European Expert Network on
Economics of Education (EENEE)

Randomized Controlled Experiments in Education

EENEE Analytical Report No. 11
Prepared for the European Commission

Adrien Bouguen and Marc Gurgand

February 2012

11

European
Commission

ABDUL LATIF JAMEEL
Poverty Action Lab
J-PAL EUROPE
TRADUIRE LA RECHERCHE EN ACTION

RANDOMIZED CONTROLLED EXPERIMENTS IN EDUCATION

Report for the European Commission

February 21, 2012

Adrien Bouguen (Paris School of Economics and J-PAL Europe)

Marc Gurgand (Paris School of Economics and J-PAL Europe)

Opinions expressed in this report are those of the authors alone and do not represent the point of view of the European Commission.

1. Introduction

Randomized controlled trials (RCTs) are becoming an important tool for the evaluation of social policies. They borrow the principle of comparing a treated group and a control group that are chosen by random assignment from medical science, where they have been a standard since the Second World War. They provide a robust and transparent way of eliciting the causal impact of an intervention. The method is not new in educational science, especially in relation to cognitive science, but it is mostly used in very controlled, quasi-laboratory, environments, and on rather small samples.

In contrast, social experiments test full-scale policies or interventions in a social context that should be as close as possible to the conditions of a scaled up program. Such RCTs, including those in education, are most developed in the US and in developing countries. For instance, an important program, the STAR experiment, tested the effects of class size reduction in the US during the 1980's. It has long been the only robust and convincing evidence in favour of such an important policy. In developing countries, the PROGRESA experiment in Mexico has demonstrated that conditional cash transfers to the poor can both ensure redistribution and encourage school enrolment. The strength of the RCT evidence has encouraged a large number of developing countries to test and adopt similar policies.

RCTs in general, and particularly in education, are much less common in Europe. In Section 5 of this report, we provide a table of some experiments that we could identify. A number have been run in France, because France began implementing RCTs at quite a large scale around 6 years ago. However, only very isolated instances could be found in other countries, and only rarely from a government initiative. We are not aware of strong support from a public or private organization except the French Experimental Fund for the Youth. Nor are we aware of any multi-country operations.

Nevertheless, as illustrated by the PROGRESA example, RCTs are a unique tool for assessing and disseminating innovations among countries with similar (if not identical) educational issues and contexts. This makes a great deal of sense at the European level, and the European Commission has a role to play in encouraging such practice among member states. Achieving this will require a precise understanding of the general logic and issues of RCTs in education.

This report is based on our experience at J-PAL Europe. It mainly draws example from one country (France), but we also illustrate our arguments with projects from other European countries. We start with the basic principles of RCTs: counterfactual, statistical precision, types of designs (Section 2). These notions may seem relevant mostly to researchers, but they shape any project and therefore must be well-understood by all parties. In particular, it is important to be aware of the possible randomization designs that can help fit into the operational constraints of a given project. We discuss how results from RCTs can be interpreted and their potential limitations, in particular the well-known issue of external validity (Section 3). We then move to more practical implementation (Section 4). We give details on how to determine the evaluation criteria, the implications of data collection, elements of budget, and the costs and benefits of an RCT to the various partners. This is important to consider, in order to anticipate their capacity to hold with the project and accept risks of failure. Finally, we present our understanding of the potential for RCTs in education at the European level (Section 5).

2. Principles of Randomized Controlled Trials (RCTs)

2.1. Counterfactual problem and randomization

When trying to assess the causal relationship between a policy and an outcome of interest, one needs to resolve the counterfactual problem: what would have happened to the beneficiaries if the policy had not been implemented? Take, for instance, the example of a remedial education program aiming to improve the educational achievement of deprived pupils at grade 1 (six years old). To rigorously assess the efficacy of this policy, one would ideally compare, after the end of the intervention, the treated pupils with exactly the same pupils had they not been treated. Of course, in a given period of time, someone is either exposed or not exposed to the same program—they cannot be both: this is where the counterfactual issue hinges. Most of the time, comparing the treated individuals with a group of non-treated individual does not solve the problem: pupils allocated to a program tend to possess a number of characteristics that intrinsically differentiate them from the non-treated individuals. Likewise, comparing the same individuals at different points in time (before and after the treatment for instance) does not usually resolve this problem.

The reasons that treated individuals cannot be easily compared to those who did not receive the treatment are numerous and depend on the context of the evaluation. Typically, when participation is voluntary, individuals who participate in the program are more skilled and/or more motivated than the ones who refuse to participate. Inversely, and returning to our previous example, if underachievers are the target of a remedial education program, they are more likely to be selected into the program: the treated group will then be relatively less skilled and/or motivated than the untreated group. Because a policy aiming at improving achievement of underprivileged students will primarily benefit underachievers, a simple comparison of treated individual with non-treated individuals will likely under-estimate the impact of the program.

Several econometric techniques attempt to generate ex post (after the policy ended) a group of individual that resembles the ones who benefited from the policy: “multivariate regression”, “statistical matching”, and “discontinuity design”. Although valuable, those techniques rely on a number of untestable assumptions that will always be subject to discussion[1]. It is hard for such methods to deal with all *unobserved* factors. In the case of program evaluation, the decision (either taken by the beneficiary or by the provider) to participate in a program is very often linked to factors that are very difficult to observe or control for (motivation, intelligence, readiness…).

To avoid such pitfalls, we would like to assign individuals to treatment independent of any factors that may affect their final outcomes. Random assignment can ensure such conditions are met. When individuals are assigned randomly, and the sample size is large enough, control and treatment groups can be expected to have exactly the same initial characteristics. Consequently, at the end of the evaluation, **any differences in the outcome of interest (test scores, drop-out rates...) can be interpreted as the treatment effect.** Returning to the remedial education example,

[1] (Schlotter, Schwerdt, & Woessmann, 2009)

the researcher can randomly assign individuals into treatment and control groups. **At the end of the program, a comparison of the mean test scores of both groups will directly give the treatment effect**.

2.2. Sample size requirements and precision

Since one only has statistical data on the actual impact, the observed difference between treatment and control groups is only an estimation of the "true" impact. It is important to assess the precision of this estimate: how sure are we that the "true" impact is close to the impact that we can estimate given our particular sample? RCTs, like other statistical evaluations, only provide estimated impacts, and confidence intervals should be considered carefully.

One obvious factor that influences precision is the sample size. As the sample size increases, the precision of the evaluation improves. The fundamental reason for this is that, as sample size grows, treatment and control samples tend more and more to have exactly the same composition. For instance, if each group has 2 individuals, it is unlikely that each will have the same gender composition. This is true of any characteristic, observed or not. With small sample sizes, differences in mean test scores would occur even if there was no actual impact of the treatment, because treatment and control groups were not similar enough at the baseline. As sample size increases, we can have more confidence that results are not due to remaining differences between the two groups.

Although often a central concern, sample size is not the only factor that affects statistical precision. It can be shown that precision is maximized when the probability of being assigned to either group is 50%. However, for practical, ethical or political reasons, the evaluator may prefer to assign more individuals to the treatment group than the control group. Although the estimation remains reliable on average (unbiased), the precision of the evaluation will be lower. Many other factors impact the precision of the estimate. The response rate, the participation rate (often called take-up), or the number of clusters of individuals, are factors that may play an important role in the determination of the precision. Suffice it to say that, from a policy perspective, **precision should be considered as a central issue**.

How much precision and what sample size should we pursue? A more precise evaluation – which generally implies a larger sample – allows us to detect a "true" treatment effect, even if it is smaller. Practically speaking, a policymaker should come up with a level of impact that he considers sufficient to justify scaling up the program. Then, **the evaluator is able *ex ante* to determine which sample size N and which design will enable the statistical detection of this minimum level of effect.**

To illustrate the way ex-ante decisions are made on sample size, let's return to the remedial education example. Imagine that the goal of a program is to reduce the proportion of first graders who get held back (grade retention). The policymaker wants the gains from the reduction in the number of children who repeat first grade to at least match the cost of the remedial education program. If the direct cost of a first grade education year is 6000 Euros, and 20% of pupils repeat it, grade repetition costs the education system 1200 additional Euros per child. Imagine now that the cost per child of the remedial education program is 500 Euros per year. What is the retention rate reduction that would match this additional cost? This is: 500/6000=8.3%. To ensure that the remedial education program is cost-effective, the evaluation should be able to detect at least an 8.3% reduction in the retention rate. Using basic statistical tools[2], we can show that, with the simplest randomization designs, the sample size should be at least 722 children, half of whom should be randomly assigned to treatment group and half to control group. Of course, the assumptions made previously are extremely optimistic: response rate is always lower than 100%, compliance is never perfect and individual randomization is probably not feasible in that context. As will be shown below, this will change precision and the relevant sample sizes. Nonetheless, this example illustrates that with relatively simple statistical tools, it is possible to determine ex-ante the sample size and the expected effect size of a policy.

