Guide on the social experimentation protocol

The principle of social experimentations is to experiment a policy intervention on a small population so as to evaluate its efficacy before deciding whether it should be scaled-up. Therefore, social experimentations require both designing a potentially relevant policy intervention and measuring its actual efficacy.

This guide is intended at policy makers willing to embark in a social experimentation in the context of the PROGRESS call for proposals. It is divided in three parts. The first part briefly lays out basic principles to follow in order to design a potentially relevant policy intervention. Those principles are illustrated with an example. The second and third parts present six methods commonly used to measure the efficacy of a social program. In the second part, the principle of each method is presented, and methods are compared from the point of view of the reliability of the results they deliver. The third part considers the costs induced by each method, and their complexity to implement in practice.

There are two main actors involved in social experimentations: policy makers and evaluation teams which are usually made up of consultants or researchers. Policy makers’ role is both to design the policy intervention and to support the implementation of the experimental protocol by the evaluation team in the administrations where the experimentation is conducted. The evaluation team might be given a say in the design of the policy intervention, but its main role is to design the experimentation protocol, to implement it and to collect and analyze the data necessary to measure the efficacy of the program.

This guide aims at giving policy makers an overview of the steps to follow to conduct a rigorous experimentation of a policy intervention. Doing so, it aims at improving their understanding of the work conducted by the evaluation team so as to facilitate mutual understanding.
Part 1: Designing a potentially relevant policy intervention

Some principles to design the policy intervention

In the beginning of a social experimentation, it is important to describe rigorously the social need the program seeks to address, and to document its existence. Then, one should describe precisely the set of actions envisaged as part of the policy intervention and explain why they might help addressing this social need. In particular, one should ensure that the different incentives, opportunities or constraints to which the population will be confronted with are identified and described, and that the program is compatible with those incentives, opportunities and constraints, to ensure that the targeted population will indeed be willing and able to participate.

The relevance of the policy intervention should be supported by a thorough search of all examples of similar policy interventions conducted either at home or in other countries. This search should bring supplementary evidence that the program might indeed help answering the social need. Scientific databases such as Jstor (http://www.jstor.org/) for studies in economics, sociology and public policy, or PubMed (http://www.ncbi.nlm.nih.gov/pubmed) for studies in public health might help to conduct such a search. One might also be interested in screening the archives of consulting firms and research teams specialized in impact evaluation such as MDRC (http://www.mdrc.org/), Mathematica Policy Research (www.mathematica-mpr.com/), the Rand Corporation (http://www.rand.org/), the Abdul-Latif Jameel Poverty Action Lab (www.povertyactionlab.org/), Innovations for Poverty Action (http://www.poverty-action.org/) or the Institute for Fiscal Studies (http://www ifs.org.uk/).

The set of outcomes on which the policy intervention is expected to have an impact should also be defined precisely.

Finally, it is important to involve all relevant stakeholders from the moment the experimentation project starts being discussed, in order to build consensus on the policy intervention, the methodology used in the evaluation and the set of outcomes which will be considered during the experimentation. A consensus should also be built beforehand on a set of conditions for scaling up, if results are positive.

Illustrating those principles through an example.

Let us illustrate these points with an example which will be used throughout this guide. It is freely inspired from an experimentation which took place in France in 2008.

There is a growing recognition in France that universities may not prepare their graduates very effectively for finding employment. In 2007, three years after they completed their course, only 70% of University graduates had a stable job. The social need addressed here is therefore to improve the insertion of young graduates on the labor market. Such low placement rates might be due to the fact that internships are not mandatory in French universities and that workshops during which companies come to present the type of positions available for young graduates are seldom organized, so that most graduates leave university
without having ever searched for a job, with little knowledge of the jobs they should apply for, with no experience of how to get prepared for a job interview and sometimes even without having ever written a resume.

Consequently, the French ministry of labor decided to experiment in 100 placement agencies a counseling program to help those young graduates having spent at least 6 months in unemployment to find a job. **The principle of the program was to increase during 6 months the frequency of the meetings between young graduates and their counselor from once per month to once per week.** This was supposed to allow the counselor to spend more time helping the young graduate to think about his professional plans, to write a proper resume, to organize his search, to prepare the next job interviews he would have during the following week…

The **constraints of the targeted population** were carefully taken into consideration. The question was to know whether young graduates would have time and would be willing to meet once per week with their counselor. To answer it, the placement agencies participating in the experiment conducted a survey to ask young graduates they followed whether they would be willing to meet once per week with their counselor, and if they thought this would help. It appeared that young graduates indeed thought the intervention to be both feasible for them and useful.

The relevance of this policy intervention was confirmed by a thorough literature search which found 6 previous studies in which counseling increased the placement rate of unemployed against 0 studies finding no impact or a negative impact. However, none of those studies specifically focused on young graduates, so that the question of whether counseling is effective within this specific population was still an open question in the beginning of the experimentation.

Finally, it was decided that the main outcome to be taken into consideration to assess the efficacy of this program would be the share of young graduates having found a durable job after those 6 months of intensive counseling. A young graduate was considered as having found a durable job when he had signed a contract for a job lasting more than six months. Hereafter, this outcome is referred to as the placement rate of those young graduates.

All relevant stakeholders were involved throughout the design of the policy intervention. In particular **a draft of the experimental protocol was sent to counselors along with a questionnaire** in which they were asked to give their opinion on the policy intervention and on the experimental protocol, and suggestions on how to improve both of them.

Before turning to the choice of a method to evaluate the policy intervention, let us finally insist that the steps described above are crucial to the success of the experimentation. Social experimentations are long and costly; they require a substantial amount of supplementary work from policy makers and civil servants working in the departments where the experimentation is implemented. Therefore, it is worth designing carefully a presumably relevant policy intervention. **There is no point in evaluating a policy intervention which**
has been shown to be useless by dozens of previous studies. Similarly, there is no point in implementing a program to which targeted beneficiaries do not want or are not able to participate in. These might sound like obvious recommendations. However, a great number of experimentations fail because the targeted population refused to participate in the program, or because far less people than expected finally enrolled. Among programs which prove ineffective in the end of a long and costly experimentation, many could have been regarded as presumably ineffective beforehand, merely after conducting a thorough literature search. This would have saved the costs of an useless experimentation.
Part 2: Measuring the efficacy of a policy intervention: some methodological considerations

The conceptual problem of impact evaluation

The efficacy of a policy intervention is its capacity in fulfilling the social need it has been designed to address. This requires measuring its impact on its recipients: what does the policy intervention change in the lives of those who benefit from it? Does the intensive counseling program allow some young graduates to find a job they would not have been able to find otherwise?

