European Commission

SOCIAL EXPERIMENTATION

A methodological guide for policy makers

Written by J-Pal Europe

At the request of

Directorate General for Employment, Social Affairs and Inclusion

Version for the Ministerial conference "Innovative responses to the social impact of the crisis" *organised by the Polish Presidency of the European Union – Wrocław, 26 September 2011*

http://ec.europa.eu/social/innovationconference - *empl-innovation@ec.europa.eu*

The principle of social experimentation is to test 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 policy-relevant intervention and measuring its actual efficacy.

This guide is intended for policymakers interested in embarking on social experimentation. It is divided into three sections.

(1) The first section briefly lays out some basic principles to follow in order to **design a potentially policy-relevant intervention**, and illustrates these principles with an example.
(2) In the second section, the principles for six commonly used methods of evaluation are presented. The methods are compared from the point of view of the **reliability of the results they deliver**.
(3) The third section considers the **costs** associated with each method, and their **complexity** to implement in practice.

There are two main actors involved in social experimentation: policymakers and evaluation teams (usually made up of consultants or researchers). The role of policymakers is to design the policy intervention and to support the implementation of the experimental protocol. The evaluation team may be asked to contribute to the design of the policy intervention, but its main role is to design the experimental protocol, to implement the experimentation, and to collect and analyze the data necessary to measure the efficacy of the program.

This guide will give policymakers an overview of the necessary steps for conducting a rigorous experimentation measuring the impact of a policy intervention. To this end, the guide aims to improve their understanding of the work conducted by the evaluation team and to facilitate mutual understanding.

1. Designing a Potentially Policy- Relevant Intervention

A. Some principles for policy intervention design

In the beginning of a social experimentation, it is important to come up with a rigorous description of the **social need** the program seeks to address and to document the nature of this need. Then, one should precisely describe the set of actions envisaged as part of the policy intervention, and explain **why they might help address this social need**. In particular, one should ensure that the different incentives, opportunities or constraints with which the population will be confronted are identified and described. The program should be 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 for examples of similar policy interventions that have been conducted domestically or abroad. **This search can also provide supplementary evidence that the program is likely to address the social need that was identified.** 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 can help in conducting these searches. One might also be interested in scanning the archives of consulting firms and research centers specializing 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/) and 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, so that they may be measured appropriately by the experimentation.

Finally, it is important to **involve all relevant stakeholders from the moment the project discussion starts**. This helps to build consensus on the design of 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 developed beforehand on a set of conditions for scaling this program up if results are positive.

B. Example: Employment Counseling for Young Graduate Job-Seekers

We will now illustrate these principles with an example which will be used throughout this guide. It is taken 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 only 70% of university graduates had a stable job three years after they completed their degree. **The social need addressed here is therefore to improve the insertion of young graduates into the labor market.** The existing low placement rates could be due to the fact that internships are not mandatory in

French universities, or the fact that workshops that facilitate interaction between students and companies are seldom organized. Most graduates leave university without ever having searched for a job, with little knowledge of the jobs they should apply for, with no experience preparing for job interviews and sometimes even without having written a resume.

Consequently, the French Ministry of Labor decided to test a counseling program to help young graduates who had been unemployed for at least 6 months. **The program increased the frequency of the meetings between young graduates and their counselor from once per month to once per week over a period of 6 months.** This was supposed to allow counselors to spend more time helping young graduates think about their professional plans, writing a proper resume, organizing their job searches, and preparing for job interviews they would have during the following week.

The **constraints of the targeted population** were carefully taken into consideration. It was not certain that young graduates would have time or be willing to meet with their counselor every week. To answer this question, the placement agencies participating in the experimentation conducted a **survey to ask young graduates whether they would be willing to meet once per week with their counselor, and if they thought this would help their job prospects.** It appeared that young graduates indeed thought the intervention would be both feasible and useful.

The likely effectiveness of this policy intervention was confirmed by a thorough literature search which found **6 previous studies in which counseling increased the placement rate of job-seekers,** against zero 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 open at the beginning of the experimentation.

Finally, it was decided that the main outcome measure to assess the effectiveness of this program would be the share of young graduates who had found a durable job after 6 months of intensive counseling. A young graduate was considered to have 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 among 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 to offer suggestions on how to improve them.

Before turning to the choice of a method to evaluate the policy intervention, it is worth reiterating that the steps described above are crucial to the success of a social experimentation. Social experimentations are long and costly; they require a substantial amount of supplementary work from policymakers and civil servants working in the departments where the experimentation is implemented. It is worthwhile to design policy-

relevant interventions carefully. **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 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 refuses to participate in the program, or because far fewer people enroll than were expected. Among programs which prove ineffective at the end of a long and costly experimentation, many could have been confirmed as ineffective beforehand, merely after conducting a thorough literature search. This would have saved the costs of a useless experimentation.

2. Measuring the Efficacy of a Policy Intervention: Methodological Considerations

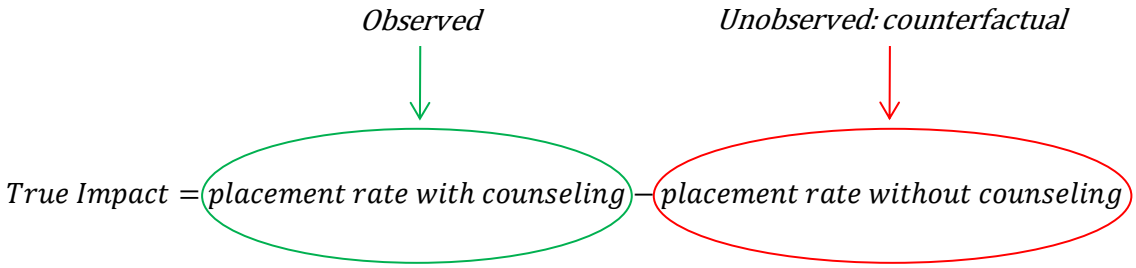
A. The conceptual problem of impact evaluation

The efficacy of a policy intervention is its capacity to address the social need it targets. To know whether an intervention was effective, one must measure its impact on program recipients: how does the policy intervention change the lives of those who benefit from it? In this example, does the intensive counseling program allow some young graduates to find jobs they would not have found otherwise?