2.3. Different types of design and ethical issues

The constraints imposed by the protocol of RCTs (sample size, proportion of treated, random assignment…) on the partners[3] are likely to cause complications during program implementation (see Section 3). For the sake of clarity, the examples of RCT design we have used so far were rather simple: individuals were simply randomly assigned (50%/50%) to a treatment or control group at the beginning of the program. They were implicitly assumed to comply with the assignment and to answer surveys or take test with no exception. Things are rarely that simple. Fortunately, many experimental designs have been developed to tackle issues regarding implementation or ethics, and in practice, the simple intuitive design is only rarely used. We now describe various alternative designs, and evaluate their interest both in terms of precision and of acceptability.

2.3.1. Oversubscription: a natural opportunity for RCTs

A natural opportunity for introducing randomization occurs when the demand for a service provided by the program exceeds the supply. This excess may be caused by implementation issue (the provider is unable to provide the program to the whole population) or to limited resources (the provider has the budget for only a certain number of beneficiaries). For instance, consider an RCT aimed at evaluating the effects of boarding schools.[4] Since the number of available dorms in a specific boarding school is limited, it is rather natural to randomly select a number of students out of the pool of eligible candidates. Practically and ethically speaking, this method is usually well-accepted by both providers and beneficiaries. Some even consider random assignment to be fairer than a selection based upon criteria that may not be accepted by all.

2.3.2. Costs and benefits from changing the unit of randomization

To alleviate legitimate ethical concerns, it is also possible to change the unit of randomization. A randomization at school level may address this issue: all eligible students enrolled in treatment schools will be offered the program, while in control schools no one will receive the program.

A second main argument in favour of changing the level of randomization is contamination and selection of the treatment group. Think of a program to test teaching methods. One could randomize half the teachers in each school to receive the training. This strategy would entail two

[2] (Duflo, Glennerster, & Kremer, 2006)

[3] The partner can be the public service in charge of the policy, an NGO, an association…

[4] Such an RCT is currently being implemented in France by J-PAL.

potential risks for the validity of the experiment. Since all teachers in a school evolve in the same workplace, treatment teachers could potentially share their teaching methods with their control counterparts. The control group would then be contaminated. Second, unless schools agree to also randomize pupils into classes, the school head may assign specific pupils to treated teachers. If he wants to favour the program, he may assign relatively better students to the treatment branch, which would produce over-estimates of the effectiveness of the program. To avoid selection and contamination, a randomization occurring at school level would be preferable.

Another reason to favour a change in the level of randomization is the detection of peer effects. When randomization occurs at school level, and only part of the school is actually treated, it is possible to analyse the effect of the program on the non-treated students in treated schools. Going back to our example, if the remedial education covers only 10 pupils per schools – that is the 10 lowest achievers in the school – the other, say, 40 non treated pupils could be indirectly affected by the program: they may benefit indirectly from the tools and advice given to the treated individuals, the regular teacher may be able to spend more time with non-treated kids, the class dynamics may be stimulated, etc. The comparison between the 40 non-treated pupils in treatment schools and the equivalent 40 non-treated pupils in control school will allow capturing the peer effects. Surprisingly, in some situations, peer effect seems to play a larger role than one might expect: in an evaluation of a simple parental meeting program in France, Avvisati et al (2010) find strong peer effects on 6th graders' behaviour.

However, it comes at a cost: If the randomization does not occur at the individual level (pupils/students in education context) then the cost in term of precision can be high. In the remedial example, while 722 pupils were sufficient to capture the 8.3% effect when randomization occurred at individual level. If randomization now occurs at the school level, sample size has to be substantially higher. In a rather favourable case, a minimum of 1,500 pupils would be needed when there are 10 pupils per school eligible to the program, implying 150 schools. The sample size would have to double[5]. Randomization can also occur at grade level: to ensure that every school benefits from the program, one grade may be randomly chosen within each school to participate in the program: all students in this treatment grade will then benefit from the treatment, while other grades will not. Since the experimental population includes many other schools, the treated first grade in one school will have corresponding control first grades in other schools. This astute design alleviates ethical concerns, will reduce the number of schools surveyed and increases statistical precision[6].

2.3.3. Phase-in design: a popular way to secure fairness

Another typical way of increasing acceptance is a phase-in design. Instead of implementing the program at its full-scale in the first year, the provider may decide to cover only a share (usually half) of the targeted population that year. Those who did not benefit from the treatment during the first year will receive treatment in the second year, while the first year beneficiaries will not benefit from the program in the second year. This type of design is particularly well fitted for educational

[5] (Duflo, Glennerster, & Kremer, 2006)

[6] Because there are typically less differences (variation) between students in the same grade than students in the same school, the precision will be higher using grade randomization.

policies. For instance, in a remedial education program for first graders, only half of the schools can receive the financial resources to implement the program in the first year. The first graders of the schools who did not participate in the program during the first year form the control group. Then those schools will have access to the treatment during the second year. This design ensures that both groups of schools receive exactly the same resources over the two years of the experiment.

2.3.4. Encouraging instead of assigning

When exclusion from treatment is really unacceptable for ethical or practical reasons, an encouragement design provides an interesting alternative. Instead of assigning individuals to an experimental group, it is possible to randomize encouragement to participate to the program: individuals in the treatment group will receive the encouragement while the control will not. This encouragement could be a letter, a commercial, a phone call, or an email. Under this design, control individuals can still participate in the program if they want to but, providing that the encouragement induce more participation, a larger share of the treatment population will be in the program. Observed outcome differences between treatment and control groups can still be used to identify the impact of the policy. While appealing for ethical reasons, this method has drawbacks: typically, a lower share of the sample will participate in the program, and thus the statistical precision of the evaluation will be much lower.

3. A theory in practice: RCTs alongside a program implementation.

Educational science has long used quasi-laboratory experiments, often RCTs, to evaluate interventions, particularly to study learning mechanisms. Social experiments that are the topic of this report have a much larger scale and evaluate interventions in a social context identical to that of envisioned scale-up. Methods developed by cognitive psychology, which have proven efficient in quasi-laboratory settings, may then be less efficient or simply inefficient when employed in the outside world. This essential feature of RCTs also generates some inconveniences. Since RCTs closely follow the implementation of a real policy, results are somehow dependent on the implementation of the policy itself. First, RCTs are not always able to understand the social, psychological or economic mechanisms that have driven the results. Second, RCTs are not always able to yield long-term results. Third, because RCTs evaluate local programs, results can be difficult to generalize to other contexts. In this section, we will look closely at each of these points and try to offer solutions.

3.1. Testing a policy or elements of a policy: Do RCTs evaluate a black box?

An education program can rarely be restricted to a single effect. It usually affects many actors (the students himself, his parents, his professors, his peers…), through different channels (influence, incentives…) and by different means (motivation, money…). Returning to one of our previous examples, a boarding school may affect boarders' achievement in many different ways: boarders are separated from their own social and cultural environment, the boarding school usually provides cheap housing, it fosters new interactions within dorms, it frees time and space in the family home. Each of these channels may positively or negatively affect the achievement of the beneficiaries. If a simple randomization is undertaken – eligible candidates are simply assigned to treatment and control branches – these various elements of the program will not be disentangled: **only the overall treatment effects – a compound of all previously mentioned potential channels – will be identified**. One of the first randomized experiments in education – the US program STAR in the 80's – faces exactly this limitation. The program aimed at evaluating a class-size reduction on pupils' achievement[7]. While authors find positive effect of the class-size reduction, the design of the STAR experiment did not allow them to precisely describe how this effect was generated: Was this effect caused by better and more numerous interactions between pupils? Or from fewer classroom disturbances? From the introduction of new teaching technique (that would have been impossible to implement in larger groups)? From a higher level of teacher motivation? Such questions could not be addressed given the experimental design.

This may or may not be an issue for the policymaker. To be sure, the policymaker may be satisfied with the overall effect of its policy. However, **if one wants to understand more thoroughly the social and cognitive mechanisms which account for schooling achievement or to evaluate which aspects of the program have the greatest effects, it might be interesting to consider multiple randomization designs**. In that type of design, several different treatments (or element of a policy) are tested simultaneously with randomizations. A recent paper by E. Duflo, P. Dupas and M. Kremer (2009), was able to identify how tracking in Kenya (class composition based on child achievement) affected pupils' outcomes and teachers' teaching technique. In that case, thanks to some features of the design and data collection, they were able to conclude that tracking affect positively both higher and lower capacity pupils by modifying the way teacher convey their knowledge to their audience. This paper goes beyond simple program evaluation and tries to account for the channels through which impacts were generated.