The impact of a social policy or more generally of a “treatment” is defined as follows. It is the difference between what happens to beneficiaries after receiving the social policy and what would have happened to them anyway. In the context of the counseling program, its “impact” can be defined as the difference between the placement rate of young unemployed after receiving some intensive counseling and the placement rate that the exact same young unemployed would have obtained if they had not received intensive counseling.

\[ \text{True Impact} = \text{placement rate with counseling} - \text{placement rate without counseling} \]

The goal of impact evaluation is to reconstruct this counterfactual scenario, i.e. the placement rate those young graduates would have obtained if they had not received intensive counseling. In order to do so, we must find a comparison group which did not receive the counseling, so as to compare the placement rate in the group which received the counseling (treated group or treatment group) to the placement rate in the comparison group. Ideally, those two groups should be similar in every respect, except that one group received intensive counseling whereas the other one did not. This would ensure that the difference in placement rates across the two groups is really attributable to the intensive counseling program and not to other differences across those two groups.

Random assignment to treatment is regarded as a “gold standard” to construct valid comparison groups. But as we will emphasize in the next part of this guide, randomized
experimentations of social programs take time and are somewhat complex to implement. Therefore, other techniques are also commonly used. They are referred to as non-experimental methods. They are usually less costly and less complex to implement than randomized evaluations, but the results they deliver are also less reliable. The reason for this is that randomization ensures that by construction the comparison group is indeed similar in every respect to the treated group. On the contrary, each non-experimental method must rely on one assumption to justify the claim that the comparison group it uses is indeed similar to the treated group. Results from non-experimental methods will be all the more credible that this assumption is credible in the context under consideration.

In the six following sections, the six most commonly used methods in the impact evaluation literature are presented. The emphasis is laid on the principle of each method and the assumption on which it relies. Methods are “ranked” according to the degree of credibility of their underlying assumption. The first two methods presented rely on fairly incredible assumptions but they are still presented here because they allow gaining a better understanding of the conceptual challenge faced when conducting an impact evaluation. The two following methods rely on much more credible but still strong assumptions. Finally the last two methods rely on either fairly innocuous assumptions or no assumption at all. The last section compares results obtained through different methods.

As illustrated through the counseling program example, results of the experimentation are highly dependent on the method used: each method will deliver its own measure of the impact of the treatment (Impact 1, Impact 2, …, Impact 6). As it has been often shown, when two different methods are used to measure the impact of a program from the same experimentation, one method might conclude that the program had a positive impact whereas the other method might find a negative impact. Hence the need to bear in mind the assumptions under which each of those methods to rely, so as to be able to assess which assumptions are the most credible in the particular situation under consideration.

Comparing participants and non participants

A first way to measure the impact of a program might be to use individuals eligible to the program but who choose not to participate as a comparison group.

Assume for instance that 10 000 young graduates have been offered to participate in the intensive counseling program but 2 000 declined this offer. To measure the impact of the counseling program one could compare placement rates among the 8 000 young unemployed who chose to participate and the 2 000 who chose not to participate. If the placement rate among participants was equal to 50% against 35% among non participants, using this methodology would conduct policy makers to conclude that the program increases the placement rate by 15 percentage points.

\[ \text{Impact 1} = \text{participants placement rate} - \text{non participants placement rate} \]
\[ = 50\% - 35\% = +15\% \]
However, for this measure to be equal to the true impact of the program, the treated group and the comparison group should be comparable in every respect. In the context of the counseling program, this means that unemployed who participated in the program should be similar to those who did not participate in terms of gender, qualifications, motivation to find a job… In this particular example, this is very likely not to be the case.

Those who chose not to participate to the counseling program might differ from those who chose to participate on observable dimensions (age, gender…). Assume for instance that males were more reluctant in participating to the counseling program than females, so that there is a larger share of males in the comparison group than in the treated group. Since in France women face higher unemployment rates and therefore lower placement rates than males, the comparison of placement rates of participants and non participants to the program will capture both the effect of the training program, and the fact that the treated group comprised more women who were less likely to find a job anyway.

More importantly, the treated group and the comparison group might also differ on dimensions which are very difficult to measure precisely (hereafter referred to as “unobservable dimensions”) such as their motivation to find a job. One could for instance argue that non-participants were probably less motivated to find a job which is the reason why they declined the offer.

Overall, this 15 percentage point difference might capture three things: the true effect of the counseling program, the fact that participants differed from non participants on observable characteristics (age, gender…), and the fact that they also differed on unobservable characteristics such as their motivation to find a job.

Before-after

Before-after amounts to using as a comparison group the exact same population which became eligible to the program once it was implemented before the program was implemented. The French counseling program has been offered to young graduates having
spent at least 6 months in unemployment in 2008. As a comparison group, one might use young graduates having spent at least 6 months in unemployment in 2007 and counseled by the exact same placement agencies. Indeed, this population did not benefit from intensive counseling but might be similar to the population which received counseling.

As per the before-after methodology, the impact of the counseling program is merely equal to the evolution of the placement rate of young graduates after the counseling program was implemented.

The before-after methodology relies on a strong assumption, which is that the placement rate of young graduates would have been the same in 2008 than in 2007 if the counseling program had not been implemented. This assumption might be violated in some instances. For example, it might be the case that an economic recession happened in 2008. In this case, one might find that the placement rate of 2007 young graduates was equal to 55% against 50% only in 2008. Using a before-after methodology in this context would lead policymakers to wrongly assess that the counseling program has a negative impact. They would indeed conclude that it diminishes by 5 percentage points the percentage of young unemployed finding a job in less than 6 months, whereas this decrease is probably at least partly due to the economic recession and not to the implementation of the counseling program.

\[ \text{Impact 2} = \text{placement rate after} - \text{placement rate before} \]
\[ = 50\% - 55\% = -5\% \]

Figure 2: Before-After

![Placement rate graph](image)
Statistical matching

Statistical matching builds upon the same intuition as the comparison of beneficiaries and non-beneficiaries. Instead of comparing all beneficiaries and all non-beneficiaries of the program, pairs of beneficiaries and non-beneficiaries resembling each other are constructed and the comparison is conducted only within those pairs.

