Designing economic experiments for evaluation purposes

Marc Willinger
Faculté d'Economie, LAMETA, Université de Montpellier 1, France

Recebido: 28/04/2011 Aprovado: 22/03/2012

ABSTRACT
The growing popularity of experimentation in economics has widened the scope for economic experiments. In this paper we question the relevance of experimental methods for economic impact assessment. The major issue of impact evaluation is answering a counterfactual question. We show that the economists' experimental toolbox can provide the appropriate method to give the right answer, especially by relying on randomized field experiments (RFEs). We contrast RFEs to other types of experiments, and discuss the limitations of experiments for evaluation purposes, by presenting 3 case studies that relied on economic experiments at the individual, local and national levels.

KEYWORDS | Experimental economics; Randomized field experiments; Economic methodology.

JEL-Codes | C9; C93; H0.

* I would like to thank two anonymous referees for useful comments that helped to improve the paper.

Desenho de experimentos econômicos para fins de avaliação

RESUMO
A crescente popularidade da experimentação na economia ampliou o espaço para experimentos econômicos. Neste trabalho questionamos a relevância dos métodos experimentais para a avaliação de impactos econômicos. A principal finalidade da avaliação de impactos é responder a uma pergunta contrafactual. Mostramos que a caixa de ferramentas experimental do economista pode fornecer o método apropriado para dar a resposta certa, sobretudo a partir do uso de experimentos de campo randomizados (randomized field experiments, RFE). Contrastamos os RFEs com outros tipos de experimento e discutimos as limitações dos experimentos para fins de avaliação, apresentando três estudos de caso que fizeram uso de experimentos econômicos nos níveis individual, local e nacional.

PALAVRAS-CHAVE | Economia experimental; Experimentos randomizados; Metodologia econômica.

CÓDIGOS JEL | C9; C93; H0.

Introduction

Economists increasingly rely on experiments to investigate various aspects of their discipline. While early experiments were tailored for the purpose of theory testing, more recent experiments were designed for evaluating new institutions and markets, testing alternative policy instruments and exploring new areas.

In this paper we question the relevance of experimental methods for economic evaluation. Evaluation covers a large spectrum of activities, including policies, institutions, professionals (researchers, doctors etc.), and impact assessment of social programs or science and technology programs, among others. Over the last decades the scope of evaluation has been broadened widely and has often been included as a requirement for various programs sponsored by public and private institutions.

There is a need for evaluation both before and after the implementation of a new program. Because resources are scarce, ex ante prospective evaluation is needed in order to enlighten the choice between competing programs. Ideally, for a given target and a given budget, the best program should be selected. Ex post, it is important to know if a selected program has reached the expected targets or not, and identify the reasons for success or failure.

Experiments are useful both for guiding the selection of projects and for evaluating the performance of the selected projects. It is useful to distinguish between *process evaluation* and *impact evaluation*. Process evaluation refers to how the program is executed and whether the procedures have been correctly implemented. Impact evaluation is the assessment of the impact of the program, i.e. how close the program's outcomes are to the expected targets, or how far they are from the state that prevailed before the program was implemented.

This paper is mainly devoted to impact evaluation and discusses how experiments can be designed towards such a goal. The major issue of impact evaluation is answering a counterfactual question. Consider a public program designed to impact some characteristic of a population of agents that could be firms, consumers, citizens, etc. The key question is how the agents who were exposed to the program would have done without the program and how the agents who were not exposed to the program would have done with the program. Of course, we do not know the answers to these questions. Why not simply compare the situation of the agents who were exposed to the program, before and after their exposure? The reason is that this is relevant only if nothing else has changed except the impact of the program. Since there is no reason to believe that nothing else has changed, it is necessary to compare agents who were exposed to the program to similar agents who were not exposed to the program. The most relevant method for doing that is a "randomized field experiment" (RFE), also sometimes called a "randomized control trial". In an RFE, some randomly selected agents are exposed to the program while others are not. Random assignment guarantees that the distribution of the agents' unobservable characteristics is the same in both sub-samples. If the two sub-samples are independent, impact evaluation of the program is obtained by measuring the difference between exposed and non-exposed agents.

An early example of this method, reported in Levitt and List (2009), is illustrated by a vaccine discovered by the famous French physician Pasteur, who was one of the first researchers to rely on a randomized field experiment in medical science.