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

$$True\ Impact = placement\ rate\ with\ counseling - placement\ rate\ without\ counseling$$

The first term in this difference is very easy to compute: it is merely the placement rate of recipients of the counseling program. But it is not possible to compute the second term because by definition it is not directly observable. Indeed, we do not know what the placement rate of the recipients of the counseling program would have been if they had not received counseling. This second term is called the **counterfactual** scenario.



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 this, we must find a **comparison group** which did not receive the counseling, so we are able to compare the placement rate in the group which received the counseling (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 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 between those two groups.**

Random assignment to treatment is regarded as the **"gold standard"** for constructing a valid comparison group. But as we will emphasize in the next part of this guide, randomized

evaluations of social programs take time and can be complex to implement. For these reasons, other techniques are also commonly used. They are referred to as non-experimental or quasi-experimental methods. They are usually **less complex** to implement than randomized evaluations, but the results they deliver are also **less reliable**. The reason for this is that random assignment to the treatment and comparison groups ensures that the comparison group is indeed similar in every respect to the treatment group. On the contrary, non-experimental methods must rely on **an assumption** to justify the claim that the comparison group they use is indeed similar to the treatment group. Results from non-experimental methods are more credible when this assumption is credible in the context under consideration.

In the following six sections, the most commonly used methods in the impact evaluation literature are presented. Emphasis is placed 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 unreliable assumptions but they are included nonetheless because they demonstrate the conceptual challenge faced when conducting an impact evaluation. The two following methods rely on much more credible, though still strong assumptions. Finally the last two methods rely on either fairly safe assumptions or no assumption at all. The last section compares results obtained through different methods.

As we illustrated with the counseling program example, the results of an experimentation are highly dependent on the method used: each method will deliver its own measure of the impact of the treatment. As evidenced by debates about public policy, when two different methods are used to measure the impact of a program, one method might conclude that the program had a positive impact while the other method finds a negligible or even negative impact. **Hence it is extremely important to bear in mind the assumptions upon which each of those methods rely, so as to be able to assess which assumptions are the most credible in the particular situation under consideration.**

C. Comparing participants and non-participants

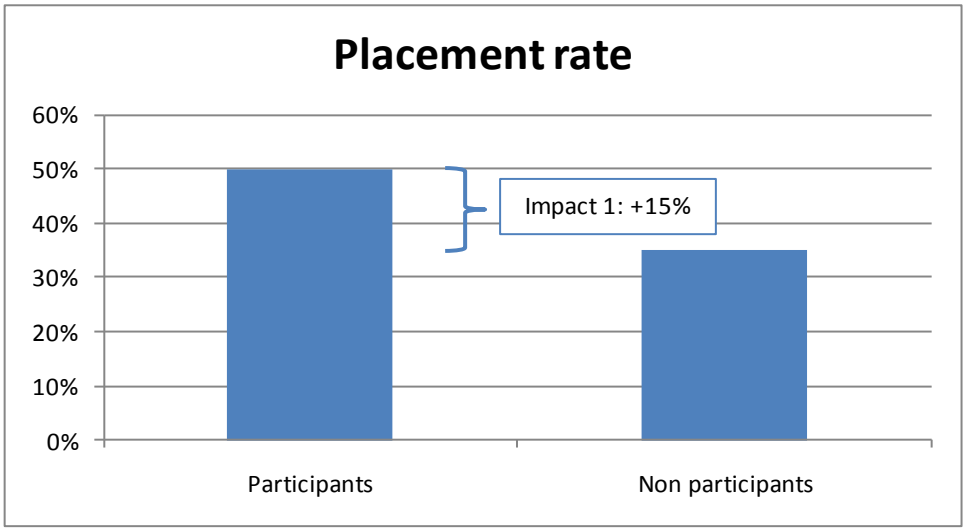
A simple way to measure the impact of a program is to use individuals eligible for the program but who chose not to participate as a comparison group.

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 50%, against 35% among non-participants, this methodology would conclude that the program increases the placement rate by 15 percentage points.

$$Impact\ 1 = participants\ \ placement\ rate - non\ participants\ \ placement\ rate$$

$$= 50\% - 35\% = +15\%$$

Figure 1 : Participants vs. non-participants



However, for this measure to be representative of the true impact of the program, the treatment group and the comparison group should be identical in every respect. In the context of the counseling program, this means that graduates who chose to participate in the program should be similar to those who chose not to participate in terms of gender, qualifications, motivation to find a job, etc. 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 such as age, gender, etc. If, for instance, men were more reluctant to participate in the counseling program than women, then there would be a larger share of males in the comparison group than in the treatment group. Since in France women face higher unemployment rates and therefore lower placement rates than men, the comparison of placement rates among participants and non- participants will capture both the effect of the training program, and the fact that the treatment group comprised more women who were less likely to find a job anyway.

More importantly, the treatment 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 that 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, and the fact that they also differed on unobservable characteristics such as their motivation to find a job.

D. Before-after

Before-after comparisons use **the exact same population which enrolled in the program, before the program was implemented,** as the comparison group. The French counseling program was offered to young graduates who had spent at least 6 months unemployed 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. This population did not benefit from intensive counseling, but might be similar to the population which received counseling.

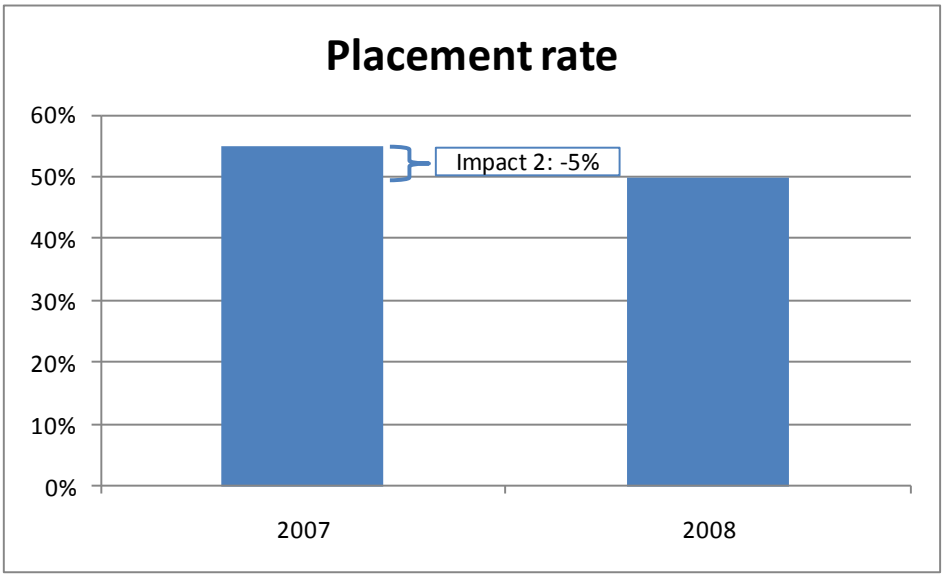
According to the before-after methodology, the impact of the counseling program is merely equal to the change of the placement rate of young graduates after the counseling program was implemented.

