar groups, and there were no threats to the randomization design, there are no other qualifying assumptions that must be made to ensure the accuracy of the impact estimate.

What are the practical constraints?

One practical constraint of conducting a randomized evaluation is the willingness and capacity of the implementing partner to randomize who receives the program and who does not. In such cases, two designs can help alleviate partner concerns: *phase-in* and *encouragement* designs. In a phase-in design, randomization determines the timing at which an eligible unit starts receiving the program, with all eligible units eventually receiving treatment.[53] Encouragement designs provide a randomly allocated subset of eligible participants with information or incentives to encourage participation in a universal program. In such cases, the randomization exploits the fact that take-up rates among different eligible units might be different, and the tested intervention is the added encouragement, rather than the program itself. Such designs are useful when the program itself is undersubscribed.

Furthermore, randomization may not be appropriate:

- When evaluating macro policies (e.g. the effect of changing a country's exchange rate)
- When it is unethical or politically infeasible to deny a program to a comparison group
- If the program is changing during the course of the experiment
- If the program under experimental conditions differs from how it would be under normal circumstances

---

[49] Paul Gertler et al., *Impact Evaluation in Practice,* 51.

[50] Paul Gertler et al., *Impact Evaluation in Practice,* 51.

[51] Abdul Latif Jameel Poverty Action Lab, "When to Conduct an Evaluation."

[52] Paul Gertler et al., *Impact Evaluation in Practice,* 56.

[53] https://www.povertyactionlab.org/sites/default/files/documents/Randomization%20Methods%20PDF.pdf

[54] https://www.povertyactionlab.org/sites/default/files/documents/Randomization%20Methods%20PDF.pdf

[55] Rema Hanna et al, "Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves"

[56] http://www.povertyactionlab.org/methodology/how/how-design-evaluation#sample

UP IN SMOKE: THE INFLUENCE OF HOUSEHOLD BEHAVIOUR ON THE LONG-RUN IMPACT OF IMPROVED COOKING STOVES[55]

Intervention: Approximately seventy percent of the Indian population rely on solid fuels such as firewood, crop residue, or cow dung, to power traditional stoves. The smoke released from solid fuels contributes to climate change and is also linked to respiratory disease and lung cancer. To reduce the number of households relying on coal and biomass fuels, Gram Vikas, an NGO focusing on rural community development, distributed improved cooking stoves to 15,000 households in the Indian state of Orissa. Gram Vikas provided the materials and paid for the construction of the stoves, while households provided the mud for the base, labor, and a Rs. 30 fee to the mason. The NGO also gave training sessions to promote the proper use and maintenance of stoves.

Study Design: A phased in randomization approach consisting of three waves of implementation was used for the 2,651 households residing in 44 participating villages in Orissa. Households were randomly assigned by a public lottery to choose the first third of households that would receive the first wave of Gram Vikas stoves. Approximately three years later, a second round lottery occurred to choose the second wave of household recipients. During the second round of stove construction, Gram Vikas repaired and rebuilt damaged stoves from the first wave of construction. The improved cooking stoves were found to have little overall improvement in smoke exposure, no effect on health status, and they did not increase standards of living. The negligible effect of the stoves was attributed to low usage of the new stoves and a failure to keep them in working order.

Threats to Validity: There are two primary threats to the validity of this study that must be considered. First, either because of chance or corruption during the lottery, the randomization process may have produced unbalanced groups statistically different from one another. Second, as the study takes place over a number of years in an area characterized by seasonal migration, attrition rates are high. Attrition becomes a big problem if it is correlated with treatment status.

Testing Assumptions: The authors used the following checks to ensure that the two threats did not jeopardize the results of the study.

a. To minimize the chance that the randomization was prone to corruption, the lotteries were publicly conducted and the research team monitored each lottery. The authors also test that the randomization created statistically similar groups by showing baseline demographics, stove use, and health outcomes, across Lottery 1 winners, Lottery 2 winners, and those who lost both lotteries. The groups are well balanced across 59 baseline characteristics.

b. To combat attrition of study participants, households that could not be located were revisited. To check whether attrition was different between treatment and comparison groups, a dummy variable for survey attrition is regressed on the treatment dummy. Findings suggest that there is no significant difference in survey attrition for households and therefore, differential attrition is not a source of bias in the analysis.