In 1882 Pasteur proved the immunity property of his new vaccine (anthrax) by publicly running a field experiment: 25 randomly selected sheep were vaccinated (the test group), while 25 other sheep were not vaccinated (the control group). After two days the 25 sheep of the control group were dead while the 25 sheep from the test group were alive. This example shows the importance of a control group for evaluating the impact of a new drug or medicine, and the importance of randomly assigning the agents to one of the two groups: test or control. After the early days of randomized field experimentation, field experiments were implemented on a very large scale in agriculture, with the influential work by Fisher (1935).

Today randomized field experiments are often considered the “gold standard” of impact evaluation, in particular for clinical application and evaluation, and for alternative therapies or pharmaceuticals. The method has spread to other disciplines, in particular to agronomy, social sciences and more recently to economics. The Organization for Economic Cooperation and Development (OECD) defines as the “gold standard” for a scientific evaluation “a randomized experiment in which the control treatment group is randomly selected to participate in a program and nonparticipants are randomly assigned to the control group”. This approach has counterfactual analysis at its core.

While RFEs are not always feasible, one should rely on such methods as frequently as possible, because they offer the best way to construct a counterfactual. In this paper we show the advantage of using RFEs compared to other methods for running economic experiments. We also illustrate more generally how a researcher who needs to carry out an evaluation (or ranking) of alternative policy programs can improve his understanding of the issues by relying on experiments. Finally, we discuss the limitations of experimental methods for evaluation purposes.

The next section describes and defines economic experiments. Section 3 discusses the advantages of RFEs compared to other types of experiments. Section 4 presents several examples of experiment-based evaluations. Finally section 5 concludes with a discussion on the limitations of experiments for evaluation.

2. Types of economic experiment

This section defines various types of economic experiment and compares them with other empirical methods, such as quasi-experimentation or field studies. Empirical methods can be evaluated according to two key dimensions: control and validity.

2.1 Objectives of economic experiments

Roth (1995) identified three types of economic experiment according to the audience to whom they speak. Experiments that “speak to theorists” are intended to test theory and research hypotheses. Experiments “searching for facts” explore new fields and issues not covered by conventional theory, and speak therefore to economists and peers. Finally, experiments that “whisper in the ears of princes” aim at advising policymakers and decision makers about possible impacts of their decisions. Very often, experiments convey several messages, both to theorists and decision makers. Therefore the list of goals can be refined. Smith (1994) provides a comprehensive list of 7 reasons for economists to run experiments.

Experiments are potentially useful for economic evaluation, although economists do not traditionally rely on them compared to other social scientists who introduced large-scale social experiments as early as the sixties. There exists a very large spectrum of methods for collecting data that can be applied within an evaluation task. At one end of this spectrum the researcher collects naturally occurring field data simply by observing facts and describing things. This includes a huge variety of data such as commodity prices, unemployment rates, stock market data, livestock, energy production, etc. At the other extreme of this spectrum are laboratory experiments, which produce data in an artificial and controlled environment with the aim of establishing causality. Between these two extremes, there are many other means of collecting data, such as randomized field experiments, natural experiments, and quasi-experiments, among others. Each of these methods of data collection has its advantages and disadvantages.

However, in contrast to experiments in physics or chemistry, where individuals behave identically (a given atom behaves just like any other atom), in life sciences and social sciences each individual is different, and therefore might react differently to a given stimulus. Since it is impossible to control for all differences between individuals, even when all characteristics can be identified, the method usually entails comparing a reference group to a test group.

2.2 Internal and external validity

Internal validity refers to the ability to establish causality based on observed correlations between facts. External validity refers to the ability to generalize the relationships found in an experiment outside the lab (e.g. to other persons, times and settings). By choosing a method of evaluation the researcher is confronted with a tradeoff between internal and external validity (ROE; JUST, 2009), which is illustrated in Figure 1.

Laboratory experiments have strong internal validity but there is a cost in terms of external validity. Control is high and causality can often be clearly established, due to the strong intervention of the researcher. Furthermore, lab experiments are replicable. But many lab experiments tend to abstract away from the context in order to reach such control. Very often they involve student subjects whose characteristics may systematically differ from those of interest in the field.