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 as in 2007 if the counseling program had not been implemented. One can imagine many scenarios which would violate this assumption. For example, **an economic recession occurred in 2008**. In this case, one might find that the placement rate of young graduates in 2007was 55%, against only 50% 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 conclude that it diminishes the percentage of young graduates finding a job in less than 6 months by 5 percentage points, whereas this decrease is probably at least partly due to the economic recession and not to the implementation of the counseling program.

$$Impact\ 2 = placement\ rate\ after - placement\ rate\ before$$

$$= 50\% - 55\% = -5\%$$

Figure 2 : Before-After



E. Statistical matching

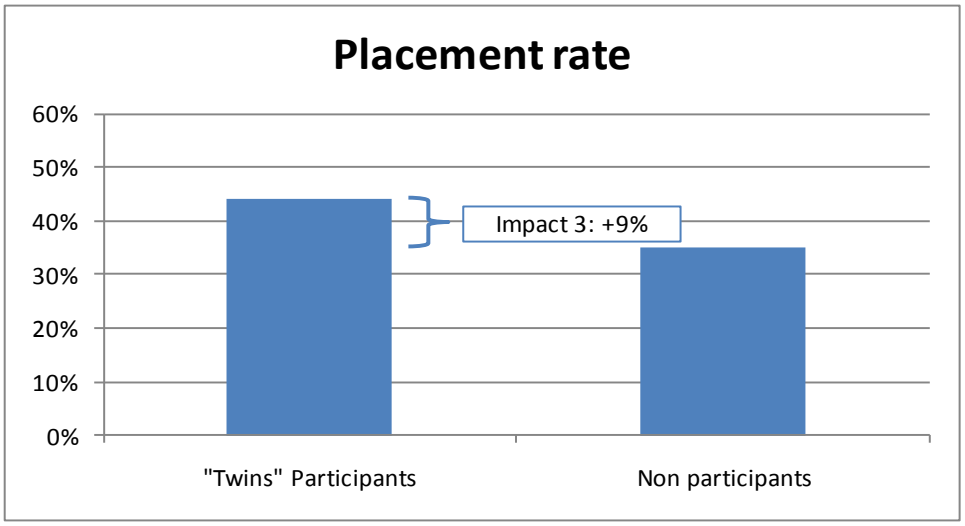
Statistical matching builds upon the same intuition as the comparison of beneficiaries and non-beneficiaries. Instead of comparing all beneficiaries and all non-beneficiaries of the program, pairs of beneficiaries and non-beneficiaries resembling each other are constructed and the comparison is conducted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

In the context of the counseling program, one should **find a non-participant for each 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, find someone of identical age, gender, previous work experience and qualifications. You would end 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 the placement rate among "twin" participants was equal to 44% (instead of 50% among all participants), against 35% among matched non-participants, then the impact of the program would be estimated as a 9 percentage point increase in placement.

$$Impact\ 3 = "twins"\ participant\ placement\ rate\ -\ non\ participant\ placement\ rate$$

$$= 44\% - 35\% = +9\%$$

Figure 3 : Matching



Statistical matching is a significant improvement on the comparison of beneficiaries and non-beneficiaries. Indeed, it ensures that by construction the two groups which are compared are **very similar with respect to important observable characteristics** used in the matching procedure, such as 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 simple fact that participants and non-participants differ in their motivation to find a job.

F. 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 enrolled in the counseling program and a group not enrolled in it, for instance uneducated young job-seekers.

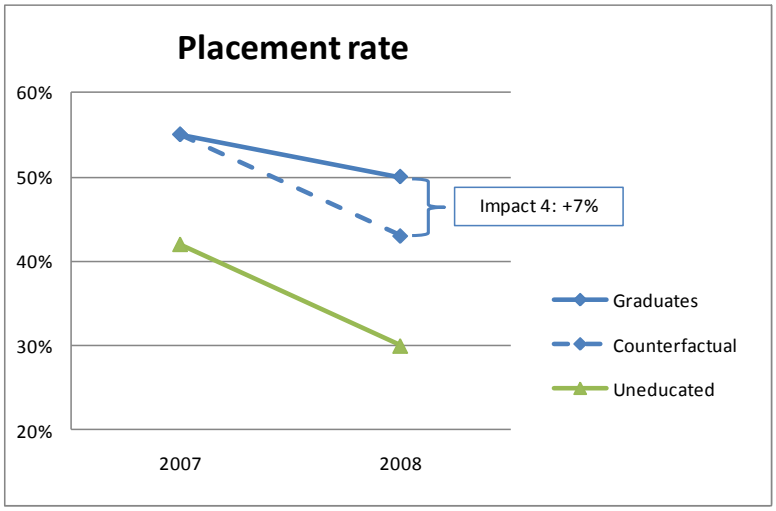
The simple change in the placement rate among the treatment group between 2007 and 2008 might not yield the true impact of the counseling program, for instance because economic conditions changed between 2007 and 2008. Therefore, to recover the true impact of the counseling program, one should compare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

