Evaluation Strategies

COLUMBIA CENTER FOR THE STUDY OF DEVELOPMENT STRATEGIES (CSDS)

Contents

1. Overview

The CSDS is engaged in a series of impact evaluations of development policies and projects in order to learn lessons that have broad application.[1] In this memo we describe the kinds of strategies that we employ.[2] We emphasize the benefits but we outline the potential difficulties – including procedural and ethical questions – associated with these strategies.

The benefits of the approaches we outline are:

- **An excellent ability to estimate real impacts of programs**. Development professionals work hard to bring about positive effects. They and their funders want to know when their efforts are successful
- Generation of knowledge about the **relative effectiveness** of different approaches
- Generation of **systematic data** to inform programming

However the approach cannot be used everywhere since it:

- requires **early integration** of evaluation considerations into program design
- requires **sufficient numbers** of "units" (depending on the program, units might be individual beneficiaries or entire communities) to allow for the generation of precise estimates
- is **sometimes more costly** than evaluations that depend on short post-implementation visits by consultants (note that we provide our support *pro bono* but systematic data collection before and after the intervention can be costly)

[1] The Center for the Study of Development Strategies (CSDS) is a new center at Columbia University launched in Fall 2009 with core support from the Earth Institute and ISERP. The mission of the center is to apply innovative research methodologies to address major policy relevant questions in the political economy of development. The substantive focus on our projects is on interventions that address strategies of governance in developing areas.

[2] The note draws on material generated by a number of CSDS members and their collaborators including Ana Arjona, Michael Gilligan, Macartan Humphreys, Elisabeth King and Jeremy Weinstein.

2. Evaluation Questions and Hypotheses

Each of our studies is structured around the examination of a set of hypotheses. These hypotheses are typically generated together with practitioners and identify expected (or possible) effects of interventions.

The difficulty in developing and stating clear and testable hypotheses is that it often requires specifying more concrete claims than are required in grant applications and other writing about projects. In particular, specifying claims through hypotheses requires focusing only on *measurable* outcomes of programming which may not capture all the aspirations that programmers have for their programs.

There are a number of different types of hypothesis. First, hypotheses can be either *qualitative* or *quantitative*. By *qualitative* we mean that they suggest the direction of change, rather than the amount of change. Using these terms to describe hypotheses does not presuppose certain methods. For example, one can use quantitative survey methods to respond to a qualitative hypothesis.

An example of a qualitative hypothesis is:

- H_1: Gaining private rights to land will result in increased investments in agriculture

An example of a quantitative hypothesis is:

- H_2: Access to energy efficient stoves will reduce wood consumption by at least 30%

In addition some hypotheses are *predictive* and some are *exploratory*. Predictive hypothesis may be used to determine whether an intervention was successful or not in light of some previously identified set of criteria. Evidence for a predictive hypothesis suggests that a program met its objectives. In many cases, however, it is important to examine whether programs have a particular set of effects even if one is not confident that the program will have those effects. For example a practitioner implementing a program that gives access to micro credit might predict gains in welfare but might also be interested in learning whether the program improves women's status or political participation. The latter questions could form the basis of exploratory hypothesis, they can be used to learn about program impact but do not provide criteria for determining whether a program was formally "successful" or not.

When are hypotheses formed? A key feature of all hypotheses that we examine is that they are generated *before* a program is implemented. The reason for this is that when hypotheses are generated after program development there is a risk of the introduction of statistical bias. In particular, there may be a tendency to generate hypotheses based upon relationships that appear in the data, even if these relations do not have causal importance.

Agreeing together on a set of hypotheses is the first step in the design of an evaluation.

3. Approaches to Evaluation

At the heart of any good evaluation is a strategy to answer the question "*what would have happened in this program area if there had not been a program?*"

Simple "before-after" comparisons are, unfortunately, not enough because even if there are improvements over time in the program areas, it will be difficult to know whether these improvements are due to the project. One might find that since the program began, things in the program communities improved, and thus conclude that the program must be the cause. Further investigation, however, could show that conditions improved in *all* communities during the time the program was running—program and non-program communities alike. This might be because of something completely unrelated to the program, such as improving economic conditions or a change in government. In fact it is also quite possible that a program succeeds but indicators actually worsen in the communities that receive the program because of adverse events unrelated to the program (for example droughts, population movements and so on). In such cases a good design should be able to show that even though things became worse, the situation in the program community is still substantially better than it would have been in the absence of the program. For this reason too, simply talking to beneficiaries of a project is not enough since although the perspectives of beneficiaries is fundamentally important gathering these perspectives does not provide sufficient information to know how beneficiaries fared relative to non-beneficiaries.