Gathering naturally occurring data provides strong external validity, but has strong limitations: replicability is often not feasible, control and treatment groups may involve systematic differences due to lack of control, causality cannot be clearly established and may be spurious because uncontrolled variables affect the control and/or the treatment group.

FIGURE 1

Tradeoffs across research methodologies according to Roe and Just (2009)

	Relative Internal Validity	Relative External Validity	Topic and Subject Limits	Replicable?
Lab Experiments	High	Low	Long duration topics, larger stakes, losses	High
Field Experiments	Medium to High	Medium to High	Limited by researcher connections	Low to medium
Natural Experiments	Medium to High	High	Limited by occurrences of nature and policy	Low
Field/market Data	Low	High	Limited by privacy, recall and trade secrets	Low to medium

Field experiments offer an attractive compromise. Most of the context is not under the control of the researcher compared to lab experiments, but the researcher can exogenously manipulate some of the dimensions and observe their impact. Consequently, the data generated by a field experiment reflects both external validity and internal validity. Field experiments therefore represent a bridge between the lab and the field (LIST, 2006). Other methods also compromise between internal and external validity, such as natural experiments. In a natural experiment manipulation the researcher takes advantage of a change that occurs without his or her intervention, in order to measure the impact of the change. This is similar to an experimental design involving a control group and a test group, but without the possibility for the researcher to manipulate the context.

While the tradeoff view provides an interesting classification of methods, Harrison (2004) argues that the researcher should be concerned with external and internal validity for any type of experiment. In his view, fundamentally there is no trade-off since no meaningful inference can be made from the data without a theory.

2.3 A taxonomy of (economic) experiments

In a recent paper, Carpenter et al. (2004) proposed a taxonomy of economic experiments that is useful both for researchers and decision makers. Their classification is based on 5 criteria: (i) the nature of the subject pool, (ii) the nature of the information and experience that the subjects bring to the task, (iii) the nature of the commodity, (iv) the nature of the task or institutional rules applied, and (v) the nature of the environment that the subjects operate in. Based on these ingredients they identify the four types of experiments detailed below:

- Conventional lab experiments *that rely on a standard subject pool (e.g. students), abstract away from any context and impose a set of rules on the participants.*
- Artefactual field experiments, *similar to conventional lab experiments but participants are non-standard subjects.*
- Framed field experiments, *similar to artefactual field experiments but involve a field context, task or information on which subjects rely.*
- Natural field experiments, *similar to framed field experiments but the environment is one where the participants naturally undertake the tasks and do not know that they are involved in an experiment.*

Besides these experiments that are considered within the taxonomy of Harrison and List (2004), there are three other types that can be relevant for evaluation: *social experiments* involving government commitment to a policy program to evaluate the impact of the program with respect to some benchmark condition; *natural experiments* involving changes or treatments not controlled by the researcher (in contrast to natural field experiments) and subjects who are unaware that they are in an experiment; and *thought experiments* representing an intellectual exercise with no implementation.

3. The "gold standard": randomized field experiments

3.1 Defining randomized field experiments

There is an important distinction between a field study and a field experiment. In a field study the researcher collects data that already exists. The difficulty is therefore to access the data and to observe it or collect it carefully without altering it. In contrast, in a field experiment the researcher participates in data production in the sense that he participates in the creation of data that does not yet exist. For example, he introduces a new opportunity for individuals to make transactions, and thereby observes the effect of this opportunity by comparing the induced changes with respect to a sample that is not exposed to this new opportunity.

A randomized field experiment involves one or several treatment groups, which are exposed to a program, and a reference group, which is not exposed to the program and to which the treatment groups are compared. We limit the discussion to the case of a single treatment group. The reference group serves as a benchmark and is supposed to behave as the agents exposed to the program would have behaved without such exposure. Thus the reference group reflects what would have happened if the exposed agents had been unexposed. By comparing the test group to the reference group, it is possible to evaluate the impact of the program since the only difference between the treatment and reference groups is the program if the assignment of agents to one of the two groups is made on a random basis within the studied population. As an illustration, consider the case of a new drug that is tested on a target population suffering from a disease (e.g. headaches). The target population is randomly split into two groups: the T group (Test) and the R group (Reference). In this example the program consists of administering the new drug to the T group for a given period, and administering a placebo to the R group for the same period. Comparison of the frequencies of headaches between the two groups guarantees that the difference is due only to the drug.