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

$$Impact\ 4 = change\ of\ placement\ among\ eligible - change\ of\ placement\ among\ ineligible$$

$$= (50\% - 55\%) - (30\% - 42\%) = -5\% - (-12\%) = +7\%$$

Figure 4 : Difference in differences

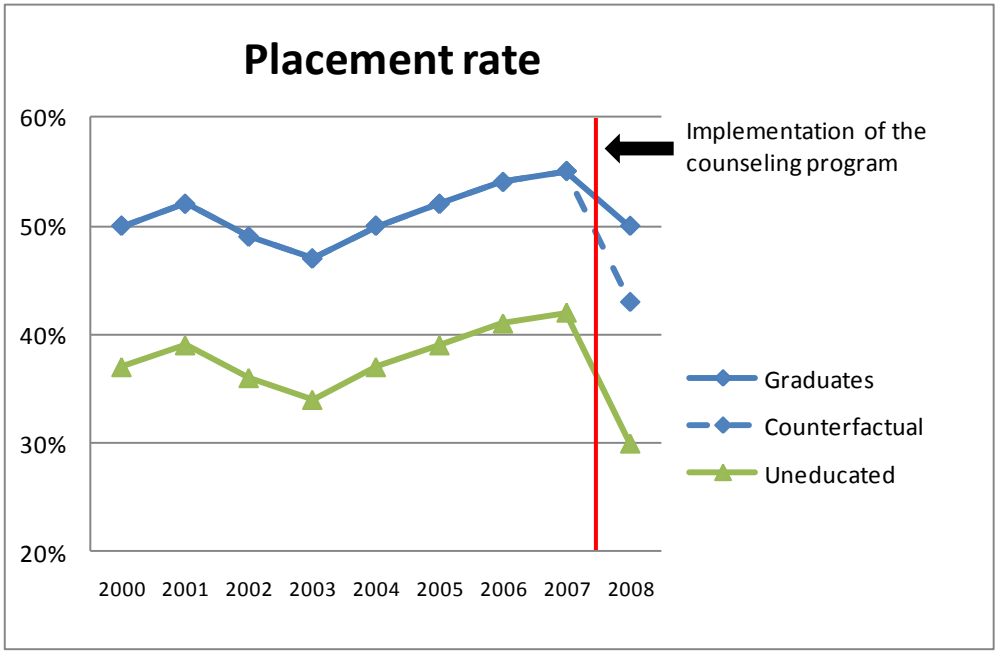


The fact that the placement rate decreased less between 2007 and 2008 among young graduate job-seekers than among uneducated young job-seekers suggests that the counseling program had a positive impact, since it allowed young graduates to suffer less from the economic recession than young job-seekers with no degree.

However, for this impact estimate 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 as the placement rate of uneducated young job-seekers. 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 paths. This is the reason why it is referred to as the "**parallel trends assumption**". This assumption could also be violated. One could, for instance, argue that the recession has probably hit the uneducated group more severely. Indeed, they might be **more vulnerable** to macroeconomic shocks because of their lack of qualifications.

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 prior to the program. 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.

Figure 5: Support for the parallel trend assumption

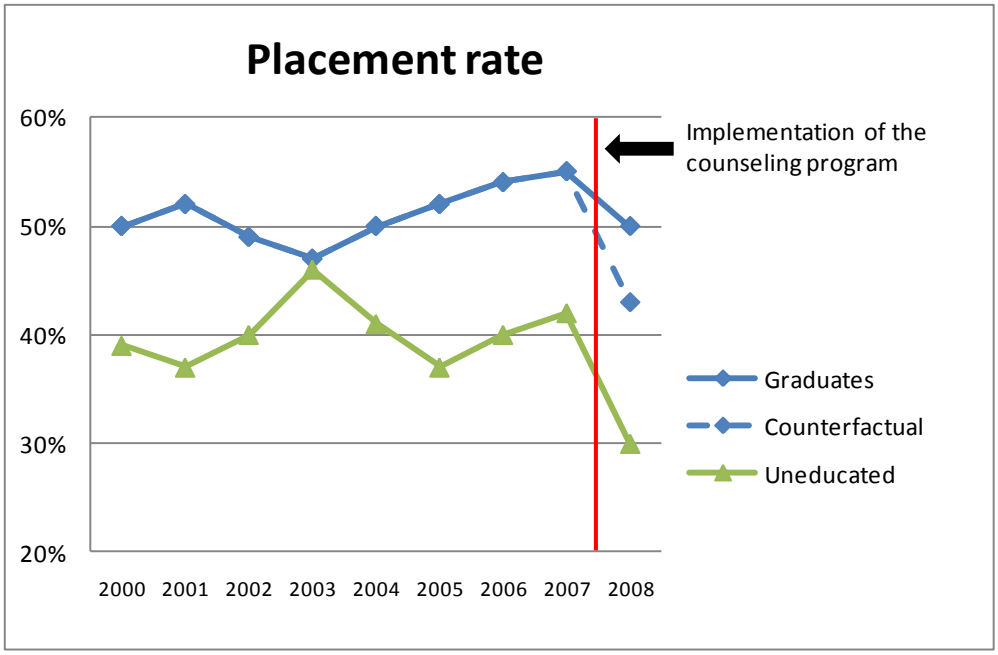


The parallel trend assumption states that placement rates in the two groups would have followed a parallel change from 2007 to 2008 if the counseling program had not been implemented. **The fact that placement rates in the two groups indeed followed a parallel evolution between 2000 and 2007 gives some credibility to this assumption.** Since

graduates and uneducated job-seekers 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 paths between 2000 and 2007 undermines the credibility of this assumption**. Since, in this scenario, educated and uneducated job-seekers 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.

Figure 6: A graph contradicting the parallel trend assumption



Such a test of the validity of the parallel trend assumption should be conducted beforehand, when designing the experimentation. Once a control group has been found, one should construct graphs similar to Figures 5 or 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.

G. Regression discontinuity

Eligibility for some programs is determined according to whether a participant is above or below a threshold for a measure like age or income. In such instances, one can measure the impact of the program through a technique called “regression discontinuity”. The principle is simply to compare program participants who are “very close” to being ineligible

because they are only slightly above the threshold to non-participants who are "very close" to being eligible because they are only slightly below the threshold.