In our analyses, we call the units which receive the project the "treatment" units (they can be sites, communities, villages, individuals etc), and the units that do not, the "control" units. These can also be referred to perhaps more simply as the "project" units and "comparison" units. The core of the evaluation will then consist in comparing the outcomes of the project units with those of the comparison units.

The key design challenge for a strong evaluation is to identify a good "control" group for each program area. A good control group should be in all ways identical to the treatment group except for the fact that it did not receive treatment. This is called *balance*.

Balance is very hard to achieve. For example in some cases project designers choose particular sites as treatment sites precisely because of some positive features, such as their investment possibilities. In such cases, unless the control sites are just as good in terms of investment prospects, the principle of balance is violated and the evaluation will not be able to tell whether differences in outcomes are really due to the program or whether they are instead due to fundamental differences in investment climates.

Factors like "investment climate" in this example are called "confounding factors." Confounding factors are characteristics that are correlated with both the likelihood that a given subject will receive the treatment and the likelihood that the treatment will succeed. Confounding factors complicate the researcher's ability to assess the effectiveness of the treatment because they

make it difficult to tell if differences between treated and control group outcomes are due to the treatment or to the confounding factor.

One might address the problem of balance by trying to think of all of the confounding factors and compare only those treated and control cases that are very similar on these factors. In that way one could plausibly claim that the only major difference between the treated and control cases is the fact that one received the treatment and the other did not. This technique is called "matching." While the matching approach is far superior to one that does not factor in the systematic differences between treated and control cases, the approach suffers from important shortcomings. First the analyst must think of *all* of the confounding factors that may affect the outcome of the treatment, a daunting task and one that is constrained by the imagination of the researcher. Second the analyst must obtain measures of all of these confounding factors. Unfortunately in many cases these factors may be unobserved so that measures are unobtainable.

Despite these challenges, two approaches hold particular promise for identifying causal effects without bias: the randomized evaluation approach and the regression discontinuity approach.

Randomized Impact Evaluations

The optimal way (from a learning perspective) to identify a good control group is through the method of "randomized intervention." Essentially the process of randomized intervention works as follows: if there are 100 people that will receive some treatment and 200 people who are eligible to receive the treatment, then 100 people are chosen randomly from the group of 200 eligible people and assigned the treatment. All 200 people, however, are tracked. The fact that the 100 are chosen randomly means that (in expectation) there is *no systematic difference* between those that did and those that did not receive the treatment – *the only systematic difference lies in the treatment itself*. Since our treated cases are drawn randomly from the whole sample, treated and control cases are just as likely to bear any particular confounding characteristic, and if the sample size and number of treated cases is large enough the treated and control cases will on average be very similar. The beauty of randomization (combined with a large enough sample) is that it renders our treated and control cases similar on average even *on factors that are unobserved and even on factors that the researcher does not know are confounding but in fact are.*

Often when researchers present evidence for the effect of a program, critics ask "but did you 'control' for this or that?" or "but how do you take account of all of the unique features of each unit?" The great advantage of a randomized evaluation approach is that one has always controlled for everything in the sense that there are no third factors that are systematically

related to treatment. As a result the findings do not depend on the idiosyncrasies of the treated units but on what they have in common – exposure to the treatment.[3]

The randomized evaluation approach is common in the study of health and technology, but is rarer in the examination of social interventions in developing countries. Yet it is an approach to which researchers and development organizations are increasingly turning in order to find reliable estimates of program impact.

The basic idea of the approach is simple. However, the approach requires more coordination between the evaluation team and implementing team than is typically the case. We describe these below.

The Regression Discontinuity Approach

In some cases (for practical or ethical reasons) randomization is simply not possible and second best strategies need to be found to identify good control units. One useful approach is the regression discontinuity approach.

This works as follows. Say that some program is going to be made available to a set of individuals. Ex ante we identify a pool of "potential beneficiaries" that is twice as large as the targeted beneficiary number.

These potential beneficiaries are all ranked on a set of relevant criteria, such as prior education levels, employment status, and so on. These criteria can be quantitative; but they can also include assessments from interviews or other qualitative information. These individual criteria are then aggregated into a single score and a threshold is identified. Candidates scoring above this threshold are admitted to the program, while those below are not. "Project" and "comparison" groups are then identified by selecting applicants that are close to this threshold on either side.