The key is randomization to prevent the selection bias that occurs when the test group is composed of agents selected because of particular characteristics. For example, participants in social programs are often volunteers and therefore exhibit differential characteristics compared to the general population (e.g. they are more educated, better informed, have more free time to spend, etc). The presence of different characteristics in the control and test groups means that the researcher might mistakenly attribute causality when there is none, or overestimate/underestimate the impact of the program.

An RFE establishes causality by creating a counterfactual (the control group), asking what would have been the outcome if the program had not existed. The method for this consists of randomly assigning units of observation (individuals, firms, villages, regions, schools, crops, …) either to the treatment group or to the control group. If this is done randomly and with a large enough sample, the control and treatment groups are identical, with respect to both observable and unobservable characteristics except treatment. RFEs provide rigorous evaluation, are often replicable and suggest ways for improvement for future studies. They are not always feasible, however. It is important to ask the evaluation question first, and then to find out what is the best method for answering.

Basic requirements for running an RFE include the possibility of defining separable units (individuals, households, villages, geographical areas etc.), collecting enough units to run meaningful statistical tests, and measuring spillover effects on control units, where this is feasible. Of course, nationwide programs (such as anti-corruption) do not satisfy these requirements and therefore must be evaluated using other methods. Quantitative measurement is not a requirement contrary to common sense. Qualitative data or categorical data is frequently used and provides satisfactory indicators for impact evaluation.

3.2 Development of RFEs

According to Levitt and List (2009), the development of field experiments has followed three major waves. The first was in the 1920s, when field experiments emerged in agricultural studies. This period was then followed after World War II by large-scale social experiments starting in Europe and spreading to the US in the late sixties. The most recent wave flows into economics, where field experimentation is largely used in naturally occurring environments and participants often ignore that they are involved in an experiment. A clear goal of these experiments is testing theory and exploring new areas of knowledge.

Social experiments provide a good example of what could happen in a near future with experiments designed to study economic issues on a large scale. Historically, large-scale social experiments were first developed in the UK to evaluate electricity pricing schemes (one study started in 1966 and lasted 6 years). In the US, large-scale social experiments were implemented in the late sixties, following the academic and political debate on the US welfare system after the publication of the Coleman report. Early experiments were targeted towards negative income taxation, an idea that became popular after Milton Friedman published his book *Capitalism and Freedom* (1962). A large-scale program involving 1300 households was set up for a period of three years. Individuals assigned to the program were guaranteed a minimum income and any dollar earned in employment reduced their subsidy by 50% of that dollar earned. A broad debate occurred about the work incentives provided by such a program, both before and after the program, reaching a consensus about the negative effect on labor supply compared to the preexisting welfare system. After this initial experiment in the US more than 235 social experiments were conducted.

4. How can field experiments be used for economic evaluation? Three examples

This section illustrates the usefulness of economic experiments for evaluation purposes by reviewing three examples: the allocation of irrigation quotas at the regional level in the US, incentives to increase crop yield at the state level in Africa, and incentives to increase productivity at the firm level in UK. Each study illustrates the applicability of experimental evaluation at some level: the individual, state or national level.

4.1 Allocating irrigation quotas

The efficient allocation of water quotas is becoming increasingly imperious in a context of global warming, especially in arid and semi-arid regions, but more generally in any region which relies heavily on irrigation. In many places, market tools are used in place of grandfathering allocation of water rights or arbitrary sharing rules (Australia, US). In 2000, the US state of Georgia passed a law requiring the use of an auction-like mechanism to allocate irrigation rights in periods of drought (the Flint River Drought Protection Act). Economists generally consider auctions efficient tools for allocation purposes. However, laboratory experiments showed that depending on the type of auction, misallocations may arise. In the case of the Flint River Drought Protection Act, the law did not specify which particular auction rule should be applied.

In response, Georgia State asked experimentalists for advice. Cummings et al. (2004) conducted a series of laboratory and field experiments to test a variety of auction procedures.

As in other countries, based on weather forecasts, the authority announces early in the year whether the current year will be a drought year or not. In the specific case of Georgia, this is announced by the Environmental Protection Division (EPD) on March 1. In a drought year the administration next decides how many acres to take out of irrigation to save enough water for other activities. Water use permits are then auctioned off by farmers who volunteer to withdraw their land from irrigation, and the auction winners receive monetary compensation. The market involves only one buyer with a limited budget (EPD) and many sellers who use their water rights for crops (corn, cotton, peanuts). Sellers are owners of land whose irrigation contains both private and common value components.