In the context of the counseling program, one should associate to each non-participant the participant who resembles her most on a number of characteristics. Those characteristics should be both easy to observe and important determinants of the chances that a young graduate will find a job. One could for instance think of her age, her gender, her previous work experience and her qualifications. One thus ends up with 2 000 pairs of “twins”, each pair being made up of one participant and one non-participant extremely similar on those four characteristics. The impact of the counseling program is then computed as the difference between the placement rate among those 2 000 participants who have been selected as twins of non-participants, and the placement rate among non-participants. If placement rate among “twin” participants was equal to 44% (instead of 50% among all participants), against 35% among non-participants, the impact of the program as per this methodology is to increase placement rate by 9 percentage points.

\[
\text{Impact 3} = \text{"twins" participant placement rate} - \text{non participant placement rate}
\]

\[
= 44\% - 35\% = +9\%
\]

Figure 3: Matching

Statistical matching allows improving a lot on the mere comparison of beneficiaries and non-beneficiaries. Indeed, it ensures that by construction the two groups which are compared are indeed very similar with respect to important observable characteristics used in the matching procedure, i.e. age, gender, previous work experience and qualifications in the counseling example. But those two groups might still differ on unobservable dimensions. In the counseling example, one might for instance argue again that the non-
participants were probably less motivated to find a job. Therefore, this 9 percentage points difference might still capture two things: the effect of the counseling program and the mere fact that participants and non participants differ on unobservable dimensions such as their motivation to find a job.

**Difference in differences (DID)**

Difference in differences is a refined version of the Before-After methodology. It amounts to comparing the evolution of the placement rate across two groups, the group eligible to the counseling program and a group not eligible to it, for instance uneducated young unemployed.

The mere evolution of the placement rate in the treated group between 2007 and 2008 might not yield the true impact of the counseling program, for instance because economic conditions changed between 2007 and 2008. Therefore, to recover the true impact of the counseling program, one should compare this evolution to the same evolution within a group which was not eligible to the counseling program in 2008 (control group). Indeed, the evolution of the placement rate within this control group will capture the mere effect of the change in economic conditions from 2007 to 2008. And the difference between those two evolutions will better capture what is specifically attributable to the program.

Assume for instance that the placement rate of uneducated young unemployed diminished from 42% to 30% from 2007 to 2008 while the placement rate of young graduates diminished from 55% to 50%. Then, as per the difference in differences methodology, the counseling program increases the placement rate by 7 percentage points:

\[
\text{Impact 4} = \text{change of placement among eligible} - \text{change of placement among ineligible}
\]

\[
= (50\% - 55\%) - (30\% - 42\%) = -5\% - (-12\%) = +7\%
\]
The fact that the placement rate decreased less between 2007 and 2008 among young graduates than among uneducated young unemployed indeed suggests that the counseling program had a positive impact since it allowed young graduates to suffer less from the economic recession than young unemployed with no degree.

However, for impact 4 to be exactly equal to the true impact, a strong assumption must be verified. It states that if the counseling program had not been implemented, the placement rate of young graduates would have diminished by the exact same amount than the placement rate of uneducated young unemployed. Putting it in other words, it states that if the counseling program had not been implemented, the blue line and the green line would have followed parallel evolutions. This is the reason why it is referred to as the “parallel trends assumption”. This assumption might also be violated. One could for instance argue that the recession has probably hit more strongly the uneducated group. Indeed, they might be more vulnerable to macroeconomic shocks because of their lack of qualification.

One way to test the credibility of the “parallel trends assumption” is to check whether placement rates in the two groups indeed followed parallel evolutions previous to the reform. Assume for instance that data on placement rates within those two populations is available from 2000 to 2008, and that placement rates in the two populations evolved as in Figure 5.
The parallel trend assumption states that placement rates would have followed a parallel evolution from 2007 to 2008 in the two groups if the counseling program had not been implemented. The fact that placement rates in the two groups indeed followed a parallel evolution between 2000 and 2007 gives some credibility to this assumption. Since graduates and uneducated unemployed have been affected similarly by all macroeconomic shocks which happened between 2000 and 2007, there is no reason to suspect that they would have been affected very differently by shocks happening between 2007 and 2008.

On the contrary, the parallel trend assumption would be challenged if the graph actually looked like Figure 6. The fact that placement rates in the two groups followed very different evolutions between 2000 and 2007 undermines the credibility of this assumption. Since in this scenario graduates and uneducated unemployed have been affected very differently by the macroeconomic shocks happening between 2000 and 2007, there is no reason why they should have been affected similarly by shocks happening between 2007 and 2008.
Such a test of the validity of the parallel trend assumption should be conducted beforehand, when designing the experiment. Once a control group has been found, one should construct graphs similar to Figures 5 and 6. If the resulting graph rather looks like figure 5, this will give some support to the parallel trend assumption. If it rather looks like figure 6, this will strongly undermine it, so that one should try to find another control group.

**Regression Discontinuity**

Eligibility to some programs is determined according to whether a participant is above or below a threshold taking a continuous set of numerical values, such as age or income. In such instances, one can measure the impact of the program through a technique called “regression discontinuity”. Its principle is merely to compare participants to the program “very close” from being ineligible because they are slightly above the threshold to non-participants “very close” from being eligible because they are slightly below the threshold.