Using this method we can be sure that treated and control units are similar, especially around the threshold. Moreover, we have a direct measure of the main feature on which they differ (their score on the selection criteria). This information provides the key to estimating a program effect from comparing outcomes between these two groups. The advantage of this approach is that all that is needed is that the implementing agency uses a clear set of criteria (which can be turned into a score) upon which they make treatment assignment decisions. The disadvantage is that reliable estimates of impact can only be made for units right around the threshold.

---

[3] Technically randomization ensures that there is balance of this form "in expectation." Especially when one has a small number of units there can be a risk that third variables are correlated with treatment and outcomes by chance. It is quite possible that the treatment works differently in different areas, if this is true the randomization approach nevertheless succeeds in returning the "average treatment effect."

For those interested in the statistical strategies behind these two approaches, the core logic of each is illustrated in Figure 1.

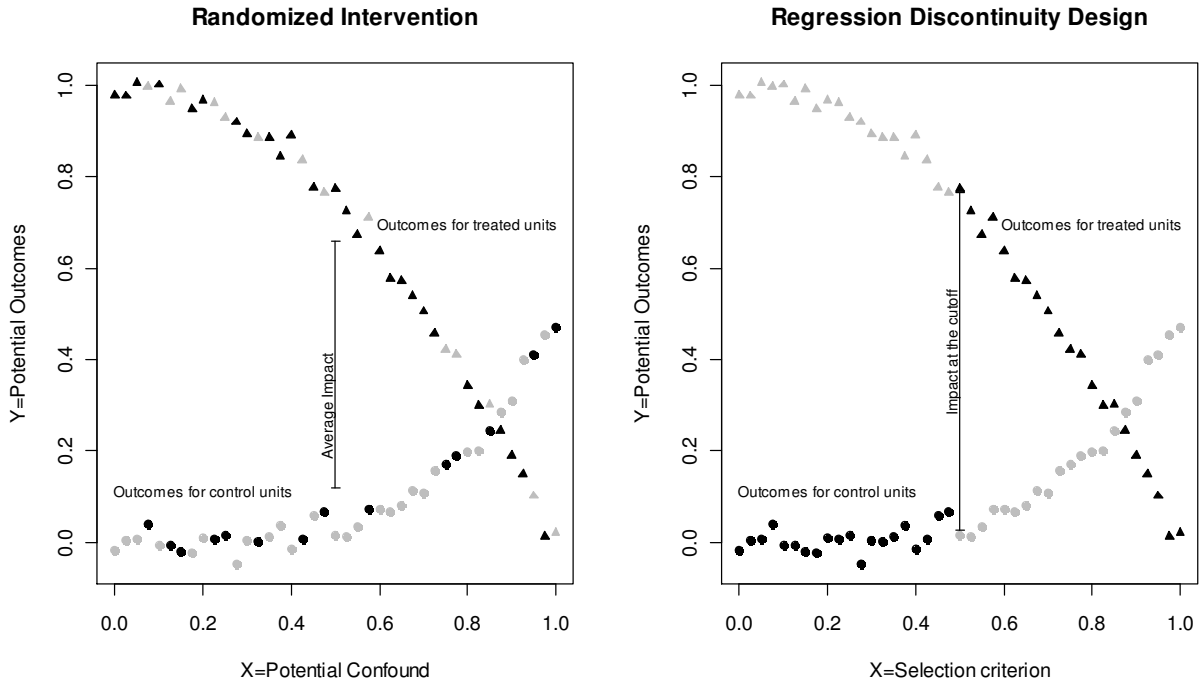

Figure 1: Illustration of randomized intervention and regression discontinuity approaches. Triangles represent the outcomes that units would have if they were treated; circles represent the outcomes that units would have if they were not. So each unit has one "potential outcome" for each condition. In fact however only one of these conditions is observed: each unit is either in treatment or in control. Black triangles then mark the actual values of treated units and black circles mark the actual outcome for control units. Hence all the dark points are *observed* and all the grey points are *counterfactuals*. Under randomization assignment to treated is unrelated to the confound X, and the difference between the average outcome for treated and control units give an estimate of the average treatment effect for all units. Under the regression discontinuity design, assignment to treatment is correlated with a confound, X, however we can still estimate the impact of treatment at the cut-off point. The estimates under these two designs need not be the same: the randomization approach provides the overall average treatment effect, the discontinuity approach estimates the treatment effect just for units close to the cutoff.