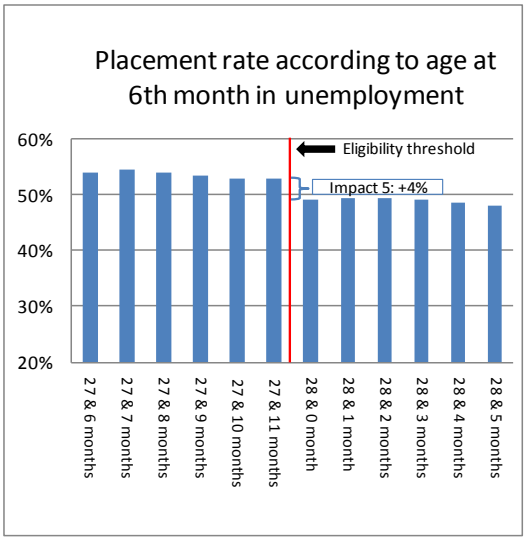
Let us illustrate this with the counseling program example. Assume that this program was accessible only to young graduates who were less than 28 years old when they reached their 6th month of unemployment. Then one could measure the impact of the counseling program by comparing placement rates of those who were 27 years and 11 months old when they reached their 6th month in unemployment to the placement rate of those who were exactly 28 years and 0 months old. The first group benefited from counseling, whereas the second 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 exactly 28 years and 0 months old, then, as per the regression discontinuity methodology, the program increased the placement rate by 4 percentage points.

$$Impact\ 5 = placement\ "slightly"\ above\ the\ threshold - placement\ "slightly"\ below$$

$$= 53\% - 49\% = +4\%$$

The underlying assumption of regression discontinuity is that the placement rate of those who reached their 6th month of unemployment at 27 years and 11 months is representative of the placement rate that would have seen among those who reached their 6th month in unemployment when they were 28 years and 0 months old if they had not enrolled in the counseling program. In this context, this seems like a fairly reasonable assumption: **there is no reason why those two groups should differ strongly**. The graph below, which plots placement rates according to age when reaching 6 month of unemployment, gives some support to this assumption. **Differences in placement rates across age cohorts are extremely small (1 percentage point at most), except between the two cohorts which reached the 6th month of unemployment at 27 years old and 11 months and 28 years old, that is to say precisely around the eligibility threshold.**

Figure 7: A graph supporting the regression discontinuity approach



However, this assumption can be violated if people are able to **manipulate the variable on which eligibility to the program is decided**. As young graduates cannot lie about their age so as to enter the counseling program, this is not a concern here. But this can be an issue in other contexts. Let us consider the example of a microcredit program in Mexico, for 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 temporarily sold part of their land to become eligible for the program**. In such circumstances, farmers owning slightly less than 2 acres of land are not comparable to those owning slightly more. Indeed, the population of farmers owning slightly less than 2 acres includes both those who actually own 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 other hand, the population of farmers owning slightly more than 2 acres only includes 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 who owned slightly less than 2 acres than farmers who owned slightly more than 2 acres of land. In theory there should be approximately the same number of farmers slightly below and slightly above that 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 to eligibility threshold**, i.e. on people who are close to 28 years old when they reach 6 months of unemployment. Policymakers might be interested in knowing the impact of this program not only among this subgroup, but among the entire population, in which case the regression discontinuity method will be useless. Therefore, when interpreting results from a regression discontinuity study, one should keep in mind that results apply only to people close to the eligibility threshold.

H. Randomized experimentations

Randomized experimentations deliver a measure of the true impact of the program

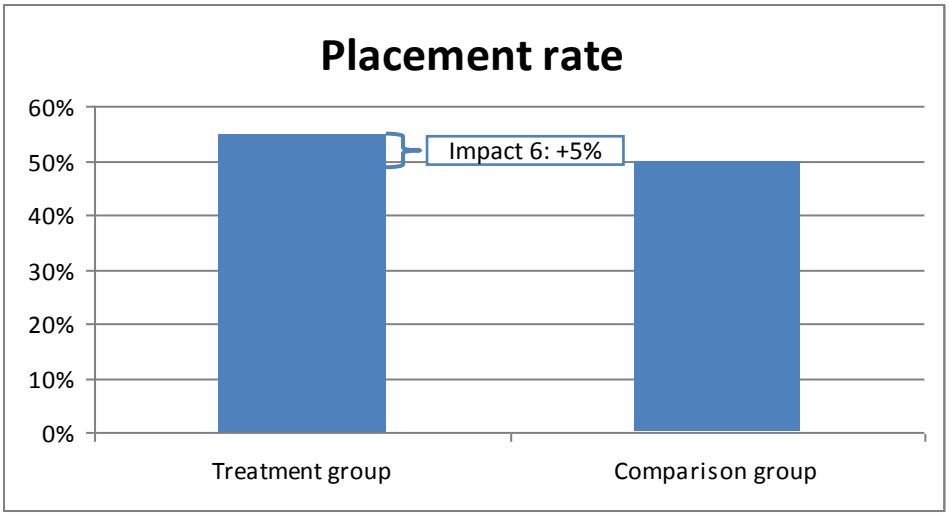
Randomized experimentations are experimentations of social policies in which assignment to treatment is based on the results of a random assignment, or lottery. Assume that the 100 French local agencies participating in the counseling experimentation followed 10 000 young graduates eligible for this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

The impact of the program will then be measured by comparing the placement rate among those two groups. Assume, for instance, that 55% of those randomly selected for counseling had found a job in less than 6 months, whereas only 50% had found jobs in the comparison group. Then, according to this randomized experimentation, the impact of intensive counseling is to increase the placement rate by 5 percentage points.

$$Impact\ 6 = lottery\ "winners"\ placement\ rate - lottery\ "losers"\ placement\ rate$$

$$= 55\% - 50\% = +5\%$$

Figure 8: Randomized experimentations



Randomization ensures that the treatment group and the comparison group are comparable in every respect (age, proportion of men/women, qualifications, motivation, experience, cognitive abilities, etc.). Indeed, **when a population is randomly allocated 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 for the counseling program, 5 000 were not motivated to find a job and 5 000 were extremely motivated. When randomly selecting the 5 000 job-seekers who will actually receive intensive counseling, it is possible that we could select 4 000 extremely motivated and 1 000 unmotivated persons. If this were the case, then the treatment group and the comparison group would not be comparable at all: the treatment group would comprise 4 000 extremely motivated job-seekers versus 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 a fair coin 5 000 times, that is to say approximately 0. Indeed, when tossing a fair coin 5 000 times, the probability of getting heads in each of those draw is equal to one half. Therefore, we expect to get heads in approximately half of the tosses, around 2 500 times. Getting 4 000 heads is so far from the scenario we expect to observe on average (2 500 heads), that the probability that it will occur is almost 0. Actually, one can compute that when tossing a fair coin 5 000 times, there is a 95% probability of getting heads between 2430 and 2570 times. Returning to the intensive counseling example, this means that when randomly selecting the 5 000 job-seekers who will actually receive intensive counseling, there is a 95% probability that the number of extremely motivated job-

seekers assigned to the intensive counseling group will be between 2430 and 2570, in which case the treatment group and the comparison group will be comparable.