Most of the time, looking inside the black box of a policy implies introducing multiple variations of a program. A typical example is Conditional Cash Transfer (CCT) programs, in which an allowance is given to families conditional on children's' school attendance. While several papers have shown the effectiveness of CCTs[8] in developing countries, few were able to disentangle the effect of the cash from the effect of the condition. In a recent paper (Baird, McIntosh, & Özler, 2011), a multiple randomization strategy was undertaken to address the question of the efficiency of conditionality in Malawi. Part of sample was randomly assigned to a treatment group – say T1 – which was given the allowance conditional on school attendance. Another treatment group T2 was given the allowance unconditionally. The last group was not provided with any allowance (control group). By comparing T1 and the control group, the researchers were able to identify the effect of the full CCT package. By comparing T2 and the control group, they were evaluating a simple cash transfer program. Finally, by comparing T1 and T2, they were able to measure the distinct effect of the conditionality. Results show that conditionality does matter to increase school enrolment.

[7] (Krueger & Whitmore, 2001)

[8] See for instance (Schultz, 2004)

3.2. Short term vs long term evaluation

Since returns to education accrue over a long period of time and investment in education is typically expected to generate benefit decades later, not being able to capture long term effects is a clear limitation for evaluations. We will first investigate the reasons why researchers and policymakers may be interested in long-term effects, and then will assess whether RCTs are able to pick up long-terms effect.

3.2.1. Interests of a long term evaluation

Depending on the context, several reasons may justify a long-term analysis. First, the body responsible for the implementation of the policy may consider that the effect of its policy will only be substantial after a relatively long period of time (i.e. more than one year). When the program is implemented over a long period of time, the evaluator should ideally follow the experimental cohorts (control and treatment group) and ensure that the protocol (i.e. assignment to treatment) is enforced as long as the program goes on. Of course, when duration is too long, the evaluator and the partner must jointly decide on a number of years of treatment that will be sufficient to detect the full effect of the program.

Second, one may be interested in testing whether an accumulation effect exists: increasing schooling performance in one year may increase school results in subsequent years. For instance, an early intervention that reduces illiteracy at the end of the first grade (age six) may increase the passing rate in grades two and three. Conversely, effects may fade out after the program ends. In that case, the policy simply accelerated the performance of the treated group but did not change their future achievement on the long run. For instance, one of the mitigating results from the previously mentioned project STAR was that the effect of the class-size reduction vanishes quickly after the end of the program. Although now contested[9], this finding completely changes the conclusion of the evaluation: class-size reduction seems to produce a structural effect only if it can be generalized to all grades.

Third, the policymaker and the research team are often interested in learning more about how school performances translate into career orientation. In that case, the evaluator must follow the experimental cohorts long after the end of a policy. In some political or institutional contexts, increasing school results (reducing absenteeism, increasing test score…) may not be enough to reach a policy objective. If the eligible population is discriminated against in the labour market, increasing schooling performance will not resolve the problem faced by the targeted population. Being able to pick up the way school results will be valued on the labour market should then be considered as a main research question.

3.2.2. Can RCTs pick up long term effects?

A long term RCT poses a certain number of difficulties. Because RCTs are implemented in the “real word”, **it is sometimes difficult to organize the follow-up of cohorts of students over a long period of time**: as time goes by, evaluation cost increases, response rates fall (especially if the individuals are followed after the end of the policy) and the probability of contamination (control group having

[9] (Chetty, Friedman, Saez, Schanzenbach, & Yagan, 2011)

access to the equivalent program) increases. Encouragingly, solutions exist. When countries already have a unique school identification number, it is theoretically possible to follow the school career on a long-term basis. In France for instance, a unique identification number is allocated to any middle school student from grade 6th till the end of his education. Provided there is good cooperation between the evaluator and the school districts in charge of education statistics, it is theoretically possible to establish a long-term follow-up. In the US, an innovative collaboration between the IRS (Internal Revenue Service) and a team of researchers has rendered possible the very long-term evaluation of the previously mentioned project STAR[10]. The research team was able to analyse how treated individuals performed in the labour market 20 years after the end of the policy. Since most European countries have well-kept national statistics, there are no technical reasons not to organize long-term evaluations in the European context.

3.3. External validity

Another legitimate concern when it comes to RCTs is their *external validity.* Thus far, we have mainly focused on issues related to the *internal validity* of the RCTs: when properly designed, RCTs generate reliable, robust and precise estimate of the impact of a program. However, since RCTs are implemented locally, some evaluations may lack "external validity" i.e. it is not clear whether an impact estimate drawn from one specific evaluation would carry over to other economic, social and institutional contexts. The reasons for this are twofold: first, when scaled-up, a program may generate effects that were not picked up in the local evaluation – this is called the *general equilibrium problem*. Second, the context of the evaluation (the place, the time, the eligible candidates, the implementer) may modify greatly the final result – RCTs then call for *replications*. It should be emphasized that the external validity problem is in no way specific to RCTs: it holds in any empirical evaluation that derives from the observation of a policy in a specific context. This is in fact a general question for the implications of empirical knowledge in social sciences. However, RCTs can bring specific means to deal with this issue.

3.3.1. General equilibrium issue

Some education programs, which have shown great success locally, may have little effect when generalized. Imagine that a group of schools is given the opportunity and the financial support to recruit the most qualified teachers. An RCT is conducted and concludes that, while costly, the program is very effective in increasing school results. Can we conclude that the same positive impact would carry on if all schools were given the same opportunity? If the number of skilled teacher is limited, the first school to be given the freedom and the resources to recruit will certainly perform well, but what about the last one? As the number of treated schools increases, the competition to recruit skilled teacher will rise. If all schools have exactly the same budget per student, the final effect – in equilibrium – should be relatively neutral. The only noticeable changes would be the distribution of wages offered to teachers: providing that skills can easily be observed (which is often not the case), the most skilled teacher will be paid higher wages. To be sure, that new wage distribution may induce positive outcomes through higher teachers' motivation. But it may also generate negative consequences: less desirable neighbourhoods may be unable to attract motivated and skilled teachers. **In any case, the *partial* results found in the local evaluation are**

[10] (Chetty, Friedman, Saez, Schanzenbach, & Yagan, 2011)

uninformative about the *general* effect generated by the scaled-up program. Depending on the type of program, the general equilibrium effect may or may not be an issue: programs deemed to be made available to the whole population are primarily concerned by this effect.

Not all interventions bear the potential of such equilibrium effects (think of learning practices). Depending on the potential general equilibrium effects of the evaluated program, policymakers should be careful in their interpretation of any RCT result. Nonetheless, if a stakeholder is interested in evaluating policies that may have such effects, some designs are able to capture the general equilibrium effect by randomizing not only the schools but also the treatment intensity within school districts, for instance. Such designs have already been tested in other contexts[11] with interesting results. Policymakers should, however, bear in mind that such designs are costly and require full cooperation of the national schooling administration.

3.3.2. Replications

Because RCTs are implemented in a specific context, the results from one evaluation cannot be necessarily transposed to another context. When an evaluation concludes that tracked schools perform better than non-tracked schools in Kenya[12], nothing can necessarily be said about the impact of a similar program introduced in a European country, or even in other developing countries. The different economic, social and institutional context renders the comparison difficult to impossible. Likewise, when a program in a country is replicated somewhere else, it rarely preserves exactly the same design: what tracking means may be very different in a European, Asian, American or African context. Previous evaluations should thus play a simple informative role in deciding whether to implement new programs or not.

This problem may be less acute in the European context. A result found in one European country should then be relatively more comparable across European countries than in other region in the word. A recent randomized experiment in Switzerland about adult education impact on labour market outcome, which shows that adult trainings have small effect when improperly targeted,[13] may have large repercussions on the way other European education system will think about this kind of policy. At minimum, it can raise doubt on similar training programs implemented in other European countries. It might also encourage governments to start their own evaluation of their national program to check whether results found in Switzerland can be generalized. The recent development of RCTs in Europe is thus a unique opportunity to accumulate useful knowledge about the efficiency of education policy in developed countries[14].

4. Implementation issues

The dialogue between a project implementer and an evaluator starts with deciding whether an intervention is well suited to an RCT and what adjustments may be required. RCT-based evaluation

[11] (Crépon, Duflo, Gurgand, Rathelot, & Zamora, 2007)

[12] (Duflo, Dupas, & Kremer, 2009)

[13] (Schwerdt, Messer, Woessmann, & Wolter, 2011)

[14] see section 5 for further details about the possible implementation of an RCT at European level

of schooling policies are appropriate in many cases because schools or classes form well-identified and separate units over which it is possible to make some intervention vary. Furthermore, the population is well-defined, can be easily sampled and is readily available for implementing many forms of measurement. This is in contrast to many social policies where even a list of potential beneficiaries may be hard to build, not to mention surveying issues. However, RCTs are not appropriate to every schooling intervention and its outcomes must be precisely defined.

