How can Randomized Controlled Trials help improve the design of the Common Agricultural Policy?

Luc Behaghel[1], Karen Macours[1] and Julie Subervie[2]

[1] Paris School of Economics, INRA;

[2] CEE-M, Univ. Montpellier, CNRS, INRA, Montpellier SupAgro, Montpellier, France

Abstract

We illustrate how randomized controlled trials (RCTs) could be used to evaluate the impact of alternative designs of the Common Agricultural Policy (CAP). We select four policy-design issues which relate to different components of the CAP and raise a wide range of economic questions: nudges, coordination failures, equity-efficiency trade-offs, contract design. Based on examples from agricultural and social policies in developing and developed countries, we show that RCTs have provided useful rigorous evidence on similar design issues, suggesting that they could also be leveraged to help improve components of the CAP.

Keywords: Common agricultural policy; Impact evaluation; Policy design; Field experiments

JEL codes: C93; Q18

Acknowledgement: Luc Behaghel and Karen Macours acknowledge the support of the French National Research Agency (ANR), under grant ANR-17-EURE-0001.

1. Introduction

Despite the renewed demand for evaluating agricultural policies in the European Union and the revival of randomized controlled trials (RCTs) in the evaluation of social policies in Europe since the mid-2000s, RCTs have not been applied to the evaluation of the Common Agricultural Policy (CAP) (Colen et al., 2016). This may seem surprising given the role played by RCTs during the "glory days" of agricultural economics with seminal contributions by Fisher and Neyman (Herberich, Levitt and List, 2009). There are undoubtedly many reasons for this absence, both on the demand side (e.g. the type of evaluations requested or accepted by various stakeholders such as governments, international agencies, farmers) and on the supply side (e.g. the incentives available to the consultants and academics operating in this area). This paper abstracts from these constraints -- each of which deserve careful analyses of their own -- and asks a preliminary question: are RCTs an appropriate tool to address issues relevant to the design of today's CAP?

By allowing direct causal inference from the comparison of a control and treatment group randomly drawn from the same population, RCTs provide a uniquely robust approach to evaluate programs implemented in the field (Athey and Imbens, 2017). Such evaluations can serve both as accountability measures (Are programs obtaining their long-term objectives in terms of anticipated outcomes, and in ways that justifies their fiscal burden?) and as a learning tool (Is the program design appropriate in obtaining expected outcomes? How can it be improved?). This paper focuses on its role as the latter, as RCTs can be particularly powerful instruments in improving understanding about the effectiveness of policies.[1] For many aspects of the CAP, the important policy question to address is not whether a certain type of policy ought to exist, but rather whether specific changes to the design of an existing policy could lead to better efficiency and/or equity outcomes. To the extent that such questions are of interest, RCTs can help provide rigorous answers by randomly introducing potentially relevant variations in the design, and testing their implications for a policy's effectiveness.

We consider the two pillars of the CAP[2], and select a list of policy-relevant design issues that could benefit from rigorous empirical evidence: (i) How could the uptake of agri-environmental measures by farmers be increased? (ii) What schemes can induce coordination among farmers on environmental

[1] As such, even though RCTs belong to the ex post evaluation approach (as they are based on observational data, measuring outcomes after they are realized, as opposed, e.g., to ex ante simulation methods), they can be used "ex ante", i.e. before scaling up actual policies.

[2] The first pillar consists of direct payments to farmers and market measures, and the second pillar is a rural development policy aimed at balancing territorial development with sustaining a farming sector that is environmentally sound, while simultaneously fostering competitiveness and innovation (European Commission, 2013).

issues in which one farmer’s defection can jeopardize the efforts of all the others? (iii) What implications could different mechanisms for direct payments have on efficiency and equity? (iv) How to design agri-environmental contracts to ensure that farmers receive sufficient compensation for their conservation efforts without inducing large efficiency losses? We discuss how these policy-design issues relate to more general economic questions, and show how RCTs have been used in comparable contexts to provide rigorous empirical evidence on these questions.

These questions illustrate that RCTs should not be primarily considered as tools to evaluate the impact of existing policies that are already being implemented universally to the entire target population. As such, they do not necessarily provide an alternative to the typical ex-post evaluations based on observational data. Instead, they can be particularly powerful to empirically test whether and to what extent potential policy changes lead to anticipated (and potentially unanticipated) impacts when implemented in real-life contexts, but prior to the universal application of the policy change. As such, they can help inform decisions regarding potential policy (design) options and provide credible and rigorous empirical evidence of impacts that complement the typical theoretical and modeling based ex-ante approaches.

The four policy-design questions are meant to demonstrate the adaptability of RCTs, and show that they can be applied both to test the adequacy of the policy design (doing the right thing) and the adequacy of the policy implementation (doing things right). It may seem hardly surprising that, transposing and extending lessons drawn from the lab by behavioral economists, experiments on information-based interventions or different types of nudges can shed light on a farmer’s propensity to adopt agri-environmental contracts. One may be skeptical, however, about what can be learned from RCTs with respect to the three other issues, namely coordination failures, equity – efficiency trade-offs, and contract design. Despite this, we believe that one of the key lessons that emerges from the recent literature is the strong adaptability of the RCT toolkit and its usefulness in addressing questions that had been deemed out of reach only a few years ago. To illustrate this claim, we complement the discussion of hypothetical experimental designs that could be applied to the CAP with actual examples from two related policy domains: agricultural policies in developing countries, and social policies in Europe and other contexts.