Therefore, randomization ensures that the treatment group and the comparison group are comparable on every dimension (age, proportion of males, qualifications, motivation, experience, cognitive abilities, etc.). The only difference between those two groups is that one receives intensive counseling and other does not. Consequently, if we observe that the placement rate is higher in the treatment 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 treatment group if it had remained untreated. **Therefore, randomized experimentations allow us to measure the true impact of intensive counseling.**

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

i. Alternative designs

There are many possible designs for randomized experimentations. 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 a **phase-in** design. Assume that sufficient money is available to hire extra teachers for all schools in an area, but that half of that money will be available this year and that the remaining half will come the next year. One could randomly decide which schools will receive an extra teacher in the first year (treatment group) and which will not receive an extra teacher until the second year (comparison group). Comparing pupils' test scores in these two groups at the end of the first year would yield a measure of the impact of giving schools an extra teacher. The main advantage of the phase-in design is its great degree of **acceptability**: the experimental protocol will seem much more acceptable to the comparison group if they are told they will have access to the program only one year later. The main disadvantage of the phase-in design is that **it precludes measurements of the long-run impact of the program**: in year two, both treatment and comparison schools will have an extra teacher.

Another design which prevents us from having to exclude anyone from treatment is the **encouragement design:** an encouragement for treatment is randomly assigned to the treatment group, while the comparison group receives no such encouragement. Consider the

example of a training program for job-seekers. Evaluating this program through an encouragement design would entail random selection of a treatment group who would receive a letter explaining the existence of the intensive training program as well as the dates and the location of the training. No letter would be sent to the comparison group. The comparison group is not be excluded from treatment: the same information as that contained in the encouragement letter would be made available to them, for instance, on the website of their placement agency, and those who find this information will still be given the right to enroll in the training. But it is likely that **because of the letter, more treatment individuals than comparison group individuals will have heard of the training.** Therefore, we will end up with two groups initially comparable in every respect thanks to random assignment, but in the treatment group a larger proportion of individuals will enroll in training. If we observe that more people in the treatment than in the comparison group manage to find a job, we will be able to attribute this to the fact that a larger share of the treatment group enrolled in the training. **The main disadvantage of an encouragement design is that it strongly reduces statistical precision.** Assume that 35% of the treatment group enrolled in the training against 25% in the comparison group. The effect of the training program would have to be extremely strong in order to find different employment rates across two groups whose only difference is that 10% more individuals enrolled in training in one group than 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 phase-in design instead of an encouragement design.**

ii. The Hawthorne effect: a threat to validity specific to randomized experimentations

One threat to the validity of the results of a randomized experimentation is the so-called Hawthorne effect. **The experimental protocol itself might have an effect on subjects' behavior, and one could accidentally conclude that an effect of the protocol was actually an effect of the program.** For instance, in a randomized experimentation subjects know they are being observed by researchers because they are part of an experimentation, which might cause them to act differently than they would have otherwise. This should not be too much of a concern: both treatment and comparison subjects know they are being observed, so that the Hawthorne effect should cancel out when comparing treatment and comparison subjects.

Usually, however, subjects also know that there is a comparison group and that they are part of the treatment group, or vice versa. This might also influence their behavior and bias the results of a randomized experimentation. 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 greater effort into their job search because they were 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 complicate the results from a randomized experimentation, 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 for such data collection. **But it is not always necessary to explain that there is a treatment and a comparison group, or to tell them which group they are in.** In the intensive counseling example, this would entail not randomly selecting *individuals* for treatment or control, but randomly selecting 50 of the 100 *agencies* for each group. 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 are served by different placement agencies.

iii. Attrition: a threat not specific to randomized experimentations

There is one last major threat to the validity of the results of a randomized experimentation, although it is not specific to randomized experimentations. Let us illustrate this with an example. "Boarding Schools of Excellence" is the name of a French educational program implemented in 2009. It is aimed at middle and high school students from underprivileged backgrounds who have strong academic skills. They are given the opportunity to study in a boarding school where they receive intensive coaching. This program is being evaluated through a randomized experimentation. Since there are more candidates than slots for this program, a lottery is conducted each year among potential candidates to determine who will be admitted to the boarding school. Admitted students make up the treatment group, while rejected students make up the comparison group. At the end of each academic year, all students participating in the experimentation take standardized tests in mathematics and French as well as some psychometric tests. The evaluation simply amounts to comparing test results from the two groups of students in those tests.

Having students take such tests is a very easy task in the treatment group: except for a small minority of students who have been expelled from school during the year, they are still in the boarding school in the end of the year. Testing is much harder among students in the comparison group. Since all of them were not assigned to a single school, a major search must to be conducted to find where they are at the end of the year.

Assume that such an effort to find these students was not undertaken. We would probably end up with around 95% of the students in the treatment group taking the tests, against only 80% in the control group. The evaluation would then compare test scores among the 95% of students who took the tests in the treatment group and the 80% who took it in the comparison group. **Despite the fact that randomization ensured that the treatment 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 the 20% of comparison group students who did not take the test were a certain type of students, such as drop-outs or those who had been expelled. This could explain why it was so hard to locate

them and have them taking the test. In the treatment group, fewer students might have dropped out: the great opportunity of being sent to a boarding school incentivized even those who would have ordinarily dropped out to stay in school. Therefore, the comparison of the 95% of students taking the tests in the treatment group against the 80% taking them in the comparison group would amount to comparing apples and oranges: **the treatment group would include weaker students who were prone to drop out from school**. Therefore, such a comparison would probably underestimate the true impact of the Boarding Schools of Excellence.

Let us formalize this principle. Randomization ensures that, by construction, the treatment group and the comparison group are initially comparable in every respect. But because some individuals participating in the experimentation might move to another city during the experimental protocol, or might no longer be willing to answer surveys, 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) cannot be measured. This share is called the **attrition rate**. Consequently we will not end up comparing the entire treatment group to the entire comparison group; instead, we will compare treatment individuals for whom we have been able to measure the outcome of interest to control individuals for whom we were able to measure the outcome of interest. It might very well be the case that those two groups are no longer comparable, for instance if the treatment caused more (or fewer) individuals to attrite from the experimentation in the treatment group. Therefore, when reading results from a randomized experimentation, one should pay a great attention that the **overall attrition rate in the experimentation 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 comparison groups**. If, for instance, the attrition rate of an experimentation is equal to 10% in the treatment group and 20% in the comparison group, one should consider the results of that randomized experimentation with some caution.