4.1. DEFINING THE OBJECT: WHAT TYPES OF POLICIES CAN BE EVALUATED USING RCT?

In order to run an RCT – or in fact any sort of impact evaluation – one needs to consider an intervention that can be *manipulated*. This means that there can be treated and non-treated subjects, and the intervention can be precisely delimited in such a way that it is clear whether someone is or is not treated. If one wants to test the impact of class reduction, one must be able to change only class-size component. Unfortunately, in many cases, programs mix various types of interventions. If it is class-size reduction *plus* teacher training, then the evaluation will capture both effects as a set. For instance, in a French parental involvement program[15], parents were invited to three meetings with the school staff and they were offered additional intensive trainings. The policy's maker goal was to evaluate the policy as the whole – meetings plus trainings. Unfortunately, very few parents followed up with trainings. Given the way treatment was originally defined, the only policy that can be evaluated here is *offering* training and meeting. In practice however, we are really learning about the usefulness of parent-teacher meetings because training participation was very marginal.

Furthermore, if the evaluation is to be useful from a policy point of view, the intervention should be formalized so as to be *reproducible*. For instance, asking schools to work with a science museum with little additional guidelines is not an intervention well suited to an RCT. One can randomize which schools are selected into treatment, but the interpretation of the impact will be very loose: if we find no impact on pupils' motivation for (or capacity in) science, is it because schools did not know how to setup well-functioning partnerships, or is it because such project did not foster pupils' interest? Rather, such a policy needs piloting, with a careful qualitative evaluation. Only when a well-defined program is available, and all treated schools would take similar action, does it make sense to run an impact evaluation.

4.2. WHAT WILL BE THE CRITERIA OF EVALUATION?

The criteria against which to judge the relevance or the desirability of an intervention are based on values and objectives that may differ according to different stakeholders, interests or points of view. It is a well-known aspect of the evaluation process that these criteria must be agreed upon beforehand, and this is, in essence, a political process.

In schooling policies, cognitive capacity measured by test scores seems to be a rather obvious outcome. It is therefore probably easier to find a general agreement over the main criteria, as compared to other social interventions. However, even this is open to debate. For instance, one can take a "literacy" perspective (typically followed by the PISA evaluations) or use tests that follow

[15] (Avvisati, Gurgand, Guyon, & Maurin, 2010)

closely the school curricula. Both strategies may imply different visions of the objectives given to the school system at large, and the measurement tools must be thoroughly discussed if the conclusions of the evaluation are to be widely shared.

Non-cognitive outcomes should not be ignored. Child behaviour is intrinsically important in the short-run, as evidenced, for instance, by the ever-increasing attention received by the issue of violence in schools. It is also important in the long-run: the effects of non-cognitive characteristics on the labour market are now well documented. Some evidences even suggest that on the long run, non-cognitive skills are more valued on the labour market than cognitive skills. Non-cognitive traits are also typically more long lasting whereas cognitive skills tend to fade out.[16]. Finally, non-cognitive outcomes such as work behaviour or intrinsic motivation for schoolwork may also influence cognitive ones. One may even wonder if cognitive outcomes are properly measured by cognitive tests. Such tests are administered for the sole sake of the evaluation but do not represent anything for the students. It is possible, for instance, that only intrinsically motivated children will exert effort while taking the test. In that case, the test it will capture to a larger extent non-cognitive traits – such as intrinsic motivation - than cognitive aspects[17].

School career is another dimension to be measured. Tracks chosen or obtained, and the decision to remain at school or enter higher education are outcomes of the utmost consequences to individuals, and they can be influenced by many school interventions. Measuring them may also imply that beneficiaries must be followed over a long period of time.

Also, distributional aspects may be considered important: it may not be enough to learn that a policy raises the average level of pupils; we may want to know if it increases of reduces inequality in outcomes. One way to assess this dimension is to explore the heterogeneity in treatment effects. This must be planned upfront because it requires larger sample sizes to evaluate many different impacts, as compared to just one average.

In the end, it is obviously preferable (though costly) to measure a large set of outcomes. In the Perry Preschool project experiment in the US[18], hundreds of outcomes were measured, even decades after the intervention. These included not only a set of cognitive and non-cognitive outcomes by the end of the intervention itself, but also college enrolment, labour market and housing condition, crime, etc.

4.3. Data collection

One specific difficulty when a program is evaluated using an RCT is data collection. When other evaluation techniques (difference in difference, multivariate regression, matching) may base their analysis on already collected data, RCTs – because they are conducted alongside the program implementation – require that data be collected as the program goes on. This has advantages and drawbacks. On one hand, the evaluator must organize a costly and sometimes complicated data collection. On the other, because the evaluator monitors directly the data collection, he/she can ensure a high level of data quality.

[16] (Heckman, Stixrud, & Urzua, 2006)

[18] (Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010)

4.3.1. Three phases of data collection

Three phases of data collection should ideally be planned: before the beginning of the treatment (baseline), in the course of the program implementation (follow up) and at the after period of treatment (endline).

In principle because randomization insures comparability between groups, collecting a full baseline database is unnecessary. Nonetheless, a minimum of information needs to be collected at baseline level. First and foremost, the evaluator should be able to identify and locate all individuals in its experimental population. In addition to the information collected during the randomization process (name surname group assignment…), the evaluator should collect any information that will ease the locations of each individuals included in the experimental population (control and treatment group). This includes contact information of the legal guardians, of the school, the friends, other family members (grand parents, siblings…), youth worker… Second, although not formally necessary, the experimental population may be surveyed at baseline level using all or part of the same instruments of measurement that will be used at endline. This can be useful for precision matters or to examine interactions between initial conditions and the impact of the program.

In addition to this initial data collection, the evaluator should collect some data during the implementation of the program (follow up data or monitoring data). First, the evaluator will have to document how the research protocol is respected: Are the treatment individuals treated? Are the control individuals non-treated? Second, the research team will try to get a sense of the intensity and the quality of the treatment. This may include informal visit in the treated schools, sending questionnaires to the treated students to ask about activities that should be included in the treatment or simply asking school how much resources have been affected for the program. Such information will help the researcher to better understand the results that he/she will find at the end of the program.

Finally, the evaluator must conduct an endline collection which includes all set of information necessary to build the indicators used for the evaluation. In the educational context, this generally includes one or several standardized tests (cognitive and non cognitive), several questionnaires administered to the teachers, the parents, the students, possibly the school administration and some administrative data (passing rate, enrolment in the next grade, information about school career…). If one is interested in the long-term impact of the program, several phases of endline surveys may be organized. For instance, a questionnaire or an administrative data collection can be organized to capture how students who benefited from one education program achieve in the labour market long after the end of the program (see 3.2.2. Short term vs long term evaluation).

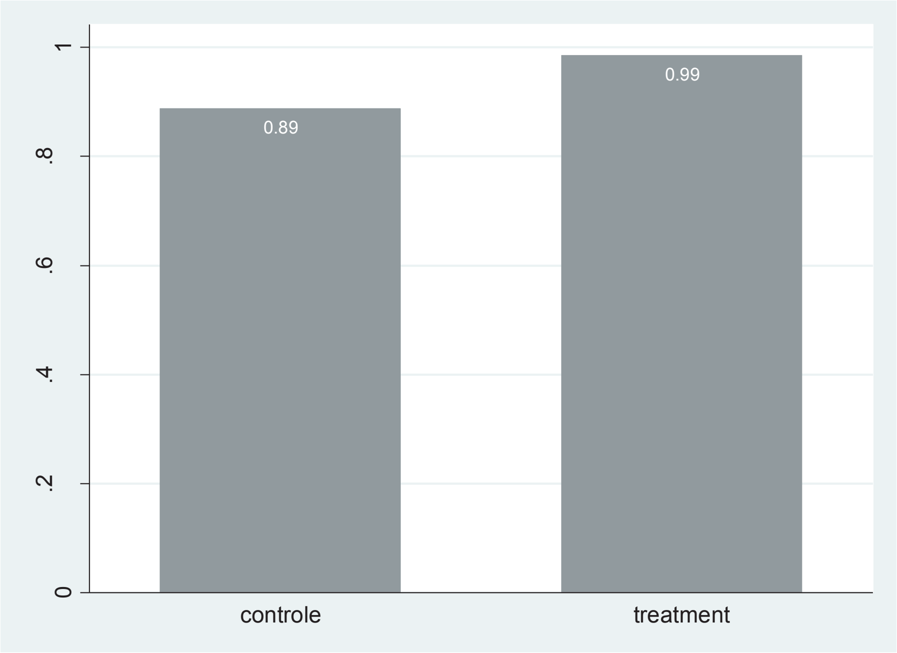
4.3.2. Data collection should be homogeneous and based on "objective" measures

A few essential rules must guide the way data is collected in an RCT. First of all, data must be collected *in a similar fashion* in the whole experimental population - that is, in both the control and the treatment groups. Randomization ensures that both groups are statistically similar. However, if measures used to compare both groups are different, the difference in data quality may cause us to spuriously attribute those differences to a program effect. If the control group is evaluated using school grade and the treatment group is evaluated using a standardize score, results won't be comparable. Likewise, the evaluator should avoid using instruments that rely too much on individual perceptions: even if the instruments are identical, individuals will report differently in the control and the treatment groups. In our example, imagine that instead of measuring achievement using different instruments, the evaluator decides to use solely teachers' grade at the end of the school year. The problem mentioned above will remain: because teachers in the treatment schools benefit indirectly from the treatment, they may over-estimate the school achievement of their students. Results found by comparing treated and controlled students will be unreliable (biased). That is the reasons why standardized tests are so commonly used in experimental and non-experimental settings.