The objective of the study was to design an auction that provided farmers with an incentive to reveal the true cost of foregoing irrigation. Conventional lab experiments were initially run with student subjects to calibrate the experimental design. Adults and farmers were involved in the field experiments.

Farmers involved in the auction were sellers owning multiple units. Units to be auctioned were irrigation permits and each seller owned several permits with differing values. Two types of auctions were compared in the experiment: a uniform price auction and a discriminatory auction (sealed-bid). The uniform price auction is the simplest mechanism and the easiest to implement for the regulating authority. The mechanism is based on a cutoff price. All offers above the cutoff are rejected while those below that price are accepted. The cutoff price is determined as the auction's closing price and can be defined either as the smallest rejected price or the highest rejected price. Only the lowest rejected price is incentive compatible: under the highest rejected price rule the possibility of strategic offers cannot be excluded. For example, if a seller has two units of differing values, it might be that his second unit can be sold just at the market clearing price, and therefore he has an incentive to announce an offer that is higher than his true cost.

Three important findings from this study led to the selection of an auction procedure. The first finding from initial test experiments was the observation of a high inefficiency-cost when auctions are run one-shot. Efficiency is achieved if the lowest value permits are sold. In the test experiments, on the other hand, many high valued permits were sold. However, by allowing repetition the misallocation disappeared when revision was possible after announcing the winning offers. In other words, sellers' regret disappeared when market participants had the opportunity to revise their initial offers. Second, the finding from lab experiments was that the offer-to-value ratio was higher than 1 for both types of auctions and tended to increase across revision rounds. Announcing the highest accepted offer in the discriminatory auction clearly tended to increase offers over rounds (even for low valued items). Finally the field experiments[1] clearly showed that participants made explicit attempts to collude, but were unsuccessful. This was an important observation, since collusion among farmers could not be excluded in the field setting. Second, moving to the field was an important target for scaling reasons. Lab experiments involved only 9 participants, while field sessions involved up to 42 participants at once in a discriminatory auction.

On the basis of these results taken together, the authors of the study recommended a discriminative price auction with revisions, no maximum accepted price announcement, and a random tie-breaking rule. The Georgia EPD implemented the recommended procedures.

4.2 Increasing crop yield

The use of fertilizer in agriculture is an important driver of development for most developing countries, as recently documented in many Asian countries. On the other hand, in regions such as Africa where the use of fertilizer has stagnated over long periods, growth and development are cumbersome. Fertilizer increases crop yields and thereby favors growth and development. Regions where fertilizer is sparsely used remain poor and impede the development of activities beyond the agricultural sector.

Many observers have therefore argued in favor of major programs to subsidize agriculture, and especially the use of fertilizer, in these regions (2% of the government budget in Zambia is spent on fertilizer). However, the traditional economic argument against subsidies is that they distort relative prices and generate misallocations of resources. Undesirable effects of subsidies therefore include input overuse and environmentally unsound practices; at a certain level, they may even become counterproductive.

[1] Field experiments involved mixed groups of subjects, including farmers. Each participant received two vouchers (permits) with a face value. If the voucher was unsold the participant received the face value in cash, and if it was sold he earned the price offered in the auction.

Duflo et al. (2008) implemented tests with farmers in Western Kenya where fertilizer use was low on the farms concerned. Their puzzling finding was that farmers did not use fertilizer although they knew how to and the estimated annual rate of return was 70%, largely covering the cost of fertilizer. "For the average farmer in our sample, who farms 0.93 acres of land, these estimates imply that using fertilizer would increase maize income net of input costs by about $9.59 to $15.68 per season, on a base of about $89.02". They showed that under standard assumptions there must be implausibly large fixed costs to prevent investment in fertilizer.

In a companion paper, Duflo, Kremer, and Robinson (2010) (DKR hereafter) showed how behavioral biases impeded the exploitation of profit opportunities such as the use of fertilizer. They argued that some individuals are present-biased and that this behavioral bias favors procrastination (O'DONOGHUE; RABIN, 1999). The existence of present-biased preferences has been widely documented by experimental research in psychology and economics, sometimes known as *hyperbolic discounting*. In the case of investments in fertilizer, the hypothesis is that farmers who are present-biased fail to realize that they might be impatient in the future because they are patient today. They therefore systematically defer investment to a future date, up to a point when it becomes too late. Why do they behave in this way? The reason is that buying fertilizer, even at a low price, induces a small cost (going to the store) that is not covered by present-biased discounted future utility.