D.2. ADDITIONAL CONSIDERATIONS

The following is a description of further issues that must be taken into consideration when conducting impact evaluation studies:

SAMPLE SIZE

If a program has a positive effect, an impact evaluation design will only be able to detect the positive effect if the sample size (number of units in the treatment and comparison groups) is sufficiently large. In the absence of a sufficient sample size it is possible that the study will not be able to detect a true effect when in fact there is one. The sample size necessary to overcome this problem is dependent upon a number of factors including the minimum effect size researchers expect to see (which should also be the minimum effect size large enough to warrant investing in the program) as well as the variability of the outcome across units in the sample.[56] The smaller the effect size or the larger the variability in outcomes, the larger the sample size needs to be to detect a statistically significant effect. As a practical matter, several different scenarios should be considered for a variety of effect sizes and population variabilities and the evaluator's choice of sample size should be justified on the basis of these calculations (known as "power calculations").

ATTRITION

Occurs when data for certain subjects of the study are not available due to nonresponse or subjects leaving the study. Attrition is a problem because it renders the existing sample less representative of the population, thus reducing the scope for generalization of results. When attrition is correlated to the intervention, the exiting sample in the group receiving the intervention and the comparison group are no longer similar, leading to bias in the estimated impact measure.

SPILLOVERS AND DIFFUSION

Evaluations must assume that treating one individual has no effect on the outcomes of other individuals. In the iron supplement example, it assumes that providing an adolescent girl with an iron supplement does not affect the hemoglobin levels of girls in the control group. This assumption is sometimes violated in social science experiments where interactions between people lead to treatments having far wider effects beyond the treated individual. This can occur for many reasons. For instance,

providing an individual with an insecticide treated bed net can affect his/her neighbors because mosquitoes killed by the net can no longer bite others.

CONTAMINATION

Contamination occurs when, for various reasons, the assignment into treatment and comparison groups did not work as intended. For example, this could happen if some control units were able to obtain the treatment (i.e. some girls assigned to the comparison group were able to obtain iron pills from the program) or some treatment units refused to take the treatment. Both of these issues are generally known as non-compliance. There are statistical fixes for non-compliance but solutions vary depending on the context and may not completely solve the problem.

APPENDIX D: DRAFTING A TERMS OF REFERENCE

To ensure that the evaluations being conducted are credible and well-planned, a few requirements should be set by the implementing organization and incorporated into the ToRs for evaluating agencies. TORs represent the basic contract with evaluators. They present an overview of the requirements and expectations of the implementing organization, and include specification of key research questions, justification of methodology, sampling, data collection, intervention monitoring, and analysis protocols.

RESEARCH QUESTIONS

The evaluating agency (the evaluator) should use the activity's theory of change as proposed by the implementing organization to identify and clarify specific research questions that closely reflect the outcome of interest. Research questions should also reflect the nature of information required by the stakeholders. Research questions therefore can be broad (e.g. "Does the new program that used fortified atta in midday meals reduce incidence of anemia among school children?") as well as more narrow (e.g. "Does the new program that used fortified atta in midday meals reduce incidences of anemia among girls attending public schools aged 12 and under?").

Evaluators also need to provide a list of indicators that they will collect to conduct the impact evaluation. This should include information on not only input, output, and outcome variables, but also data on a number of contextual variables such as program participant demographics and socioeconomic information. Indicators should be SMART. They should be Specific to the research objective, detailed, focused and well-defined; Measurable through objective quantitative or qualitative means, Achievable and can actually be collected at an acceptable cost, Relevant to the information needs of stakeholders, and Time-bound, able to measure changes in a specified and reasonable time frame.

EVALUATION METHODOLOGY

Once the research questions are defined, the methodology used to estimate impact must be selected and justified. Justification for using the methodology will involve elaborating on the identification strategy and what the features of the study design are, as well as the implementation location and participant eligibility that allow for the use of this methodology.

SAMPLING

The evaluator has to conduct power calculations to decide the sample size that is required to detect an impact. As mentioned above, a key determinant to power is the expected size of the impact of the program. The implementing agency should provide insight to the evaluator on the magnitude of change on an outcome of interest that the program is expected to cause in the timeframe given (this is known as the effect size).

The required sample size and sampling frame must be discussed prior to the start of any evaluation related activities, including data collection.

DATA, DATA COLLECTION, AND ENTRY

A variety of data sources can be used for the evaluation. The two main types of data used are administrative data and primary data, both of which are collected by the evaluator. If administrative data are being used, the evaluator should provide information on the source, authenticity, and accuracy of the data.