To be fair to randomized experimentations, it is important to mention that attrition is a serious threat to the validity of all evaluation methods. Consider 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. Young graduates in 2008 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 multiple dimensions. Not only did the 2008 cohort experience 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. The 2007 cohort would only be made up of the first type of individuals.

The only reason why we mention attrition as a threat to validity when presenting randomized experimentations is that attrition is, by and large, the only serious threat to the

validity of randomized experimentations, whereas all other methods suffer from several threats to validity.

I. Comparing results obtained through different methods

When attrition rates are sufficiently low, or at least balanced between the comparison and the treatment group, and there is no reason to suspect a strong Hawthorne effect, **one can consider that randomized experimentations deliver a measure of the true impact of a program** (allowing for statistical imprecision). There exists a large academic 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 experimentation. This is exactly the exercise we conducted with 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 that differs 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 a wide margin, and it seems very difficult to find criteria which allow us to predict the circumstances in which matching will do well or poorly.

Before-after comparisons usually fall far from the true impact, except when the outcome under consideration is extremely stable over time. This could be the case if, for instance, the outcome was not very sensitive to variations in the economic conditions.

Difference in differences provides significantly better results than before-after comparisons. Sometimes it will deliver a measure close to the true impact, though sometimes it will fall wide of the mark. When data is available over a long period of time, so that the parallel trends test exemplified in figure 5 can be verified over many periods, 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.**

3. Measuring the Efficacy of a Policy Intervention: Some Practical Considerations

Having considered the underlying principles and the reliability of these six methods of evaluation, this section first considers some practical issues related to data collection that are common to all six methods. It then reviews the respective advantages and disadvantages of each method in terms of cost, complexity to implement and acceptability. Methods are ranked from the most 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.

A. Measuring the outcomes

All methods require measuring outcome(s) in the treatment 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, and a specific **survey** must be conducted in order to collect it. Whether measuring outcomes will require specific surveys or not usually depends on the richness of the set of outcomes the experimentation will consider. 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 that one use questions from existing surveys which have already been administered to large populations,** and not design one's own questions. This will ensure that questions are properly formulated and easy for respondents 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 a 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 to measure such psychological traits. Examples can be found here: http://www.er.uqam.ca/nobel/r26710/LRCS/echelles_en.htm . These scales go through a well-defined "validation" process which involves having a large number of subjects answer 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 scale for higher education should demonstrate that high school students scoring high on their scale are indeed more likely to pursue 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 avoid designing this kind of questionnaire ("Do you have a good self-esteem?" or "Do you feel motivated?").

In the counseling program example, only one outcome was examined: 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 an administrative database and required no further data collection. In the Boarding Schools of Excellence example, policymakers and the evaluation team agreed that the treatment could have an impact on students' academic performance, as well as their psychological well-being. Therefore, several outcomes were selected: students' standardized test scores for 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, etc. None of those outcomes was already available in an administrative database, so a specific survey was designed: students participating in the experimentation took tests in French and Mathematics and answered psychometric questionnaires during two sessions of one and a half hours each. Questions probing 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 experimentations do not merely require measuring outcomes among "naturally occurring" treatment and control groups. They also require constructing these groups, which implies a somewhat more complex experimental protocol, though this complexity is often exaggerated. A last example is regression discontinuity, which is not suitable to evaluating all programs. This method will work only for programs whose eligibility rules include a continuous numerical criterion such as age. We will now review all methods emphasizing their respective advantages and disadvantages in terms of cost, complexity to implement and acceptability.

B. Randomized experimentations

Implementing an experimental protocol

Randomized experimentations rely on an experimental protocol which requires some cooperation between policymakers and the evaluation team. This usually entails some supplementary work from civil servants working in areas where the experimentation is conducted. Such operational burden should not be denied: it should be acknowledged beforehand by policymakers, by the evaluation team and by civil servants. But it should not be exaggerated either: in most cases, it is possible to keep the extra workload at a very reasonable level.

The impact of the counseling program described above was measured through a randomized experimentation which lasted for a year. At 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 of unemployment during that month to the evaluation team. In each of those 100 lists, the evaluation team randomly selected some young graduates who would be offered the intensive counseling scheme and some who would be excluded from it. Then the team sent back the list to each of the placement agencies, who offered job-seekers selected for the treatment group the opportunity to participate in the counseling scheme. The experimentation 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 a list to the evaluation team each month, and then to ensure that only treatment group individuals participated in the counseling scheme.

Implementing such experimental protocols is sometimes costly. The total cost implied by the randomized experimentation of the Boarding Schools of Excellence is around 500 000€. This includes the costs of following students participating in the experiment during two years to have them taking standardized tests, as well as surveying their parents once per year. This should be put into the perspective of the total cost of the program if scaled-up, which should be around 500 million €.

Ethical issues

Randomized experimentations also raise some **ethical issues**. Most evaluation teams conducting randomized experimentations try to abide by the following ethical rule: conducting a randomized experimentation 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 eligible young graduates, but that the evaluation team had requested that it be withheld from half of them in order to have a comparison group. Therefore, randomized experimentations are usually conducted when there are more candidates than spaces for a program.

In many social programs, a rule must be used to allocate the program, whether it is a lottery or some other method. **Opponents to randomized experimentations argue that the lottery is not a fair allocation rule because social programs should be allocated to those who need them 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 experimentation 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 could 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 were offered the program!

In medicine, applications to market new drugs must contain a randomized experimentation demonstrating that drug's 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 later compared 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 job-seekers from a potentially helpful counseling scheme. **But doctors still consider that the benefits of such experimental protocols (the ability to precisely measure the effects 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 unclear whether men will benefit more from intensive counseling than women. Therefore, an allocation rule targeting those presumably most in need of the program most could end up allocating the program to those who least need it, if prior beliefs prove wrong. But randomized experimentations enable policymakers to identify subpopulations which will benefit most from the program, so as to optimize eligibility criteria if the program is scaled up.

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

C. Regression discontinuity

Regression discontinuity does not require an experimental protocol. Consequently, it introduces virtually no additional 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 why this method is not used often in practice.