DKR developed a model with a dichotomous population involving both present-biased farmers and "regular" farmers. Their model predicted under-investment by farmers who were present-biased, entailing sub-optimal levels of input use (fertilizer). They used their model to evaluate two policies. The standard policy (policy 1), which implements heavy subsidies for fertilizer, can induce present-biased farmers to buy fertilizers in advance, but on the other hand it can also induce regular farmers to overuse fertilizers, ending in counterproductive use. The alternative policy (policy 2) offers a small discount on fertilizers immediately after harvest (when farmers have cash). According to their model, such a policy induces fertilizer purchases of the same order of magnitude as the purchases induced by high subsidies later in the season.

The predictions of the model were tested on the basis of an RFE in Kenya. Some farmers were randomly allocated to policy 1 and the others to policy 2. Policy 2 consisted of offering free delivery to farmers early in the season. Policy 2 increased fertilizer use by 47% to 70%, depending on the farm. The authors showed that the impact was greater than could be obtained by a policy offering free delivery and a 50% rebate on fertilizer later in the season. They also showed that about 70%

of the farmers were present-biased and about 60% never bought fertilizer. Their conclusion was that a "paternalistic libertarian" approach (THALER; SUNSTEIN, 2008) involving "small time-limited discounts could yield higher welfare than either laissez-faire policies or heavy subsidies, by helping stochastically hyperbolic farmers commit themselves to investing in fertilizer while avoiding large distortions in fertilizer use among time-consistent farmers, and the fiscal costs of heavy subsidies".

4.3 Designing managerial incentives

The theory of incentives has largely demonstrated the superiority of performance-based payment schemes over fixed wages to incentivize workers and managers to exert high effort and increase the overall profit of firms.

Bandiera et al. (2009) studied the effect of social connections between workers and managers on productivity in the workplace. To evaluate whether the existence of social connections is beneficial for a firm's overall performance, they explored how the effects of social connections varied with the strength of managerial incentives and workers' ability. To do so, they combined two types of data: panel data on individual workers' productivity from personnel records and data collected from a natural field experiment. The interesting part is how the field data was produced: the authors could observe an exogenous change in managerial incentives from fixed wages to bonuses, based on the average productivity of the workers managed. Although the counterfactual is missing in this type of study, the effects observed after manipulating the incentive schemes are sufficiently strong to remove any doubts about causality.

There are two key findings of this study: first, when managers are paid fixed wages, they tend to favor workers with whom they have social connections irrespective of the workers' ability, and second, when managerial incentives are based on performance, managers tend to favor high-ability workers irrespective of their social connections with them.

The authors relied on detailed personnel data from a leading UK farm that switched from a relative incentive payment scheme to payment by piece rate. The workers' daily task consisted of fruit picking. The analysis compared workers' productivity under relative incentives and piece rates. Under relative incentives, daily pay depends on the ratio of individual productivity to average productivity among all co-workers on the same field and day. Under piece rates, individual pay depends on individual productivity alone.

Under relative incentives, individual effort imposes a negative externality on co-workers. Under piece rates individual effort has no impact on co-worker pay. If there is a difference of productivity between the two schemes, it means that the workers internalize the negative externality that their individual effort imposes on co-workers.

To analyze the impact of the change in incentives, the authors observed the daily productivity of the same workers before and after the introduction of piece rates. By looking at within-subject productivity variations, they controlled for time-invariant sources of unobservable individual heterogeneity. Further control was provided by keeping an identical work environment throughout: tasks, technology, management, and other practices were the same before and after the change of incentive schemes.

The change from relative incentives to piece rates increased average productivity per worker by at least 50%, a significant and permanent effect. This is consistent with the assumption that workers internalize the negative externality they impose on co-workers, leading to lower average productivity, although this internalization is partial.

According to social preference theory, this means workers put some weight on the benefits of others in their utility (the average estimated weight is 0.65% of own weight, i.e. 1).