4.3.3. Non response (attrition) should be minimized

When does attrition pose a problem for the accuracy of the RCT estimate?

If the quality or the quantity of data collected is dependent on the treatment status, then attrition will pose a problem. To illustrate this point, let's consider that the main criterion in our evaluation of boarding schools is the passing rate on the final high school exam (*Baccalauréat*). In France, only students who pass the B*accalauréat* are allowed to enrol in tertiary education institutions. Let's first consider a situation where the research team is able to collect information on 100% of the experimental population. At the end of the data collection, the following results are found:



Because the random assignment ensures that both treatment and control are alike, the difference between the average pass rate in the treatment group and the average pass rate in the control group is 99%-89%=10%. The high school boarders seem to have outperformed the non-boarders.

Now imagine that the evaluator was not able to collect the *Baccalauréat* results for some control group students (say, 5% of them). This may well be the case. While it is rather easy to collect data on the treated individuals who are enrolled in the boarding school[19], the control individuals, who are scattered in different control schools, are much more difficult to locate. For that reason, the attrition rate will be typically higher for the control group. The attrition is said in this case to be *differential.* **As a consequence, control and treatment will not be comparable anymore and estimated results will likely be biased.**

Solutions to deal with attrition

As we have just seen, because attrition may unbalance control and treatment groups, it may undermine the efforts undertaken to randomize assignment prior to the beginning of the program. **Attrition is thus considered as one of the most serious threats to analysis**. Consequently, the objective to cover 100% of the experimental population is of utmost importance. Several actions can be undertaken by the research team. Ex-ante (at baseline level) the evaluator should pay attention to the collection of contact information. When anticipated attrition is high, the evaluation team must collect as much contact information as possible, ideally from different sources (administration, questionnaire…). Ex-post (during the endline collection), when the research team observes a low and unbalanced response rate, he/she can organize an intensive tracking effort. Of the number of students for whom the data is missing, the research team can re-randomize a subset of students who will be intensely tracked. For these individuals, the research team, the schools and schools district can devote financial and human resources to reach a 100% response rate. This may imply transporting a surveyor to the new location of the student, organizing online tests, taking a research team to another country… Since it only concerns a very small subset of the population, this solution can be financially affordable. This technique reduces precision but conserves the comparability between groups. The final results will remain reliable. Other econometric techniques aiming at either reweighting the sample, predicting missing values (through multivariate regression, matching) so as to account for attrition do not resolve the problem satisfactorily, because they require untestable, often strong, assumptions. In any case, such techniques should be used only as a last resort when all other possible measures have been undertaken to minimize attrition.

4.4. Costs

Many evaluations based on observational data are inexpensive at the margin: a policy is in place and data already exist (whether administrative or survey data). For instance, the evaluation of the impact of class size by Angrist and Lavy (1999) is based on a regression discontinuity design. In Israel, schools must open an additional class when they reach a certain maximum pupils per class threshold: this generates exogenous changes in class-size that happen in the system without a specific intervention. Furthermore, outcomes are measured with a national testing program that is unrelated to the class-size issue. Evaluating the impact of class-size reduction in this context incurs very little specific cost. But, it is only by chance that it proved possible to do so. The story would be entirely different with an experimental class-size evaluation. In the absence of a convincing quasi-

[19] On the one hand, they are less likely to enrol in another school institution since the boarding school provides them with a relative higher level of school services. One the other hand, because boarders remain on site, they are less likely to be absent (this might not be a problem for collecting B*accalauréat* results but may be if we give a cognitive test to the control and the treatment group on a specific day : absenteeism should be higher in control schools)

experiment, it is the only way to learn about the impact of class size. But this implies a specific budget for the operation itself and another one for the evaluation.

4.4.1. THE COST OF THE INTERVENTION

The operational cost of an experiment can be very small or very large. Consider a class-size experiment on the order of the US STAR experiment (2,000 pupils). A back-of-the-envelope calculation is as follows: with a cost per primary school student of about 6,000 Euros, as it is in many European countries, and 90% of this cost corresponding to teacher wages, an additional teacher costs about 160,000 Euros. With 1,000 pupils in control classes (24 pupils) and 1,000 in treatment classes (12 pupils), this involves about 40 additional teachers, or 6,4 million Euros.[20] Other interventions are much less costly. An on-going evaluation about the returns to schooling awareness in London only involves the fixed cost of setting-up a web site to disseminate information. Likewise, thanks to its small sample size (250 students), the financial incentives experiment implemented at the University of Amsterdam[21] involved a maximum potential payment of 75,000 Euros, but given the actual pass rates (for which the incentive was paid), the cost of the operation was finally about 17,000 Euros. Notice that these kind of costs are not specific to RCTs: any experiment can be costly, however poorly evaluated. When the cost of a treatment is very high, then this is a strong argument in favour of planning a well-designed evaluation.

4.4.2. THE SPECIFIC COST OF THE EVALUATION

Costs specific to the evaluation include the cost of monitoring the protocol, which usually requires one part-time research assistant, and the cost of data collection, which is the largest element of cost. In many education policy experiments, data collection is relatively simple and costless: tests are administered in class by the teachers, subjects are not too difficult to track (as opposed to in labour policy experiments), and administrative data collection is taken over by the school or district administration. However, tests need to be prepared, sent, re-collected, marked and entered into computers. When possible, on-line tests can spare much of this cost. Overall, our experience with J-PAL education projects is that a survey cost of 8 Euros per pupil for a test score is a reasonable order of magnitude (provided this is not for face-to-face tests).

Administrative data are not necessarily readily available either. They must be collected in schools and it requires a lot of research assistant work to obtain answers from 100% of schools (which is the necessary goal). Also, one may need to survey parents, and this may not be efficient unless a telephone survey is run. 30 Euros per family would be a typical cost for an 8 minutes interview (which is quite long already). Overall, an experiment with 2,000 subjects, baseline and endline tests, a parental survey and administrative data collection may cost 140,000 Euros. When tests scores are sufficient for the evaluation, the cost is much lower.

Experimenting with policy and evaluating it can thus be a costly process (not to mention the cost it imposes on people and organizations), although this can vary substantially across projects. But this

[20] Krueger (1999) reports that the STAR experiment cost was US$ 12 million over four years (class size was reduced 4 years in a row) in the 1980's, which is a comparable order of magnitude.

[21] (Leuven, Oosterbeek, & van der Klaauw, 2010)

cost must be viewed as a high-return investment. A driver's license subsidy program implemented in France costs about 1,000 Euros per young beneficiary, and this program covers 10,000 youth, for a total cost of 10 million Euros. Many people think that driver licenses are important for youth labour market entry process, though this is not supported by any evidence. In this program, youth have been randomly allocated to a treatment and control group, and the whole evaluation costs about 300,000 Euros. This is a significant budget, but it may be worth paying 3% of the program cost *once*, so as to learn if it is wise to spend 10 million annually on such intervention.

In France, an Experimental Fund for the Youth was set up in 2009 with a budget of 230 million Euros, about 25% of it provided by private donors[22]. This fund pays for up to 50% of program costs and 100% of evaluation costs, provided that they are experimental and properly evaluated. This is a way to encourage program evaluation, the financial burden of which most project holders cannot face by themselves. It is also a way to generate knowledge about what works and what doesn't, and disseminate this information.

4.5. Relationships between the partners in an evaluation

Setting up a randomized evaluation is an institutional challenge. The evaluation may fail to get started, or it may collapse during the experiment. An RCT has costs and benefits, but they may not be properly perceived, and they may not be identical for the various partners. As a new feature in European countries, its costs tend to be overestimated and its benefits not well understood. In setting up an experiment, one should assess those two terms for each partner and try to optimize them.

An experimental project has potentially several types of stakeholders. The evaluator is an obvious one. Evaluators are typically academic researchers, consulting firms or public statistical services. On the implementation side, project implementers can be very different bodies. In education, typical project implementers are schools, school boards or the central administration. It depends very much on the national education system and on the nature of the policy. For instance, in the UK, it is possible to work directly with schools. Sandra McNally, from the London School of Economics, is evaluating a program that provides secondary school pupils with information on the returns to schooling. She has directly asked many schools in the London area if they would like to participate in the experiment, and she then randomized among volunteer schools. Similarly, a program examining the effect of financial rewards on student achievement was implemented within the University of Amsterdam, and evaluated by affiliated researchers (Leuven, Oosterbeek, & van der Klaauw, 2010).