Let us illustrate this with the counseling program example. Assume that this program was accessible only to young graduates who were strictly less than 28 years old when they reached their 6th month in unemployment. Then, one could measure the impact of the counseling program comparing placement rates of those who were 27 years and 11 months old when they reached their 6th month in unemployment to the placement rate of those who were 28 years and 0 month old. Indeed, the first group benefited from counseling whereas the other one did not. If the placement rate of those 27 years and 11 months old was equal to 53%, against 49% for those who were 28 years and 0 month old, then, as per the regression discontinuity methodology, the program increased the placement rate by 4 percentage points.
The underlying assumption to regression discontinuity is that the placement rate of those who reached their 6\textsuperscript{th} month in unemployment at 28 years and 0 month is representative of the placement rate that would have obtained those who reached their 6\textsuperscript{th} month in unemployment when they were 27 years and 11 months old if they had not followed the counseling program. In this context, this seems like a fairly reasonable assumption: there is no reason why those two groups should strongly differ. The graph below, which plots placement rates according to age when reaching 6 month in unemployment gives some support to this assumption. Indeed, differences in placement rates across age cohorts are extremely small (1 percentage point at most), except between the two cohorts which reached 6\textsuperscript{th} month in unemployment at 27 years old and 11 months and 28 years old, that is to say precisely at the eligibility threshold.

However, this assumption might be violated when people can manipulate the variable on which eligibility to the program is decided. In the counseling program example, young graduates cannot lie on their age so as to enter the program so that this is not a concern. But this can be an issue in other contexts. Let us consider the example of a
microcredit program in Mexico to which only farmers owning strictly less than 2 acres of land were eligible. It happened that to become eligible to this program, many farmers owning slightly more than 2 acres sold temporarily part of their land to become eligible to the program. In such circumstances, farmers owning slightly less than 2 acres are not comparable to those owning slightly more. Indeed, the population of farmers owning slightly less than 2 acres include both those really owning less than 2 acres and those who used to own more but who were clever enough to sell part of their land to get into the program. On the contrary, the population of farmers owning slightly more than 2 acres only include those who were not clever enough (or did not want) to sell part of their land to benefit from the microcredit. Still, there are simple tests to detect such manipulations. In the Mexican example, researchers found that there were many more farmers owning slightly less than 2 acres than farmers owning slightly more than 2 acres. In theory there should be approximately the same number of farmers slightly below than slightly above the threshold. Therefore, this gave a good indication that the threshold had indeed been manipulated.

Another limitation of results obtained from a regression discontinuity is that it measures the impact of the program only on people close from the eligibility threshold, i.e. on people close from 28 years old when reaching their 6 months in unemployment. Policy makers might be interested in knowing the impact of this program not only among this subgroup but among the whole population, in which case regression discontinuity will be useless. Therefore, when interpreting results of a regression discontinuity study, one should keep in mind that results apply only to people close from the eligibility threshold.

**Randomized experiments**

*Randomized experiments deliver a measure of the true impact of the program*

Randomized experiments are experimentations of social policies in which assignment to treatment is based on the results of a lottery. Assume that the 100 French local agencies participating in the experimentation of the counseling program followed 10 000 young graduates eligible to this program in 2008. Evaluating this program through a randomized experiment requires to select randomly 5 000 young graduates who will indeed receive intensive counseling (the treatment group) and 5 000 who will not receive it (the comparison group).

The impact of the program will be measured comparing the placement rate among those two groups. Assume for instance that 55% of lottery winners had found a job in less than 6 months against 50% only among the comparison group. Then, as per this randomized experiment, the impact of intensive counseling is to increase the placement rate by 5 percentage points.

\[
\text{Impact } 6 = \text{lottery "winners" placement rate} - \text{lottery "losers" placement rate} = 55\% - 50\% = +5\%
\]
Randomization ensures that the treated group and the comparison group are comparable in every respect (age, proportion of males, qualifications, motivation, experience, cognitive abilities…). Indeed, when a population is randomly split into two groups, the two groups will have extremely similar characteristics provided the population is sufficiently large.

To understand why, assume that among the initial pool of 10,000 young graduates eligible to the counseling program, 5,000 were not motivated at all to find a job and 5,000 were extremely motivated. When selecting randomly the 5,000 unemployed who will indeed receive intensive counseling, we could end up selecting 4,000 extremely motivated and 1,000 unmotivated unemployed. If this were to be the case, then the treated group and the comparison group would not be comparable at all: the treated group would comprise 4,000 extremely motivated unemployed against only 1,000 in the control group.

But the probability that this scenario happens is the same as the probability of getting 4,000 heads when tossing 5,000 times a fair coin, that is to say approximately 0. Indeed, when tossing 5,000 times a fair coin, the probability to get heads in each of those draw is equal to one half. Therefore, we expect to get heads in approximately half of the draws that is to say 2,500 times. Getting 4,000 heads is so far from the scenario we should observe in average (2,500 heads) that the probability that it indeed happens is almost 0. Actually, one can compute that when drawing 5,000 times a fair coin, one has a 95% probability to get between 2,430 and 2,570 heads. Getting back to the intensive counseling example, this means that when selecting randomly the 5,000 unemployed who will indeed receive intensive counseling we have a 95% probability that the number of extremely motivated unemployed we send to the intensive counseling group will be included between 2,430 and 2,570, in which case the treated group and the comparison group will be comparable.
Therefore, randomization ensures that the treated group and the comparison group are comparable on every dimension (age, proportion of males, qualifications, motivation, experience, cognitive abilities…). The only difference between those two groups is that one receives intensive counseling and not the other one. Consequently, if we observe that the placement rate is higher in the treated group than in the comparison group after 6 months, this means that intensive counseling is effective. Because randomization ensures that the two groups are comparable in every respect, the placement rate in the comparison group is representative of the placement rate that we would have observed in the treated group if it had remained untreated. Therefore, randomized experiments allow measuring the true impact of intensive counseling.

A nice feature of randomized experiments is that they enable determining easily whether the effect of the program is the same within different subgroups of the population so as to optimize eligibility criteria when scaling up. For instance, the comparison of placement rates among males in the treatment and in the comparison group yields a measure of the impact of the intensive counseling program on males. The same comparison among females yields a measure of the impact of the program on females. If one were to find that the program had a very large impact on females and virtually no impact on males, this would allow policy makers to restrict access of the program to females only when scaling up. Such a comparison of the effect of the program across subgroups is also possible with other methods but will require even more assumptions.