4. What are the steps of an evaluation process?

The evaluation process has the following elements:

- **Design**. CSDS develops hypotheses, research questions and measures together with implementing partners.
- **Funding**. If funding is not already in place, funding for the evaluation is sought.
- **Protocol development.** Terms of collaboration between researchers and program implementers are developed and design is sent out for peer review and consultation with an appropriate ethics review board.
- **Unit Identification**. Identify population of *potential* treatment units; typically this number will be about twice as large as the ultimate beneficiary population.
- **Baseline Measurement**. Once *all* potential beneficiaries are identified, collect a set of baseline measures for all units. Note that these baseline measures should ideally be collected after all potential units are identified, but *before* actual treatment and control units are determined.
- **Lottery / Selection**. The project sites are then randomly chosen in the randomized trial design (or the cut-off is identified in the regression discontinuity design) and the project is implemented.
- **Follow up Measurement**. Finally, after the completion of the project, a second set of measures is collected in both project and comparison communities. These can include quantitative, qualitative and behavioural components.
- **Analysis**. The core analysis consists in comparing the conditions in the sites where the program was run to conditions in the sites where it was not run. Depending on the exact design and type of data collection, the method can also let see what elements of the program are most effective.

5. What is the motivation of the research team?

The fundamental motivation of the research team is to find answers to major development challenges. We believe that with a better understanding of what sorts of interventions work where we can gain deeper knowledge into fundamental social processes, and also contribute to a knowledge base that can help inform future development strategies. Our motivation here is as researchers, not as consultants.

This motivation has implications for how and where we work.

- **Project Selection**: We focus on projects that have a strong potential for generating knowledge that has implications for other projects and processes around the world.
- **Remuneration**: In order to maintain independence CSDS typically undertakes work of this form on a *pro bono* basis. That is, we would not any seek remuneration from implementing agencies for taking part in this work.
- **Publication Rights:** While we will not seek remuneration, we will want to retain unrestricted rights to analyze and publish from data generated from the research. We will in all cases share all writing with the implementing partner for commentary and feedback before publication but to maintain the integrity of research, academic output is not subject to approval by partner organizations. These are elements that we would put into a memorandum of understanding.
- **Project costs**: We do not finance project or research costs (e.g. travel, data collection, dissemination of results) beyond our time contribution. In some cases evaluation budgets are written in to project grants; in others independent funding needs to be sought.

6. Where might we run in to problems?

From working on related interventions elsewhere we think we can see ahead of time where difficulties are most likely to area. We highlight these here:

- **Statistical Power**: For these approaches to work one needs a fairly substantial number of units (individuals or communities, depending on the project). This gives what we call statistical "power". Typically one needs on the order of about 100 treated and 100 control units although numbers depend on the likely size of effects. These approaches cannot be used with numbers in the order of 5 to 10 units. Also power depends to some extent on how close units are to each other; if units are too close and there is a risk that "control" units learn from "treated" units then there is a risk of underestimating the effectiveness of a program.
- **Timing**: Time is needed between the identification of all potential sites and the selection process (whether it be through lotteries or through identification of selection criteria and a cut off point) in order to undertake a baseline survey (where relevant). Depending on the scale this could take perhaps 1 to 3 months. This pause can sometimes be inconsistent with the natural schedule of a project.
- **Randomization**: The strongest evaluation strategy requires randomization of program assignment at least for some (though perhaps not all) units. Some people do not like the idea of using randomization as a part of development processes. Indeed in some cases randomization is not appropriate, but there are however quite strong ethical reasons for why in some cases randomization might be an especially fair approach. Moreover there are different ways of doing randomization, in some cases simply varying the *timing* of which communities are worked with when can be sufficient. If partner organizations are uncomfortable with the idea of randomization then this approach will not work.
- **Distrust**: Sometimes there can be distrust between practitioners and researchers especially if practitioners think researchers are trying to make judgments on how well practitioners are doing their job or if practitioners feel that researchers are only interested in abstract questions that are not relevant for the organization. To avoid this it is best if there is clear understanding of *why* CSDS researchers are willing to get involved and also to make sure that the questions being addressed really are of joint interest.

All of these difficulties are things that can and have been overcome in many projects. We raise them early to anticipate potential problems and how they may be addressed.

7. Practical Implications for Programming

A robust evaluation strategy involves a substantial commitment of both time and money, and requires that the evaluation team works closely with the project team throughout the lifespan of the project. Unlike many evaluations that are put together after programs are completed, a randomized or regression discontinuity evaluation is built into a program from the very start. This has implications in a number of areas.