In contrast, interventions that must be large-scale or involve some political decision require higher-level partners. This would obviously be the case for a class-size experiment. In some countries, schools would not enter any randomization program without formal approval of the school authority. In that case, there is a distinction to be made between the authority that starts the program and the schools that are the experimental groups. Their perspectives may differ.

[22] www.**experimentations**ociale.fr/

Finally, there may be other stakeholders. First, the financer may be different from the project managing authority. This would be the case for a program financed by the European Commission. It is also the case for projects financed by international funds in developing countries, or by the French Experimental Fund for the Youth (see 4.4.2). Second, an NGO may be the actual implementer of the policy.

4.5.1. Bringing stakeholders in: costs and benefits

In this section, we go through the objectives of each partner and discuss the difficulties at getting them to build and maintain a successful experiment together.

The evaluator usually wishes to maximize the methodological quality of the experiment. He therefore focuses on having a large sample sizes, a design that ensures high statistical precision, and a treatment or a set of treatments that are innovative and may produce new academic knowledge. But he should also be ready to adjust his strategy, given operational constraints in the field. Other stakeholders must be aware that adjustments are possible: as explained in the previous sections, there is a set of possible randomization designs that can adapt to various constraints. But they should also be aware that a serious evaluator will not engage in an RCT that has low statistical power, and such an undertaking would be a mistake for all stakeholders in any case.

Costs

Institutional partners face many costs. A fundamental one is to introduce change in the system, and the difficulty at raising interest among innovations. This is not specific to RCTs, and it is not clear at the outset whether an experimental innovation will be more easily accepted just because it is an occasion to generate knowledge. On the contrary, RCTs bring additional constraints as discussed below. In reality, it is often difficult to find even a few dozen schools ready to enter an experiment, and it requires a lot of effort from either the researcher or the school authority.

Certainly, it is easier to bring innovations that have long been debated, that are considered very likely to be adopted at some point or that do not raise strong opposition among unions or public opinion. Generally, one should expect that innovations would be more easily accepted in educational systems that favour autonomy and project-based education, although this will be less true in strongly centralized systems.

In the latter case, strong political support is required, and political risk must be assessed very carefully. At the Paris School of Economics, we started an experiment that involved offering entire high school classes a budget for a class-project that was based on actual pupils' presence in class. The program was abandoned because of a strong opposition in public opinion to the fact that participation in class should be encouraged by any kind of material incentive, even in an experiment. It is clear that in the first instance, starting such a project was a political decision that could only be taken at the government level. Fortunately, this project started with a very small pilot: if started at full scale, the project would probably have collapsed on its way and important resources would have been wasted.

In some cases, an entity - say a government - may like to test something without having any strong stake at the results. This is probably the case for the parental meetings intervention in France, which had no strong political implication but nevertheless was understood to require an RCT to rigorously determine the impacts. This would seem a sound way to build up policies, but unfortunately, we do not see such an approach as a frequent one. It requires a long-term research program, rather than a short-term political agenda. In other cases, as in the British experiment mentioned above, the interventions was set up at a researchers' initiative and it so happened that a sufficient number of schools felt that returns to schooling awareness was an interesting intervention. But programs built entirely by researchers' priorities would be rather limited in scope.

A randomized experiment is costly to the field-implementing partner, typically schools. The randomization design is a constraint to the operations: rather than inviting all parents to meetings, one has to convey information to one group only and face control group parents who would like to participate; rather than moving some kids to small classes (the most disruptive for instance), one has to follow the random list, and so on. Also, data collection is a burden to the schools: administrative data must be collected from staff, and teachers have to spare time for the standardized tests and surveys. Options to ruin the design or data collection, either voluntarily or by lack of involvement, are many. This is all the more difficult when the entity that wishes an experiment to take place is distinct from the organization that has to run it.

Randomization itself is perceived as a cost. It may be difficult to convince schools to volunteer for an experiment with only 50% chances to actually implement the program (even when the only alternative is that no one would have the program). At very least, a lot of information and justification should be provided to potential partners, and it is often necessary to use a phase-in designs where randomization is over the date the program starts. Even in that case, communication is critical. In a program run with a French NGO, schools were asked to volunteer for the program in June, given that only a randomized two-thirds would start on the first year. But the schools were not properly informed of the randomization, and not even of the two-thirds rule. All were ready to begin in September when they suddenly learnt about the design. This was a source of disappointment to them and several left the program, while others had to be convinced to remain, sometimes with much difficulty. Lack of initial information resulted from the fear that schools would not volunteer in the first place if they knew about the design: this was clearly a short-sighted strategy that came close to killing the whole experiment.

When randomization is to take place between classes or even pupils, the cost may sound even larger to the schools: they have to explain why some are not treated and to deal with frustrated parents. If the program is obviously rationed, a lottery may be justified, but this has to be explained thoroughly and may depend on national traditions. A decisive aspect here is that the schools feel they know which pupils would benefit best from the program. By hypothesis, if an RCT is run, one is not even sure that it benefits *anyone*, but the opposite conviction is often strongly grounded.

Benefits

The social benefit of a well-run RCT is generating knowledge. It is unclear whether contributing to such a public good is sufficient to compensate for the costs of engaging in an experiment, even in the eyes of such a highly educated population as teachers. The incentives each type of agent has to participate into the experiment must be carefully examined, especially in the case when different actors have different views.

The incentive to run an RCT may be intrinsic: some NGO or a school authority wants to demonstrate the efficiency of a program. A very clear example can be taken from the field of employment policies. In 2007, the French Public Service Employment (PSE) administration had to face competition from private firms at providing job seeker counselling. The unemployment insurance scheme, which is a separate entity, claimed that PSE inefficiency was costly, and that relying on private firms would increase efficiency. In this situation, the PSE strongly wanted an RCT of public- vs. private-provided counselling to be run, so as to demonstrate very clearly its larger cost-effectiveness (which it did![23]). In France, an NGO called Apféé provides 2 hour tutoring sessions to groups of 5 first grade pupils every evening after school, in order to help them at learning and reading. This is a very expensive intervention (about 1,000 Euros per year/child) partly financed by the State and partly by donors, and the actual efficiency of this scheme had been questioned. All partners, State, donors and the NGO itself, wanted a clear demonstration of the efficiency of the program (or lack thereof) to take place, and therefore agreed on an RCT. Actually, an evaluation of this schemed published a few years before concluded that it had no value added, but this finding was challenged because it was not RCT-based. Of course, in such a case, the evaluated institution has to accept fair risks, namely that the program will be found to be ineffective. In that perspective, experimenting programs that are politically challenged or whose effectiveness is quite unlikely, may be particularly risky, and such projects may never get started.

Another incentive is *conditionality*: public or private finance of a program, at least in its experimental form, can be conditioned on a rigorous evaluation. The French Experimental Fund for the Youth mentioned above, finances both the evaluation and part of the intervention itself, provided that a rigorous evaluation is run. In this specific case, this is an incentive not only to NGOs but also to public services, including government bodies, that may lack the budget for starting programs. It requires that such a fund is either a private entity or is politically autonomous, so it is capable of imposing constraining and potentially risky evaluations upon other organizations in exchange for its funds. To a certain extent, this would be the case of the European Commission. However, the implementing partners must view the evaluation as an opportunity, not as an obligation, and it would be naive to think that an RCT can be run with an unwilling partner.

Agreement with the evaluation project must be met at all levels. It is not sufficient that some school authority is strongly involved: as argued, much of the burden will be borne by schools or teachers, or by intermediary levels (school boards), and their involvement is needed. They may not feel directly concerned with the policy and its experiment, especially when they are in the control group. In very centralized systems, the authority may be able to impose much on the schools and school personnel. However, it is extremely difficult to ensure good design implementation or data collection with unwilling field partners Therefore, a lot of informational effort is required to explain the usefulness of the whole operation. This information should make very clear that one is evaluating the policy, not the people who implement it. Also, schools often appreciate getting their own results on tests back, or having further access to the testing instruments for their own future use. But it is fair to say that immediate or future (in case of phase-in design) benefit of a valuable program is the strongest encouragement.

[23] (Behaghel, Crepon, & Le Barbanchon, 2011)

In some cases also, a third partner may be present, which has limited interest in the evaluation. Again in case of the Apféé NGO, municipalities are involved in the administrative organization of the scheme, and some did not agree on the design. In some instances, some of the burden of the design and data collection falls on a third party that has no direct interest in the program or its evaluation. For example, specific social agencies (*Missions locales*) happen to direct young people towards a French young adults training program (*Ecoles de la deuxième chance*). An evaluation design planned to randomize individuals from the lists of those agencies: this put a burden on their daily work, whereas they had no direct interest in the evaluation of the training program. After several months of piloting, the evaluation had to be abandoned in large part for that reason.