**Alternative designs**

There are alternative designs to randomized experiments. For instance, one might not be willing to exclude anyone from treatment. In such circumstances, it is possible to implement what is referred to as the phasing-in design. Assume that sufficient money is available to hire extra teachers for all schools in an area, but that half of the money will be available this year and that the remaining half will come the next year. One could randomly sort out schools which will hire their extra teacher the first year (treatment group) and those which will hire them the following year (comparison group). Comparing pupils test score in the two groups in the end of the first year would yield a measure of the impact of giving schools an extra teacher for a year. The main advantage of the phasing-in design is its great degree of acceptability: the experimental protocol will seem much more acceptable to the comparison group if they are being told they are temporarily deprived from treatment but will get access to it soon. The main disadvantage of the phasing-in design is that it prevents from measuring the long-run impact of the program: in year two, both treatment and comparison schools will have an extra teacher.

Another alternative design enabling not to exclude anyone from treatment is the encouragement design: an encouragement for treatment is randomly assigned to the treatment group whereas the comparison group receives no such encouragement. Consider the example of a training program for unemployed. Evaluating this program through an encouragement design would amount to choose at random a treatment group to whom a letter explaining the contents as well as the dates and the location of the training is sent, whereas no
such letter is sent to the comparison group. The comparison group will not be excluded from treatment: the same information than those contained in the letter should be made available to them for instance on the website of their placement agency, and those who will manage to find this information out will also be given the right to follow the training. But it is likely that because of the letter, more treatment group than comparison group unemployed will have heard of the training. Therefore, we will end up with two groups initially comparable in every respect thanks to the lottery, but with a group among which a large share of individuals followed the training against a smaller share in the comparison group. If we are to observe that more people in the treatment than in the comparison group manage to recover a job, we will be able to attribute this to the fact that a larger share of the treatment group followed the training. The main disadvantage of the encouragement design is that it strongly reduces statistical precision. Assume that 35% of the treatment group followed the training against 25% in the comparison group. The effect of the training program should be extremely strong to find different job recovery rates across two groups whose only difference is that 10% more of individuals followed the training in one group than in the other. Therefore, if one is not willing to exclude anyone from treatment to increase the acceptability of the experimental protocol, we strongly recommend using a phasing-in design instead of an encouragement design.

The Hawthorne effect: a threat to validity specific to randomized experiments

A first threat to the validity of the results of a randomized experiment is the so-called Hawthorne effect. The experimental protocol might have an effect per se on subjects’ behavior, and one might end up calling an effect of the treatment what is merely an effect of the protocol. For instance, in a randomized experiment subjects know they are being observed by researchers because they are part of an experiment, which might induce them to adopt a behavior different from the behavior they would have had otherwise. This should not be too much of a concern: both treated and comparison subjects know they are being observed so that this first Hawthorne effect should cancel out when comparing treatment and comparison subjects.

But usually subjects also know that there is a comparison (resp. treatment) group and that they are part of the treatment (resp. comparison) group. This might also influence their behavior and bias the results of a randomized experiment. To understand why, consider the example of the intensive counseling program. If young graduates in the treatment group are being told they have been selected to receive intensive counseling after a lottery, they might feel a responsibility to put a great amount of effort in searching a job because they have been lucky enough to benefit from some supplementary help. On the contrary, those in the comparison group might feel depressed because of their bad luck which could undermine their efforts to find a job. Overall, the comparison of the placement rate in the treatment and in the control group would capture both the effect of intensive counseling and the effect of the experimental protocol.

To minimize the risk that such Hawthorne effects pollute the results from a randomized experiment, researchers usually tell subjects as little as possible about the
experimental design. It is unavoidable to tell subjects that some data is collected on them and to get their agreement on this. **But it is not always necessary to explain them that there is a treatment and a comparison group and to tell them in which group they are.** In the intensive counseling example, this would for instance require not to sort out graduates who will receive intensive counseling within each placement agency, but to sort out 50 placement agencies out of 100 which will offer intensive counseling to all young graduates they follow while the remaining 50 agencies will offer it to none. This will ensure that subjects in the treatment group do not realize they are being treated differently from other subjects in the comparison group since the two groups go to different placement agencies.

**Attrition: a threat to validity not specific to randomized experiments**

There is a last major threat to the validity of the results of a randomized experiment which is not specific to randomized experiments. Let us illustrate this with an example. “Boarding schools of excellence” is the name of a French educational program implemented in 2009. It is intended at middle and high school students from underprivileged backgrounds with strong academic skills. They are given the opportunity to study in a boarding school where they receive intensive coaching. This program is evaluated through a randomized experiment. Since there are more candidates than seats for this program, a lottery is conducted each year among potential candidates to determine who is admitted to the boarding school. Students admitted make up the treatment group; other students make up the comparison group. In the end of each academic year, all students participating in the experiment take standardized tests in Mathematics and French as well as some psychometric tests. The evaluation merely amounts to comparing results of the two groups of students in those tests.

Having students taking those tests is a very easy task in the treatment group: up to a small minority which has been expelled from the school during the year, they are still in the boarding school in the end of the year. This proves to be much harder for students in the comparison group. Since all of them were not assigned together to a single school contrary to what happened to their treatment group counterparts, a major search effort has to be conducted in order to find where they are at the end of the year.

Assume that such massive effort was not conducted. We would probably end up with let us say 95% of the students in the treatment group taking the tests against only 80% in the control group. The evaluation would then amount to comparing test scores among those 95% of students who took the tests in the treatment group to those 80% who took it in the comparison group. **Despite the fact that randomization ensured that by construction, the treated group and the comparison group were initially comparable in every respect, those two subgroups might no longer be comparable.** Indeed, it is very likely that those 20% of students who did not take the test in the comparison group were very specific students, for instance those who had dropped out from school or who had been expelled, which is the reason why it was very hard to locate them and to have them taking the test. In the treatment group, fewer students dropped out: the great opportunity which was given to them by sending them to a boarding school incentivized even those who would have been prone to drop out in a normal school to stay in school. Therefore, the comparison of those
95% of students taking the tests in the treatment group against those 80% taking them in the comparison group would amount to comparing apples and oranges: the treated group would comprise more weak students prone to drop out from school. Therefore, such a comparison would probably underestimate the true impact of boarding schools of excellence.