Primary data can be collected using a variety of methods such as firm, household, or individual surveys, individual cognitive or ability-based tests, focus group discussions, observation, etc. Certain data collection protocols should be followed to ensure quality of primary data collected. The first step in collecting primary data involves designing the data collection instruments, which should draw heavily on the theory of change. Protocols such as piloting, translation, and back translation must be followed while finalizing the instrument. If tests are being administered, reports on the validity of the tests in the actual context need to be provided.

Data collection instruments must be administered by trained surveyors. Activities of surveyors should be regularly monitored and audited. There are specific protocols to ensure accuracy in health measurements. Evaluators must also track the number of nonresponses and have a strategy to address attrition of human subjects.

Certain data entry protocols, such as double-data entry and error rate checks, need to be followed if data entry is required for data collected using paper questionnaires.

The evaluator must provide the data collection instruments and inform the implementer about the practices and protocols being followed to ensure the data collection and entry process are well managed and data quality is maintained. Data collection instruments should be reviewed and discussed by the implementing agency and the evaluator to make sure relevant information is collected in meaningful ways.

INTERVENTION MONITORING

If the planned evaluation is an RCT, the intervention should be monitored to ensure the integrity of the study design is maintained.

The evaluator must inform the implementing agency about intervention-monitoring practices being followed to ensure the integrity of the study design. If the possibility exists to align intervention monitoring with existing program monitoring structures, this should be explored.

ANALYSIS PLAN

Prior to conducting any analysis, it is important to have an analysis plan. The analysis plan should provide detailed information on the way the analysis will be conducted and the statistical models that will be run. Ideally this analysis plan will be placed in the public domain to generate transparency.

The evaluator must inform the implementer about their analysis plan to ensure that they are conducting the analysis as agreed and addressing the specific research questions.

ANJALI GUPTA | MATH GAMES' FUNDERS (UBS BANK, SWITZERLAND) VISITING A PRATHAM BALWADI (PRE-K) LOCATED IN KONDLI, NEW DELHI - VIEWING (AND PLAYING) MATH GAMES

REFERENCES AND RESOURCES

REFERENCES:

1. The World Bank Group. 2012. "Impact Evaluations: Relevance and Effectiveness." https://openknowledge.worldbank.org/handle/10986/13100

2. Abdul Latif Jameel Poverty Action Lab. "Impact Evaluation. Accessed on October 30, 2013. https://www.povertyactionlab.org/research-resources/introduction-evaluations

3. Patricia Rogers, and Better Evaluation. 2012. "Introduction to Impact Evaluation." https://www.interaction.org/sites/default/files/1%20-%20Introduction%20to%20Impact%20Evaluation.pdf

4. USAID Evaluation Policy. 2011. "Evaluation: Learning from Experience." https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf

5. Guhit, Irene. 2010. "Capacity Development in Practice, Chapter 21: Accountability and Learning." http://snv-website-2015.live.dpdk.com/public/cms/sites/default/files/explore/download/capacity_development_in_practice.pdf

6. Imas, Linda and Ray Rist. 2009. *The Road to Results.* World Bank Publications.

7. Gertler, Paul, et al. 2010. "Impact Evaluation in Practice." http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf

8. USAID. "USAID/India Country Development Cooperation Strategy, (2012-2016)." 2013. https://www.usaid.gov/sites/default/files/documents/1861/India_CDCS.pdf

9. Independent Evaluation Group. "Writing Terms of Reference for Evaluation: A How-to Guide." http://siteresources.worldbank.org/EXTEVACAPDEV/Resources/ecd_writing_TORs.pdf

10. Murnane, Richard J and John B. Willett. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* Oxford University Press.

11. J-PAL Executive Education Course. 2013. "Why Randomize?" http://www.clearsouthasia.org/workshopcourse-content/

12. Bamberger, Michael, Jim Rugh, and Linda Mabry. 2006. *Real World Evaluation.* Sage Publications.

13. World Bank Website. "Evaluation Designs." Accessed November 15, 2103. http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20188242~menuPK:412148~pagePK:148956~piPK:216618~theSitePK:384329,00.html

14. Judy Baker. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners.* http://siteresources.worldbank.org/INTISPMA/Resources/handbook.pdf

15. Shadish, William, Thomas Cook, and Donald Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin.

16. Community Interventions for Health. "Process Evaluation." Accessed November 20, 2013. http://www.oxha.org/cih_manual/index.php/process-evaluation.

17. Hempel, Kevin and Nathan Faila.2012. "Measuring Success of Youth Livelihood Intervention: A Practical Guide to M&E." http://www.iyfnet.org/sites/default/files/gpye-m&e-report.pdf

18. Abdul Latif Jameel Poverty Action Lab. "Methodology Table." http://www.povertyactionlab.org/sites/default/files/documents/Experimental%20Methodology%20Table.pdf.