4.5.2. Guidelines for setting up an RCT

Based on the above arguments, the main guidelines for setting up an experiment include the following:

Select a program. Relevant topics for an RCT can be both simple and costless interventions, where it is unclear they can have any effect at all, or interventions with strong potential impact, where their cost-effectiveness must be assessed carefully. They must be sufficiently well-accepted that no strong opposition to even testing them should be encountered. Programs that are known to have worked well in other countries or settings may raise interest, even if they are unusual in the national system considered.

Pilot the intervention. The program should be tested on a very small scale (in 4 classes, for instance) to ensure that it will not face implementation challenges when enlarged to a scale relevant for an RCT. This is also an occasion to test measurement instruments, the anonymity process, etc. This is not always possible due to political agenda but should be encouraged as much as possible.

Define the randomization design. It must be defined between the evaluation team and all the implementing partners. They must agree on a design that will optimize on the constraints and objectives of the partners. As seen above, there are many options. A pilot is helpful in figuring out the operational constraints and it can bring data that is useful to the power calculations. Sometimes, the financing institution must define the general principles of the randomization design and only then an evaluator is recruited on a tender, to implement those principles. In that case, the financing institution must rely on experts that are familiar with RCTs, and it must also anticipate that the recruited evaluator will have to suggest some fine-tuning on the design (partners must be ready to that).

Inform all partners. Information on the RCT must be given far in advance. Many partners will understand neither the usefulness nor the implications of the project: it has to be explained very carefully. When randomization is among volunteers, they must be identified in advance and be made aware of their odds to enter the program and at what time. When randomization is not among a volunteer population, the implications and the ethics of the program must be made clear so as to maximise actual take-up.

Implement the design. The evaluator is responsible for ensuring that the design is properly implemented (treated are effectively treated for instance, see 4.3.1) and that data is exhaustively collected; this requires a lot of fieldwork. The evaluator usually has no power over agents so the relevant authority should be ready to intervene. Unbalanced (between treatment and control) data collection, for instance, can ruin the whole project, and so much care (and political support) should be given to certain decisive dimensions.

Publicize the results. When the results of the evaluation are available, they should be widely shared with all partners and the general public. It is particularly important that agents who have supported the burden of implementation should be kept informed and be able to understand what their efforts have produced. A meeting with their representatives would be recommended. When the results are positive, the option for scaling-up must be considered carefully, especially in view of the external validity issues: the RCT may have answered some of the important questions but not all of them. Replication may be warranted. In some cases, further experiments are needed to understand failures or unexpected results: all partners must be ready to accept that building knowledge is not always a one-shot thing and may take several years.

5. Randomized control trial in the context of the European Union

5.1. RCTs implemented in Europe

Few RCTs have already been implemented in Europe, especially in the domain of education. As a general statement, in developed countries, the number of large-scale evaluations of education projects remains limited, and the vast majority of empirical evidence available in education today derives from non-experimental evaluations. In recent years, however, an increasing number of research projects have used randomized design in Europe. Nonetheless, the number of critical results that allow building strong evidence on the efficiency of education policies in European countries has not yet been reached. An expansion of the RCTs is thus needed. The following table is the result of our investigation about RCTs in Europe. It summarizes the randomized evaluations that have been brought to our knowledge[24] on the topic of education.

This obviously excludes RCTs more related to cognitive science and it concentrates on *policy* experiments. One of those experiments (Teaching methods for reading, France 2011) has tested methods that are expected to be very efficient based on many small-scale experiments from cognitive science. In a large-scale environment (40 schools), where precise control of a year-long teacher intervention is not easy, it found no effect at all. Additional interventions are ongoing to understand what field conditions make implementation of these methods efficient or not. This is an interesting illustration of the specifics of policy experiments.

[24] By no means do we pretend that this constitutes a comprehensive list of randomized experiences of education policies in Europe. This is based on information gathered from members of the EENEE network. Budget information is available only for projects we have been involved in. They refer to the evaluation part only.

Title	Place	Randomization Strategy	Objectives		Results	Budget
Getting Parents Involved[25]	Suburban Paris (France) (2007)	The sample is composed of 37 schools. Within each school, grade 6 was randomized (5000 students). Only the volunteer families were accessible for the	Evaluate the effect of a policy aiming at stimulating parental involvement.	completed	Results in term of parental involvement, students' behaviours and school results were found. Peer effects are typically large.	220 k € over one year
Teacher training in Didactics and Classroom Management on Student Learning Outcome[26]	Denmark (2011)	26 schools, 60 classrooms are assigned to two treatment arms: one offering didactics training to the teacher the other class management training.	Evaluate whether a training package offered to first grade teachers improve non cognitive and cognitive skills.	On going	.	
Information about return to education[27]	London (2010)	54 schools in total. Students in the treatment arm benefit from an information campaign on the return of education (website, video and school materials).	The goal is to modify the perceived return to education and to evaluate whether such change translate into higher investment in education	On going	.	
Classroom training for unemployed and employed workers[28]	Denmark (1994)	Individual randomization. Sample size 118,000 unemployed and employed workers	Effectiveness of an individual traineeship on probability of employment	Completed	Surprising negative effect of the traineeship on both employment rate and unemployment	

[25] (Avvisati, Gurgand, Guyon, & Maurin, 2010)

[26] Contact researchers: Anders Holm (Aarhus University)

[27] Contact researcher: Sandra Mc Nally (LSE)

[28] (Rosholm & Lars, 2009)

Title	Place	Randomization Strategy	Objectives	status	Results	Budget
Education voucher for adults[29]	Switzerland (2005)	2437 individuals in the treatment branch, 17234 in the control.	Do lifelong learning increase employment, wages and subsequent education of adults?	Completed	No effect of the voucher program on employability or wage. Take-up is low among who potentially benefit the most from the treatment.	
Financial reward at university[30]	The Netherlands (2001)	249 students from the economics department of the University of Amsterdam	Do passing rate increase if students are financially rewarded?	Completed	No significant effects are found.	
Remedial education for first graders	France (2010)	5000 students from 120 schools in suburban Paris. Only the fifth lowest achieving first graders are eligible to the program in the treatment schools.	Efficiency of a remedial education program for first graders on reading and writing skills	On going	.	350 k € over two years
Teaching methods for reading	France (2011)	80 schools, half randomized into the program.	Test methods based on cognitive science literature that emphasize phonics and phonemics and group-level work.	Completed (report on going)	No effect in spite of high predictions from cognitive literature. Illustrates differences between lab and field experiment.	

[29] (Schwerdt, Messer, Woessmann, & Wolter, 2011)
[30] (Leuven, Oosterbeek, & van der Klaauw, 2010)

Title	Place	Randomization Strategy	Objectives	status	Results	Budget
Boarding school of Excellence	Suburban Paris (2009)	445 applicants to the boarding schools were randomly assigned.	The goal is to test the efficiency in term of cognitive, non cognitive and school career of boarding schools	On going	.	460 k € over three years
Children and Parental information on school choices	Suburban Paris (2010)	37 schools in the sample were randomly assigned to receive the treatment. In total 1000 students were eligible.	The goal is to evaluate the efficiency of a school career information campaign on school choices	On going	.	420 k € over two years
Mentoring at high school	Suburban Paris (2008)	22 schools compose the sample. Eligible students in treatment schools are offered a mentoring program.	Is mentoring a proper way to change school choices and investment in higher education? Improve school involvement and Behaviour.	On going	.	390 k€ over three years
Mentoring program for underprivileged graduates	Parisian Universities. (2010)	Individual randomization of 600 students. Students belonging to the treatment arm benefit from a mentoring program.	Impact of mentoring on social capital and professional integration.	On going	.	200 k€ over two years

5.2. THE TRANSNATIONAL EXPERIMENT

The above tables show that it is realistic to run randomized experiments in order to evaluate education policies in Europe. It also shows that much more could be done, and that there is no coordination taking place so far. Why would a Europe-wide approach be relevant to such a program? We believe one can imagine two approaches.

The first one aims at testing a similar program in different countries simultaneously and in a coordinated way. The interest of that approach is to overcome the external validity issue. An RCT run in one country has strong internal consistency: the causal impact is demonstrated for that country and the context of the experiment. But it may not be valid in other contexts. As discussed already, one way to overcome this issue is replication. Such a replication process is ongoing worldwide. IPA (Innovations for Poverty Action), an NGO closely related to the J-PAL global network, is responsible for more than 300 evaluations worldwide, many of them replications of successful J-PAL programs. By running a similar program in different countries, one may learn more about the robustness of a program and better understand the elements of context that make it more or less effective.