Let us formalize this. Randomization ensures that by construction, the treated group and the comparison group are initially comparable in every respect. But because some individuals participating in the experiment might move to another city during the experimental protocol, or might no longer be willing to answer surveys for instance because they were disappointed not to be selected for treatment, there will always be a share of the initial population for which the outcome we were interested in (test scores in the boarding school example) could not be measured. This share is what is called the attrition rate. Consequently, we will not end up comparing the entire treatment group to the entire comparison group, but individuals for whom we have been able to measure the outcome of interest in the treatment group to individuals for whom we have been able to measure the outcome of interest in the control group. It might very well be the case that those two groups are no longer comparable, for instance if treatment induced more (resp. less) individuals to attrite from the experiment in the treatment group. Therefore, when reading results from a randomized experiment, one should pay a great attention that the overall attrition rate in the experiment is not too high (as a rule of thumb let us say not above 20%), and above all that the attrition rate is not statistically different in the treatment and in the comparison group. If for instance the attrition rate is equal to 10% in the treatment group, and to 20% in the comparison group, one should consider the results of the randomized experiment under consideration with some caution.

To be fair to randomized experiments, one should finally mention that attrition is a serious threat to the validity of all evaluation methods. Consider for instance the Before-After method. Assume that in the intensive counseling example, placement rates of the 2007 and 2008 cohorts were measured through a survey. It is very likely to find that the response rate to this survey was higher in 2008 than in 2007. 2008 young graduates received more intensive counseling, and probably felt grateful to their counselors who dedicated a lot of time to them, which incentivized more of them to spend some time answering the survey. Consequently we will end up comparing two populations which differ on two dimensions. Not only the 2008 cohort experienced different labor market conditions than the 2007 cohort, but it also comprises both individuals who would have answered the survey anyway and individuals who answered because they felt grateful to their dedicated counselor, whereas the 2007 cohort was only made up of the first type of individuals.

The only reason why we mention attrition as a threat to validity when presenting randomized experiments is that attrition is by and large the only serious threat to the validity of the results from a randomized experiment, whereas all other methods suffer from other threats to validity.
Comparing results obtained through different methods

When attrition rates are sufficiently low, or at least balanced between the comparison and the treatment group, and that there is no reason to suspect a strong Hawthorne effect, one can consider that randomized experiments deliver a measure of the true impact of a program (up to statistical imprecision). There has been an important literature comparing measures of the impact of the same policy intervention obtained through each of the 5 non-experimental methods to the true impact of this policy intervention measured through a randomized experiment. This is exactly the exercise we conducted through the intensive counseling example. The main lessons of this literature are as follows.

Comparing participants to non participants will almost always yield a measure of the impact of the policy very far from the true impact. Matching will improve things. In some instances, it might even deliver a measure close from the true impact of the policy. But sometimes it will miss it by much, and it seems very difficult to find criteria which allow predicting beforehand in which circumstances matching will do well and in which circumstances it will do poorly.

Before-after usually falls far from the true impact, except when the outcome under consideration is extremely stable over time, for instance when it is not very sensitive to variations in the economic conditions.

Difference in differences will improve a lot on a simple before-after. Sometimes it will deliver a measure close from the true impact, sometimes it will fall far. When data is available over a long period of time, so that the parallel trends test displayed on figure 5 can be conducted over many periods, and that the control group by and large passes this test, one can be reasonably confident that difference in differences should yield a measure of the impact of the policy close from its true impact.

Finally, when the eligibility threshold cannot be manipulated, regression discontinuity most often delivers a measure of the impact of the policy close to its true impact.
Part 3: Measuring the efficacy of a policy intervention: some practical considerations

After considering the principles and the reliability of those six methods, this part first considers some practical issues related to data collection and which are common to all six methods. Then it reviews the respective advantages and inconvenient of each method in terms of cost, complexity to implement and acceptability. Methods are ranked from the more to the least complex in terms of implementation. It should not come as a surprise that the two most reliable methods are also somewhat more complex to put in place, whereas less reliable methods are easier to implement.

Measuring the outcomes

All methods require measuring the outcome(s) in the treated group and in the comparison group. In some instances, this measure is directly available from administrative data sets and requires no further data collection. In other instances, it is not available so that a specific survey must be conducted in order to collect it. Whether measuring the outcomes requires specific surveys or not essentially depends on the richness of the set of outcomes considered in the experimentation. If the experimentation considers only one very simple outcome, it is very likely that it will be available in some administrative data set. If the experimentation considers several complex outcomes, it will probably require specific data collection.

If the experimentation requires designing a survey, then it is highly recommended to use questions from existing surveys which have already been administered to large populations and not to design one’s own questions. This will ensure that questions are properly formulated and easy to understand. Such questionnaires can be found on the websites of international organizations such as the OECD (http://www.oecd.org/) or of national offices for statistics such as the INSEE in France (http://www.insee.fr/).

Among complex outcomes, a category of particular interest is psychological traits. It might be especially interesting to measure the impact of the program on a number of participants’ psychological traits. But people usually think that self-esteem, confidence and motivation are not things which can actually be measured through a questionnaire. However, psychologists have designed a great number of psychometric scales intended at measuring such psychological traits. Examples can be found here: http://www.er.uqam.ca/nobel/r26710/LRCS/echelles_en.htm. Those scales go through a well-defined “validation” process. It amounts to have a large number of subjects answering the questionnaire to ensure that the questions are easily understandable and that subjects’ answers to the different questions are consistent. In some instances, it is also required that psychologists show that subjects obtaining high scores to their scale indeed have a great propensity to adopt behaviors consistent with the psychological trait measured. For instance, psychologists designing a motivation for higher education scale should demonstrate that high school students scoring high on their scale are indeed more prone to undertake undergraduate studies. Therefore, one should not refrain from measuring the impact of a policy on various psychological dimensions since tools are available in order to do so. On the other hand, unless there is really no validated psychometric scale available to measure the
psychological trait one is interested in, one should really avoid designing its own questionnaire (“Do you have a good self-esteem?” or “Do you feel motivated?”).