- **Criteria for selecting units**. Before any sites are selected, criteria need to be established upon which all units will be selected. We can help establish these criteria, but will of course look to you to lead this process.
- **Number of sites to be identified**. Based on the criteria, typically twice as many sites need to be identified as will eventually be selected.
- **Lotteries**. In the case of randomized interventions, once all potential sites are identified, lotteries are held to determine which sites will be project sites. All sites must have an equal chance of receiving the project. This is crucial! If units are selected in advance of the lottery then they cannot form a part of the evaluation (If there are some areas that are clear priority sites these should naturally be removed from the public lotteries; in this case however these sites cannot be a part of the impact evaluation). This lottery approach is a transparent and equitable approach to identifying sites as long as all sites in the lottery pool are approximately in equal need of the project. In some contexts political and organization pressures may be such that the outcomes of lotteries are ignored and politically "favored" sites are selected. If this occurs, estimates from the evaluation will not be valid. If the lotteries are to be held in public then we will work with you to develop clear protocols for running these lotteries.
- **Data collection**. We will need to collect data on *all* of the treatment and control communities. If we were to collect data from only some of the communities, the power of our findings would be significantly diminished. We would not have enough statistical power in our data to draw meaningful conclusions. There are two implications of this part of the process. First we will look to you to help develop measures that are appropriate to the local context. Second we will need to coordinate schedules to ensure that the baseline data collection happens *after* the determination of the pool of sites but *before* the actual selection.

8. A Few Questions & Answers

Here are answers to a few common questions on impact evaluation. We welcome any more you have!

Q: Why is the randomization process so important?

A: The random selection is so important because we want the program and comparison communities to be as similar as possible at the outset. That way, if we see that by the end of the program, conditions have improved in the program communities, we can conclude that the cause of the improvement must have been the program, since the program and comparison communities were in all other ways the same.

We recognize that randomly choosing treatment communities from a larger population makes the implementation process a little more complicated than it would be otherwise. However, it is very important to be able to have comparison communities. Sometimes you will see evaluations where the evaluator has only tracked progress in the program communities, ignoring all other areas. The evaluator might see that since the program began, things in the program communities improved, and thus conclude that the program must have helped. Further investigation, however, could show that conditions improved in *all* communities during the time the program was running—program and non-program communities alike. This might be because of something completely unrelated to the program, like improving economic conditions or a change in government.

More commonly however people track outcomes in communities that they want to use as comparison areas but which were not part of a lottery process. This can cause important problems for assessing impact since there are typically some systematic features that distinguish the project communities from the comparison communities. For example it could be that one area was selected and another not because it is worse off, or more prone to conflict and so on. But then, if things look no better in the first community than in the second at the end of the project we could be wrong to infer that the program had no affect; it is possible that it did have an effect precisely by making the worse off community as well off as the better off community.

Q: Is this evaluation strategy ethical? Isn't it unfair to choose some communities and not others? Will choosing some communities lessen the chances for the comparison communities to one day receive projects of their own?

A: Ethics is very important to us. All development projects are implemented in some places and not in others. The strategy here does nothing to prevent projects happening in any areas or to affect the total number of projects. Rather the focus is on identifying specific areas where the project will not take place in order to learn about the areas where it does take place. A couple of principles are important however. First: care should be taken to ensure that the set of areas

that are identified for the lottery are all needy areas; areas that are not needy should not be included just to make up the numbers. Second: if it turns out that there are some areas that are clearly needier than all the rest, it may be inappropriate to include these in a lottery. In this case the program should be implemented in these areas and the lottery should be held over all the other, equally needy areas. In this case the neediest areas will not be integrated into the evaluation. Third it should be noted that random selection among equally needy areas is often more equitable than other forms of program allocation. If the selection process is by public lottery, community members will know how and why sites were chosen; they will know that they had the same chance as the others and that they were not discriminated against. Finally, being a ”comparison” community does not mean that community should in any way be prevented from receiving programming of their own from any other practitioners or from other initiatives.

Q: Has this kind of an evaluation strategy been done before?

A: This method of randomization has been used for project evaluation in a host of recent development projects including community driven development interventions by IRC and DFID in Liberia and Congo, by the World Bank in Indonesia and Sierra Leone, and by the NSP in Afghanistan. It is being very well received by the policy and donor community. This type of evaluation strategy is growing in popularity and is now strongly promoted by some donors, such as DfID, for major projects.

Q: Is this approach consistent with other approaches such as qualitative approaches or participatory rural appraisals?

A: Yes. The core to this approach is working out what the appropriate comparisons are in order to work out impact. Actual data collection and analysis can then be done using multiple approaches. We expect that a quantitative approach will play a central role; however combining this perspective with qualitative and ethnographic research in the study area can often lead to more insightful results.