19. Abdul Latif Jameel Poverty Action Lab, "When to Conduct an Evaluation." Accessed November 16, 2013. http://www.povertyactionlab.org/methodology/when/when-conduct-evaluation.

20. Khandker, Shahidur, Gayatri Koolwal, and Hussain Samad. 2010. *Handbook on Impact Evaluation.* Washington, DC: The World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/2693/520990PUB0EPI1101Official0Use0Only1.pdf?sequence=1.

21. World Health Organization. "Workbook 4: Process Evaluation." http://apps.who.int/iris/bitstream/10665/66584/5/WHO_MSD_MSB_00.2e.pdf

22. Institute of Education Sciences and National Science Foundation. 2013. "Common Guidelines for Education Research and Development: A Report from the Institute of Education Sciences, U.S. Department of Education and the National Science Foundation." http://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf.

23. Bloomquist, John. 2003. "Impact Evaluation of Social Programs: A Policy Perspective." Working Paper.

24. European Commission, Directorate General Regional and Urban Policy. 2013. "Guidance for Terms of Reference for Impact Evaluation: The Programming Period 2014-2020." European Regional Development Fund and Regional Fund.

25. OECD. "Outline of Principles of Impact Evaluation." http://www.oecd.org/development/evaluation/dcdndep/37671602.pdf.

26. Coalition for Evidence Based Policy. 2013. "Practical Evaluation Strategies for Building a Body of Proven-Effective Social Programs: Suggestions for Research and Program Funders." October 2013. http://coalition4evidence.org/wp-content/uploads/2014/05/Practical-Evaluation-Strategies-2013.pdf

27. ICCO Policy Department. 2004. "Guidelines: Terms of Reference for Evaluations."

28. Abdul Latif Jameel Poverty Action Lab. Randomization Designs. Accessed at http://www.povertyactionlab.org/sites/default/files/documents/Randomization%20Methods%20PDF.pdf

29. AusAid: Office of Development Effectiveness.: 2012. "Impact Evaluation: A Discussion Paper for AusAID Practioners." Accessed at: http://www.perfeval.pol.ulaval.ca/

30. Goldstein, Markus. 2012. "DFID's Approach to Impact Evaluation, Part I." Accessed at http://blogs.worldbank.org/impactevaluations/dfids-approach-to-impact-evaluation-part-i

31. United Nations Development Programme. "Annex 3: Evaluation terms of reference template and quality standards." *Handbook on Planning, Monitoring and Evaluation for Development Results.* Accessed at http://web.undp.org/evaluation/handbook/Annex3.html

32. United Nations Evaluation Group. 2010. "UNEG Quality Checklist for Evaluation Terms of Reference and Inception Reports." Accessed at http://www.uneval.org/papersandpubs/documentdetail.jsp?doc_id=608.

33. Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2006. "Using Randomization in Development Economics Research: A Toolkit."

34. Coalition for Evidence Based Policy. 2012. "Which Comparison Group (Quasi-Experimental) Study Designs are Most Likely to Produce Valid Estimates of a Program's Impact." Accessed at http://coalition4evidence.org/video-comparison-group-studies/.

35. Freedman, David A. "Statistical Models and Shoe Leather."1991. Sociological Methodology. Vol 21. 291-313. Accessed at http://siteresources.worldbank.org/INTISPMA/Resources/handbook.pdf

ADDITIONAL RESOURCES:

Glennerster, Rachel and Takavarasha, Kudzai. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.

Cartwright, Nancy and Hardie, Jeremy. 2012. *Evidenced-Based Policy: A Practical Guide to Doing It Better.* Oxford University Press. (Introduction and Chapter One).

SOURCES FOR FREE ACADEMIC PAPERS:

Abdul Latif Jameel Poverty Action Lab. http://www.povertyactionlab.org/evaluations

National Bureau of Economic Research Working Papers. http://www.nber.org/.

International Initiative for Impact Evaluation. http://www.3ieimpact.org/en/evidence/.

The Abdul Latif Jameel Poverty Action Lab (J-PAL) is a network of more than 140 affiliated professors from over 40 universities. Our mission is to reduce poverty by ensuring that policy is informed by scientific evidence. We engage with hundreds of partners around the world to conduct rigorous research, build capacity, share policy lessons, and scale up effective programs.

For more information, visit povertyactionlab.org.