In the counseling program example, only one outcome was selected: graduates’ placement rate after six months. This outcome was very simple and it appeared that placement agencies collected it themselves, so that it was available in one of their administrative database and required no further data collection. In the boarding schools of excellence example, policy makers and the evaluation team agreed that the treatment could have an impact on students’ academic performance, and on their psychological well-being. Therefore, several outcomes were selected: students’ score to standardized tests in French and Mathematics, as well as measures of their self-esteem, their motivation to go to school, the quality of their relationships with their professors… None of those outcomes was already available in some administrative database so that a specific survey was organized: students participating in the experiment took tests in French and Mathematics and answered psychometric questionnaires during two sessions of 1h30 each. Questions to investigate the quality of the relations between students and their professors were taken from the PISA survey conducted by the OECD. Students’ self-esteem and motivation to go to school were measured through two validated psychometric scales.

Despite the fact that all methods require measuring the outcome of interest in the treatment and in the comparison group, each method has its own peculiarities in terms of costs and complexity. For instance, the measurement of outcomes will be more costly with methods using larger comparison groups, because they require gathering data on a larger population. Moreover, randomized experiments do not merely amount to measuring outcomes among “naturally emerging” treatment and control groups. They also require constructing such groups. This will imply a somewhat complex experimental protocol, even though such a complexity is too often exaggerated. A last example is regression discontinuity which is not suitable to evaluating all programs. It will work only for programs whose eligibility rules include a continuous numerical criterion such as age. I now review all methods emphasizing their respective advantages and inconvenient in terms of cost, complexity to implement and acceptability.

Randomized experiments

Implementing an experimental protocol

Randomized experiments rely on an experimental protocol which requires some cooperation between policy makers and the evaluation team. This will usually imply some supplementary work from civil servants working in administrations where the experimentation is conducted. Such operational burden should not be denied: it should be acknowledged beforehand by policy makers, by the evaluation team and by civil servants. But it should not be exaggerated neither: in most cases, it will be possible to contain the extra workload at very reasonable levels.

The impact of the counseling program has been measured through a randomized experiment. The experimentation lasted for a year. In the end of each month, the 100 local
agencies had to send to the evaluation team a list of all young graduates who had reached their sixth month in unemployment during that month. In each of those 100 lists, the evaluation team randomly sorted out young graduates who had to be offered to participate in the counseling scheme and those who had to be excluded from it. Then the team sent back the list to each of the placement agencies, who offered lottery winners to participate in the counseling scheme. The experiment undoubtedly implied some supplementary work from placement agencies, but one should also recognize that the amount of extra work was reasonably low: the experimentation only required them to send each month listings to the evaluation team, and then to pay attention to offer lottery winners only to participate in the counseling scheme.

Ethical issues

Randomized experiments also raise ethical issues. Most evaluation teams conducting randomized experiments try to abide by the following ethical rule: conducting a randomized experiment should not diminish the total number of recipients of the program. This means that in the intensive counseling example, money was available to provide intensive counseling to 5,000 young graduates and to 5,000 only. This ethical rule would have been violated if sufficient money had been available to provide intensive counseling to all 10,000 young graduates eligible but that the evaluation team had requested it were provided to 5,000 of them only in order to have a comparison group. Therefore, randomized experiments are usually conducted when there are more candidates than seats for a program.

In such instances, some rule must be used to allocate the program, whether it is a lottery or some other rule. Opponents to randomized experiments argue that the lottery is not a fair allocation rule because social programs should be attributed to those who need it most. Because of the lottery, some young graduates very much in need of intensive counseling might be deprived from it. Supporters usually answer that the whole point of conducting an experiment is to determine whether the program is useful or detrimental, so that beforehand it makes little sense to claim that lottery losers will be disadvantaged with respect to winners: if the program proves detrimental (counseling might not help young graduates to find a job and might just be a waste of time), lottery winners will end up being disadvantaged with respect to the comparison group because they did not get the chance to escape from the program!

In medicine, application for marketing authorization of each new drug must contain a randomized experiment demonstrating its efficacy. A lottery is conducted among a pool of sick patients to determine those who will receive the new medicine and those who will receive a placebo. Recovery rates in the two groups are compared after some time to determine the efficacy of the new drug. Depriving sick patients from a potentially helpful treatment raises far more serious ethical issues than depriving young unemployed from a potentially helpful counseling scheme. But doctors still consider that the benefits of such experimental protocols (measuring precisely the efficacy of new drugs) far outweigh their ethical costs.

Moreover, even when there is a strong suspicion that the program is indeed helpful, it is extremely hard to tell beforehand who will benefit the most from it. For instance, it is
fairly unclear whether males will benefit more from some intensive counseling than females. Therefore, an allocation rule targeting those presumably needing the program most could end up allocating the program to those who least need it if prior beliefs on who needs the program most prove wrong. On the contrary, as mentioned above, randomized experiments enable identifying subpopulations which benefit most from the program, so as to optimize eligibility criteria when scaling up.

Finally, let us mention that randomized experiments are usually submitted to a human subjects ethical committee for approval, and that alternative designs such as the phasing-in design presented above strongly reduce ethical issues.

**Regression discontinuity**

Regression discontinuity does not require an experimental protocol. Consequently, it introduces virtually no supplementary complexity and therefore no supplementary costs with respect to running the program the way it would be run if scaled up. However, regression discontinuity requires that the program under consideration meets two restrictive criteria which is the reason why it is not used so often in practice.

Firstly, **eligibility to the program should be based upon a numerical criterion taking a continuous set of values such as age or income**. Consider the following fictitious example of a program intended at increasing the share of young mothers getting back to work after their pregnancy. Suppose this program amounts to giving a financial aid to mothers who give birth to a child if they resume to working in less than 6 months after they gave birth. Assume also that only mothers giving birth to their third child at least are eligible. In this context it is not possible to use a regression discontinuity type of analysis to evaluate the impact of the program, because the eligibility rule is based on a discrete numerical criterion: number of children can take only integer (1, 2, 3…) values, not decimal ones (3.45, 5.72…). Therefore, regression discontinuity in this context would amount to compare activity rates of women just below to women just above the eligibility threshold, that is to say to compare women who gave birth to their second child to women who gave birth to their third child. Women giving birth to their second child are probably too different from women giving birth to their third child (they are probably younger for instance) to serve as a credible comparison group and for their activity rate to be representative of the activity rate that women giving birth to their third child would have experienced if they had not benefited from the program.