Firstly, **eligibility for 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 to increase the share of young mothers returning to work after a pregnancy. Suppose this program entails giving financial aid to new mothers if they resume work less than 6 months after they gave birth. Assume also that only mothers who gave birth to their third or subsequent child 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 comparing the employment rates of women just below the eligibility threshold to women just above it- that is to say, comparing 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 likely to be too different from women giving birth to their third child (they are probably younger, for instance) to serve as a credible comparison group.

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 in the narrow margin slightly below and slightly above the threshold, so comparisons are based on a small number of individuals. To illustrate this, let us consider the counseling example. The 100 local agencies followed 10 000 young job-seekers eligible for the program in 2008. Assume that all of them were aged between 25 years old and 0 months and 27 years and 11 months. This means that they belonged 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 old. Despite the fact that the initial population is large, **regression discontinuity will result in comparisons between two groups of only 300 individuals each, whereas in a randomized experimentation two groups of 5 000 individuals each would have been compared.** This will result in 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. Indeed, measuring the impact of the intensive counseling program on men alone will amount to comparing the placement rate among 300 / 2 = 150 men just below the threshold to the placement rate among 300 / 2 = 150 men just above it. Similarly, the measure of the impact of the program on women only will amount to comparing two groups of only 150 individuals each. Thus, the impact of the program on both men and women will be estimated very imprecisely, making it very difficult to determine if one effect is larger than the other.

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

D. Difference in differences and before-after

Those two methods consist of 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 for before the program was implemented, as well as after its implementation. In the intensive counseling example, this would mean that placement data for young graduates followed by placement agencies in 2007 should be available. This would not be an issue in the intensive counseling example, because placement agencies always collect placement data for people they follow. If this were not the case, a specific survey would need to have been conducted to measures outcomes, and **the experimentation would have needed planning far in advance**.

In addition to the necessary components of a before-after comparison, difference in differences requires that you find a control group who were excluded from the program both

before and after its implementation because they did not meet the eligibility criteria. Data on the relevant outcome will also need to be collected for this control group. 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.

E. Statistical matching and participants vs. non-participants

Those methods require measuring the outcome of interest for all individuals eligible for the program (i.e. both participants and non-participants) for the duration of the program. In the intensive counseling example, this means that placement data would need to be collected for all young graduates eligible to receive 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 beyond that of 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 Experimentation to Scale-Up

Designing a social experimentation first requires defining a relevant policy intervention. In order to do this some basic steps must be followed, including (1) rigorous description of the social need the policy seeks to address, (2) a precise statement of all the actions which will be part of the policy, (3) an examination of all the constraints faced by the target population to ensure that they will be able to enroll in the experimentation, (4) a thorough literature search of all evaluations of similar policy interventions to gather 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 in evaluating a policy intervention which has been shown to be useless by dozens of previous studies, or to test a program which beneficiaries are not able or willing to participate in.** A large number of experimentations fail because targeted participants are not actually willing or not able to participate. Among programs which prove ineffective at the end of a long and costly experimentation, many could have been confirmed as ineffective beforehand if a thorough literature search had been done.

Once a presumably policy-relevant 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. Section 2 compares those methods from the point of view of the validity and reliability of the results they deliver, while Section 3 compares them from the point of view of their respective complexity to implement in practice.

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

A cheap, not particularly innovative program, for which a large number of studies exist indicating that similar programs are usually cost-effective, **might not need to be evaluated with a randomized experimentation.** However, since the additional cost of evaluating it through a difference in differences methodology relative to a before-after comparison will probably be very low, it is often worth bearing those supplementary costs. **But an expensive and very innovative program should be evaluated through a randomized experimentation, or through a regression discontinuity.**

In the end of an experimentation, results finally come. They enable calculation of the cost-effectiveness of the program, and yield valuable information which subgroups benefit the most from a program. Based on this, policymakers can decide whether or not to scale up the program, and which eligibility criteria should be used. When doing so, they must bear in mind

that **results of the experimentation hold only within the population that participated in the experimentation.** 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 experimentation. Despite the fact this program proved efficient among young graduates who had been looking for a job for more than 6 months, it might not be as efficient among young job-seekers with no degree, or among senior job-seekers.

Finally, methods which yield robust results make the scale-up decision easier and more acceptable to all stakeholders. When a program has been convincingly shown to be ineffective after a well-conducted randomized experiment, it is much easier to make the decision not to scale it up than if some doubts remain on the true effectiveness of the program because it has been evaluated through a less reliable method such as matching. Indeed, more robust methods such as randomized experiments and regression discontinuity are also more transparent to all stakeholders involved in the experimentation, which makes it easier to reach a consensus on whether or not to scale-up.

Appendix: some examples of experimentations which led to scaling-up

Police Skills Training

Training police officers in personality development skills and scientific techniques of investigation can improve victim satisfaction and investigation quality. Evidence from a randomized study has contributed to the scale-up of police skills training for 10 percent of police personnel in Rajasthan state.

http://www.povertyactionlab.org/scale-ups/police-training

Remedial Education

Remedial tutoring for children who have fallen behind academically can improve learning outcomes. Evidence from a randomized study has contributed to the scale-up of NGO Pratham's Read India program in 19 states in India. In 2008-09, 33 million children benefited from remedial education through the Read India program.

http://www.povertyactionlab.org/scale-ups/remedial-education

The Canadian "Self-Sufficiency Project"

Welfare recipients who find jobs may remain poor. The "make work pay" approach rewards those who work by boosting their income. This strategy was the centerpiece of the Self-Sufficiency Project (SSP), a large-scale demonstration program in Canada that offered monthly earnings supplements to single parents who left welfare for full-time work. Launched in 1992, SSP was evaluated through a randomized experiment. SSP substantially increased full-time employment, earnings, and income and reduced the poverty rate - all at a low net cost to the government. The program also improved the school performance of enrollees' elementary school-aged children, a benefit that - unlike the positive economic effects - persisted even after parents stopped receiving the supplement.

http://www.mdrc.org/publications/46/execsum.pdf

Getting parents involved

Parental involvement campaigns significantly increased parents' interaction with schools and improved student behavior in France, and effects spilled over onto classmates whose parents did not participate in the program.

http://www.povertyactionlab.org/publication/getting-parents-involved