Why do workers care about others? The authors address two plausible reasons that might explain their behavior: altruism and collusion. To address this issue the authors compared a subgroup of workers where monitoring was not possible with another subgroup where monitoring was feasible. They found that the negative productivity impact of relative incentives was observed only when monitoring was available, excluding the altruistic preference explanation. Workers only care about others when they can be monitored and can monitor others.

Can we conclude that piece rates are optimum payment schemes in the presence of externalities among workers? No, because if workers have social preferences, incentives affect productivity but can be positive or negative. Some incentive schemes may produce positive externalities on co-workers, which would therefore increase individual productivity. The authors show in a thought experiment that if the firm moved from piece rates to group incentives, assuming the same externality among workers, productivity could increase by 30%.

The main conclusion is obvious: optimum incentives depend on workers' preferences. Understanding those preferences is a key to increasing workers' productivity.

5. Discussion

The growing popularity of experimentation in economics has widened the scope for economic experiments. In this paper we discuss the relevance of experimental methods for economic impact assessment of institutions, social programs and alternative policy instruments. We contrast RFEs, considered the "gold standard" of impact evaluation, and other types of experiments such as natural experiments and field experiments in terms of control and validity (internal and external). Although RFEs offer an attractive compromise between validity and control, this method has limitations, since RFEs are not always feasible.

The article discusses 3 case studies that relied on economic experiments at the individual level (incentives to increase productivity at the firm level in the UK), state level (incentives to increase crop yield in Africa) and regional level (allocation of irrigation quotas in the US). In the case of irrigation quota allocation (regional level), a field experiment was used, and moving from lab experiment to field experiment was very important because the behavior of the lab participants in auctions was unreal compared to the real participants. In the case of the incentives to increase crop yields at the state level in Africa, RFEs were used in an ex-ante evaluation of two different policies with very good results because it was possible to define separable units and collect enough data.

The natural field experiment to evaluate the impact of incentives to increase productivity at the firm level in the UK did not involve a counterfactual but showed strong causality between measures of daily productivity before and after the change of incentive schemes. RFEs are not a good method for evaluation of this policy because the spillover effects on control units are high.

In sum, the analysis shows that RFEs are well suited to evaluations where it is possible to define separable units and to collect enough units to run meaningful statistical tests (including the possibility of measuring spillover effects on control units).

Our findings are in line with the discussion that took place at a conference co-organized by the International Initiative for Impact Evaluation (3ie), the African Evaluation Association (AfrEA), UNICEF and the Network of Networks on Impact Evaluation (NONIE) at Cairo in April 2009 on approaches to assessing development effectiveness. The organizers of the conference motivated interest in a session entitled "Designing impact evaluations: different perspectives" by noting that "debates get stuck when they remain at the conceptual level, but that a greater degree of consensus can be achieved once we move to the specifics of the design of a particular evaluation" (see KARLAN, 2009). In the spirit of this quote the key session of the conference involved four experts with different backgrounds who were asked to propose an evaluation methodology for three different types of interventions: (i) a conditional cash transfer, (ii) an infrastructure project and (iii) an anti-corruption program. While the examples were taken as issues in development economics, the answers given by the panelists were more general and relevant for other fields.

The exact framing of the three questions was as follows:

- *A conditional cash transfer in a Central American country, in which households receive a monthly payment if females of school age remain in school and meet specified attendance and performance requirements.*
- *A transport sector program in a South Asian country that includes port rehabilitations, trunk road rehabilitation, and new investments in rural feeder roads.*
- *An anti-corruption commission (ACC) in an African country. The program includes helping develop guidelines, infrastructure upgrading and study tours. Similar programs are being implemented in six countries.*

For the first issue, the Conditional Cash Transfer (CCT), panelists agreed that the appropriate method would be an RFE. The program is targeted and allows assignment of individuals who benefit from the program on a random basis. Moreover, spillover effects are limited. The third issue, the anti-corruption commission (ACC), requires a nation-wide evaluation as corruption is pervasive. Spillover effects are large, and, as one of the panelists wrote, "transparency in the evaluation approach really matters!" He cited the example of Brazil where municipal audits were televised (as in Indonesia). For the second issue, the transport sector program, the authors agreed that it represented a hybrid program between the CCT and ACC which required government commitment. An RFE can be implemented for specific parts of the program, but spillovers are large. Ports and trunk roads typically require process evaluation methodologies, in the form of program monitoring. For feeder roads it is possible to make a randomized development plan, when spillovers are conceivably limited.