The second limitation of regression discontinuity is that it **often yields statistically imprecise estimates of the impact of the program**. Indeed, there are usually few individuals slightly below and slightly above the threshold so that the comparison is based on a small number of individuals. To illustrate this, let us consider again the counseling example. The 100 local agencies followed 10 000 young unemployed eligible to the program in 2008. Assume that all of them were aged between 25 years old and 0 month and 27 years and 11 months. This means that they belong to 36 different year & month cohorts of age. If this population is evenly distributed according to age, we can expect to have approximately $10 000 / 36 = 300$ of them who were 27 years and 11 months. Despite the fact that the initial population is large, regression discontinuity will finally amount to comparing two groups
bearing only 300 individuals each, whereas in a randomized experiment two groups of 5 000 individuals each would have been compared. This will entail a lack of statistical precision.

This will also make it very difficult to compare the impact of the program across different subgroups so as to optimize eligibility criteria before scaling up as is usually done when randomized experiments are used. Indeed, measuring the impact of the intensive counseling program on males only will amount to comparing the placement rate among 300 / 2 = 150 males just below the threshold to the placement rate among 300 / 2 = 150 males just above it. Similarly, the measure of the impact of the program on females only will amount to comparing two groups bearing only 150 individuals each. Therefore, the impact of the program on both males and females will be estimated very imprecisely making it very difficult to determine if one is larger than the other.

Therefore, when a social program is experimented using a regression discontinuity type of analysis, two requirements should be met. Eligibility to the program should be based on some continuous criterion which cannot be easily manipulated by participants to the experiment such as age. Moreover, a statistician should have made statistical power computations beforehand to ensure that given the number of expected participants in the experiment, it will have sufficient power to measure the effect of the program with reasonable statistical precision.

**Difference in differences and before-after**

Those two methods consist in measuring the evolution of the outcome of interest before and after the program is implemented. Therefore, they require that data on the population of interest is available also before the program was implemented and not only after its implementation. In the intensive counseling example, this means that placement data for young graduates followed in 2007 by placement agencies participating in the experiment should be available. This will not be an issue in the intensive counseling example because placement agencies always collect placement data of people they follow. If it were not the case, a specific survey should have been conducted to measures outcomes, and then the experiment should have been planned much in advance, for instance in 2006, so as to interview also the 2007 cohort and to compute its placement rate.

With respect to a simple before-after, difference in differences only require to find a control group excluded from the program both before and after its implementation because it does not meet the eligibility criteria (e.g. young unemployed with no degree in the intensive counseling example), and to collect relevant outcome data for this control group also. Therefore, this will increase survey costs (if it is necessary to conduct a survey to measure outcomes) since information on a larger sample of individuals will have to be collected.

**Statistical matching and participants VS non-participants**

Those methods require measuring the outcome of interest for all individuals eligible to the program, i.e. both participants and non-participants, during the duration of the program
only. In the intensive counseling example, this means that placement data needs to be collected for all young graduates eligible to intensive counseling in 2008. Therefore, those two methods are the simplest in terms of data collection and experimental protocol.

With respect to a simple comparison of participants and non-participants, the only supplementary requirement of statistical matching is that data analysis should be conducted by a skilled statistician since matching procedures are somewhat complex to implement. This will induce an increase in the cost of data analysis.
Conclusion: from the experimentation to scaling-up

Designing a social experimentation first requires defining a relevant policy intervention. In order to do so, some basic steps must be followed such as rigorous description of the social need the policy seeks to address, a precise statement of all the actions which will be part of the policy, an examination of all the constraints faced by the targeted population to ensure that it will be able to enroll in the experiment, a thorough literature search of all experimentations of similar policy interventions conducted at home or abroad to gather some evidence that the policy under consideration will indeed address the social need it seeks to address… Even though few pages are dedicated to describing those steps in this guide, we insist that they are crucial to the success of the experimentation. **There is no point to evaluate a policy intervention which has been shown to be useless by dozens of previous studies, or to experiment a program to which beneficiaries are not able or willing to participate in.** A large number of experimentations fail because targeted participants are actually not willing or not able to participate. Among programs which prove ineffective in the end of a long and costly experimentation, many could have been regarded as presumably ineffective beforehand after conducting a thorough literature search.

Once a presumably relevant policy intervention has been designed, one must choose how to evaluate its impact on beneficiaries. Most of this guide is dedicated to presenting and comparing the 6 most commonly used methods in the impact evaluation literature. Part 2 compares those methods from the point of view of the validity and of the reliability of the results they deliver while part 3 compares them from the point of view of their respective complexity to implement in practice.

Choosing the right method is merely a **trade-off** between the **cost of the experimentation** in terms of time and money and the **cost of scaling-up an ineffective program** (resp. renouncing to an effective policy) because the experimentation erroneously concluded the policy to be effective (resp. ineffective). To make an optimal trade-off, one should bear in mind four elements: the cost of the experiment, how this cost varies depending on the evaluation method used, the cost of the program when scaled up, and an assessment of the degree of uncertainty on the program effectiveness.

**A cheap, not very innovative program,** for which there are already a large number of studies indicating that similar type of programs are usually cost-effective, **might not need being evaluated through a randomized experiment.** However, since the supplementary cost of evaluating it through a difference in differences than through a before-after or a comparison of participants and non-participants will probably be very low, in most instances it will be worth it bearing those supplementary costs. **But an expensive and very innovative program should be evaluated through a randomized experiment or through a regression discontinuity.**

In the end of the experimentation, results finally come. They enable computing the cost-effectiveness of the program and determining the subgroups on which it has been the most effective. Based on this, policy makers can decide whether or not to scale up the program, and which eligibility criteria should be used. Doing so, they must bear in mind that
results of the experiment hold only within the population of the experiment. The experimentation of the intensive counseling scheme showed this program to be effective and relatively cost efficient. But this might not be true outside the population of the experiment. Despite the fact this program proved efficient among young graduates having looked for a job for more than 6 months, there is no reason why it should also work among young unemployed with no degree or among senior unemployed. Consequently, scaling-up does not mean extending the program to all French unemployed, but only to all French young graduates having looked for a job for more than 6 months. Determining whether it should be extended to young unemployed with no degree would require another experiment.