In this model, the European Commission could identify interventions that are potentially relevant for a number of countries and publish a call for projects with detailed evaluation guidelines. Countries would submit their project with an evaluation protocol that would imply some form of randomization within the country itself. Several evaluations of the program would then be available and could be compared. One interesting aspect is that the decentralized research teams could have a common practice in terms of data collection. Agreement on the instruments used to capture the outcomes of the experiments (questionnaires, cognitive test, non-cognitive tests) is necessary. The experience of PISA has shown that it is possible, although not necessarily easy, to measure learning skills in various countries.

To sum up, this approach would require: several truly motivated schooling administrations in various countries; a topic that is equally appealing for each local context; a centralized research team responsible for creating standard instruments, enforcing a common protocol and ensuring that the highest standards of analysis are carried out in each country; decentralized research teams in each country responsible for the local implementation of the program. The European Union could be an interesting actor for promoting comparable experiments in various countries in Europe. Being able to rely on one central entity – the European Union – would be an interesting way to standardize a program.

We must stress however the possible difficulties that such endeavour would entail. Member States may not have the capacity or sufficient experience to set up an experiment, and coordination of programs and evaluation protocols may be very costly. At any rate, this would be realistic only with a very small number of countries altogether.

Also, the topic of the evaluation must be equally appealing in each European country. That certainly limits the number of eligible programs. Programs that depend excessively on the structure of the education system would be hardly evaluable at European level. For instance, teacher recruitment processes would be difficult to evaluate: in some countries schools are not in charge of hiring teachers. Conversely, programs promoting teaching technique (class management, cognitive learning methods…), providing new teaching tools (computer, internet) or offering specific service or benefit (Early childhood intervention, Conditional cash transfer, linguistic or cultural trips) might be more suitable for coordinated Europe-wide evaluations.

5.3. The on-going dissemination process

A different approach is one where the European Union, rather than pushing for a specific program in a coordinated way, would identify topics that have proved interesting vectors for educational reform or innovation, and help disseminate results by promoting RCT-based replications in other European countries. This would be somewhat in the spirit of the Open Method of Coordination. The Mexican PROGRESA experiment gives an illustration of such a process: it has demonstrated that conditional cash transfer programs are effective. Based on this observation, many developing countries have considered such a program for themselves, though not without testing it in their context first. This is an ongoing process that has taken time and, in this case, was not centralized (although the World Bank certainly encouraged some experiments to take place). Likewise, an RCT of a parental involvement program that we have mentioned already has proved very efficient at reducing truancy and improving student behaviour in France, with strong peer effects. This program is becoming well-known and is now considered for replication both in Chile and in South Africa, using a randomized methodology. It would certainly make even more sense to adapt and test it in other European countries.

According to that model, the European Commission could either identify an existing national program that has the potential to tackle issues that are a concern in several Member States; or promote the initial evaluation of an original relevant program in one country. It would then encourage other States to test this program with an RCT. The projects would be built one by one, possibly with the technical assistance of previous experimenters. The difference with the previous approach is that there is no trying simultaneous implementation of a program; but the dissemination process takes more time and learning is slower.

This approach would raise far fewer coordination problems. Also, it addresses the fact that replicated programs are not exact copies of the original programs; instead replications may be adapted to the political and social context in which they are implemented, sometimes with significant adjustments. Moreover, researchers could improve their understanding of the workings of the programs by testing new hypotheses, which can only occur when an intervention is tested recursively. Finally, acceptability of innovations to school authority and schools themselves would be stronger, provided that local actors are aware of the progress made in other European countries, and can be inspired by other experiences.

Our experience is that national school systems can be quite unaware of policies implemented elsewhere. The European Union, as well as international organizations like OECD, is keen to spread information and encourage adaptation. The promotion of RCTs seems a natural way to support such a process. Being able to show that one specific program has interesting results in one European country may induce other countries to replicate the same program, even though this would not have been possible in the first place. Some country may be a leader in one specific type of education program and thus encourage others to enact the same types of policies. We thus think that implementing evaluation at the European level may be a very strong tool to instil change in each country’s educational systems.

6. Conclusion: the case for EU intervention

This report shows that RCTs have been used worldwide to test programs in education. They have generated fundamental pieces of knowledge. For instance, the STAR experiment (US) has confirmed that class-size reduction has the potential for improving achievement; the Perry pre-school project (US) has proved that intensive early childhood intervention can raise very poor children's outcomes, even until adulthood; the PROGRESA experiment (Mexico) has shown that incentives are effective at raising school attendance.

But none of those findings is directly applicable to European countries, or even to any country in particular. They point to general principles that require further testing and adaptation. Furthermore, additional innovations should be tried (and existing practice should also be evaluated), based on European specifics and priorities.

Educational choices usually have huge budgetary consequences, and it would be reasonable to invest even significant resources to rigorously testing them with RCTs. As with any knowledge building, externalities are strong, so it is reasonable to avoid paying the same fixed cost of learning several times. Therefore, it would be efficient for the European Union to encourage and coordinate experimental programs: it would maximize both spending and knowledge spillovers. This would require funding but also support from a network of experts and practitioners, so as to identify and promote programs to be tested, and counsel Member States on appropriate testing methods.

This report has not shied away from the limitations of RCTs, and the difficulties of setting up and running them. But experience proves that these difficulties can be overcome. It requires time, determination, and a lot of information for decision makers and schools to make them incorporate innovations. In themselves, RCTs are strong arguments in favour of change.

Bibliography

Angrist, J., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quaterly Journal of Economics , 114* (2).

Avvisati, F., Gurgand, M., Guyon, N., & Maurin, E. (2010). Getting Parents Involved:A Field Experiment in Deprived Schools. *CEPR working paper 8020* .

Baird, S., McIntosh, C., & Özler, B. (2011). Cash or condition? Evidence from a cash transfer experiment. *The Quarterly Journal of Economics , 126* (4).

Chetty, R., Friedman, J. N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings: Evidence from project STAR. *NBER Working Paper No. 16381* .

Crépon, Duflo, Gurgand, Rathelot, & Zamora. (2007). Counseling and Job Placement for Young Graduate Job Seekers in France. *working paper* .

Duflo, E., Dupas, P., & Kremer, M. (2009). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *NBER Working Paper No. 14475* .

Duflo, E., Glennerster, R., & Kremer, M. (2006). Using Randomization in Development Economics Research: A Toolkit.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, a. A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative economics , 1* (1), 1-46.

Heckman, j., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and non cognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics , 24* (3), 411-482.

Krueger, A., & Whitmore, D. (2001). The Effect of Attending a Small Class in the Early Grade on College Test Taking and Middle School Test Results: Evidence from Project STAR. *The Economic Journal , 111* (468).

Leuven, E., Oosterbeek, H., & van der Klaauw, B. (2010). The Effect of Financial Rewards on Student's achievement: Evidence from a Randomized experiment. *Journal of the European Economic Association,* (6), 1243-1265.

Rosholm, M., & Lars, S. (2009). Is labour market training a cruse for the unemployed? Evidence from a social experiment. *Journal of applied econometrics , 17*, 338-365.

Schlotter, M., Schwerdt, G., & Woessmann, L. (2009). Methods for Causal Evaluation of Education Policies and Practices: An Econometric Toolbox. *EENEE Analytical Report No. 5* .

Schultz, T. P. (2004). School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics , 74*, 199-250.

Schwerdt, Messer, Woessmann, & Wolter. (2011). Effects of Adult Education Vouchers on the Labor Market: Evidence from a Randomized Field Experiment. *IZA DP No. 5431* .

EENEE Analytical Reports

11	Adrien Bouguen Marc Gurgand	Randomized Controlled Experiments in Education
10	Torberg Falch Hessel Oosterbeek	Financing lifelong learning: Funding mechanisms in education and training
9	Reinhilde Veugelers	A Policy Agenda for Improving Access to Higher Education in the EU
8	Giorgio Brunello Martin Schlotter	Non Cognitive Skills and Personality Traits: Labour Market Relevance and their Development in E&T Systems
7	Eric A. Hanushek Ludger Woessmann	The Cost of Low Educational Achievement in the European Union
6	George Psacharopoulos Martin Schlotter	Skills for Employability, Economic Growth and Innovation: Monitoring the Relevance of Education and Training Systems
5	Martin Schlotter Guido Schwerdt Ludger Woessmann	Methods for Causal Evaluation of Education Policies and Practices: An Econometric Toolbox
4	Martin Schlotter	Origins and Consequences of Changes in Labour Market Skill Needs
3	Martin Schlotter Guido Schwerdt Ludger Woessmann	The Future of European Education and Training Systems: Key Challenges and their Implications
2	George Psacharopoulos	The Costs of School Failure – A Feasibility Study
1	Ludger Woessmann Gabriela Schuetz	Efficiency and Equity in European Education and Training Systems