This discussion shed some light on how economic field experiments can be used for evaluation purposes. First, RFEs cannot always be implemented, but when feasible such methods should be preferred because they produce the right counterfactual. As a tentative conclusion, experimental methods are fit for evaluation purposes and can be useful guides for selecting projects. However, there are also serious limitations, especially when one considers nation-wide projects or local projects that have large spillover effects. In the latter case, experimentation is no longer the relevant tool for evaluating alternatives, and probably the debate about which project to implement has to be carried to the political arena.

References

BANDIERA, O.; BARANKAY, I.; RASUL, I. Incentives for managers and inequality among workers: evidence from a firm-level experiment. *Econometrica*, v. 77, n. 4, p. 1047-1094, July 2009.

_________. Social preferences and the response to incentives: evidence from personnel data. *Quarterly Journal of Economics*, v. 120, n. 3, p. 917-962, 2005.

CARPENTER, J.; HARRISON, G.; LIST, J. Field experiments in economics: an introduction. In: CARPENTER, J.; HARRISON, G. W.; LIST, J. A. (Eds.). *Field experiments in economics*. Greenwich, CT: JAI Press, Research in Experimental Economics, v. 10, 2004.

CUMMINGS, R.; HOLT, C.; LAURY, S. Using laboratory experiments for policymaking: an example from the Georgia irrigation reduction auction. *Journal of Policy Analysis and Management*, v. 23, n. 2, p. 341-363, 2004.

DUFLO, E.; KREMER, M.; ROBINSON, J. How high are rates of return to fertilizer? Evidence from field experiments in Kenya. *American Economic Review* (Papers and Proceedings), v. 98, n. 2, p. 482-488, 2008.

_________. *Nudging farmers to use fertilizer: theory and experimental evidence from Kenya*. Working paper, 2010.

FISHER, R. A. *The design of experiments*. Edinburgh: Oliver and Boyd, 1935.

FRIEDMAN, M. *Capitalism and freedom*. University of Chicago Press, 1962.

HARRISON, G. W.; LIST, J. A. Field experiments. *Journal of Economic Literature*, n. 42, p. 1009-55, 2004.

HARRISON, G. W. Field experiments and control. In: CARPENTER, J.; HARRISON, G. W.; LIST, J. A. (Eds.). *Field experiments in economics*. Greenwich, CT: JAI Press, Research in Experimental Economics, v. 10, 2004.

KARLAN, D. *Thoughts on randomized trials for evaluation of development*: presentation to the Cairo evaluation clinic. Yale University, Innovations for Poverty Action, Jameel Poverty Action Lab, 2009.

LEVITT, S.; LIST, J. What do laboratory experiments measuring social preferences tell us about the real world? *Journal of Economic Perspectives*, v. 21, n. 2, p. 153-174, 2006.

_________. Field experiments in economics: the past, the present, and the future. *European Economic Review*, n. 53, p. 1-18, 2009.

LIST, J. Field experiments: A bridge between lab and naturally occurring data. *The B.E. Journal of Economic Analysis & Policy*, v. 6, n. 2 - Advances, article 8, 2006.

O'DONOGHUE, R. M. Doing it now or doing it later. *American Economic Review*, v. 89, n. 1, p. 103-124, 1999.

ROE, B.; JUST, D. Internal and external validity in economics research: tradeoffs between experiments, field experiments, natural experiments and field data. *American Journal of Agricultural Economics*, v. 91, n. 5, p. 1266-1271, 2009.

ROTH, A. *The handbook of experimental economics*. John H. Kagel and Alvin E. Roth, editors, Princeton University Press, v. 1, 1995.

SMITH, V. Economics in the laboratory. *The Journal of Economic Perspectives*, v. 8, n. 1, p. 113-131, 1994.

THALER, R.; SUNSTEIN, C. *Nudge:* improving decisions about health, wealth, and happiness. New Haven, CT: Yale University Press, 2008.

CORRESPONDENCE ADDRESS:

Marc Willinger – willinger@lameta.univ-montp1.fr
LAMETA, Université de Montpellier 1
Avenue Raymond Dugrand, C.S. 79606
34960 Montpellier Cedex 2 – France