This article contributes to a small prospective literature on the role that RCTs can play in the evaluation of agricultural policies. Heberich et al. (2009) build on observations over the past century to argue that agricultural economists could take “a new train at the depot,” in the form of RCTs inspired by laboratory experiments and economic theory. Colen et al. (2016) provide a detailed review of the state of the art in the evaluation of the CAP, and discuss the comparative advantages of different experimental methods, and their complementarities with other methods. Both papers take an optimistic view of the potential of the RCT toolbox while acknowledging that this potential has not yet been realized. To complement these discussions, we highlight potentially fruitful examples of evaluations, inspired by recent successes in neighboring areas. The next section provides more background on the debates regarding RCTs and policy-making. The subsequent sections consider the four policy design questions in turn, and the last section offers several concluding remarks.

2. Background: The debated role of RCTs and policy-making

Existing CAP evaluation relies on a wide range of methods: simulation models, statistical and econometric analysis of survey data, qualitative evaluations, and, more recently, experimental methods. Colen et al. (2016) argue that the new policy instruments introduced over the past three decades call for an increased reliance on experimental methods, which are particularly adapted to analyze farmers' heterogeneous responses to decoupled payments, or enrolment decisions in voluntary schemes such as agri-environmental contracts, for instance. Moreover, the greater discretion granted to Member States in the 2014-20 CAP programming period gives room to regional or State-level policy experiments that can be accompanied with experimental evaluation. Colen et al. (2016) provide several examples of how discrete choice experiments and laboratory experiments have been used in that context to help design and evaluate the CAP. There also exist a large number of positive mathematical programming models covering individual European States or regions; and a number of them have full EU coverage.[3] Randomized controlled trials, however, have remained absent from the CAP evaluation toolbox. As noted above, this comes in sharp contrast with the rise of RCTs in the evaluation of agricultural policies in developing countries, and in the evaluation of social policies in the European Union. If the law of diminishing marginal returns applies, this fact suggests that introducing a little bit of RCTs to evaluate the CAP may have high returns. Yet another fact calls for caution: the rising popularity of RCTs in policy-making has come with continuing debates on the pros and cons of the approach:[4] does this mean that RCTs have been overvalued in other policy areas, providing a good reason not to use them for agricultural policy-making? While a complete summary of the debates on RCTs is beyond the scope of this section, we use them to derive "points of attention" when designing RCTs, which we then apply to the specific examples considered in this paper: what threats against the validity of RCTs should we consider? What are the conditions for their feasibility and acceptability?

[3] See for example to the CAPRI-FT model (Gocht and Britz, 2011), the AROPAj model (De Cara and Jayet, 2011) or the IFM-CAP model (Louhichi et al., 2017; Louhichi et al., 2018).

[4] For recent conflicting views, see Deaton and Cartwright (2018) and the comments in Imbens (2018) as well as in the special issue of *Social Science & Medicine* 210 (August 2018); Ravallion (2018) also revisits the debates.

Internal and external validity of RCTs for policy-making

Following Roe and Just (2009), RCTs can be viewed as an intermediate approach to the fundamental trade-off between internal and external validity in empirical research.[5] While laboratory experiments achieve high internal validity by fully controlling the context of the experiment, the specificities of the laboratory context limit its external validity. At the other end of the spectrum, naturally occurring data can be representative of the context of interest, so that results may have high external validity, but their internal validity may be questionable in the absence of exogenous sources of variation to identify causal effects. By contrast, RCTs – when well-designed and carefully implemented – may provide both relatively high internal and external validity: in RCTs, the researcher manipulation of a naturally occurring context induces the exogenous variation needed for causal identification while the context remains policy-relevant, and is more likely than a laboratory to resemble other policy-relevant contexts. However, naturally occurring contexts cannot be fully controlled, which can pose threats to internal validity; and unless the RCT is implemented at the same scale as the policy of interest, without further assumptions external validity is limited to its context, as in any empirical research (see Deaton and Carwright, 2018). Viewing RCTs as an intermediate approach in the trade-off between internal and external validity has two important consequences. First, it implies that ranking RCTs against other empirical methods will involve putting weights on these two criteria, which is not straightforward. As noted by Athey and Imbens (2017), claims in favor of RCTs based on the primacy of internal validity have been challenged by other researchers arguing that there is no particular priority for internal validity over external validity (Manski, 2013). Second, we should not expect all RCTs to reach the same point in this trade-off. Trivially, large-scale RCTS are more likely to have high external validity, but this may come at the expense of internal validity if compliance with the design is less controlled. More generally, the degree of internal and external validity of an RCT depends on the context of the study, the experimental design, and the execution of the experiment.

Among the different threats to validity identified in the literature, let us discuss two that have direct consequences for the specific examples of CAP impact evaluations considered below: violation of the "Stable unit treatment value assumption" (SUTVA) and differential attrition in follow-up data.[6] SUTVA requires that the (potential) outcomes of control units be unaffected by the fact that other units get treated, so that they reflect what would have occurred to treatment units in the absence of the

[5] Following Shadish, Cook and Campbell (2002), a study has internal validity if the observed relationship between a treatment and an outcome reflects a causal relationship; and external validity "concerns inferences about the extent to which a causal relationship holds over variation in persons, settings, treatments, and outcomes" (p. 83). See also Athey and Imbens (2017).

[6] A more complete list would notably include imperfect compliance with treatment assignment (and the distinction between intent-to-treat effects and local average treatment effects), and "experimental effects" such as the Hawthorne and Henry effects, by which treatment and control units change their behavior when they are aware of participating in an experiment.

program evaluated. This rules out general equilibrium effects, peer effects and any sort of externalities. However, SUTVA is by no way required for RCTs: when it holds, simple experimental designs (e.g. simple individual randomization) can be used, but alternative designs can be used in cases where it is likely to be violated. For instance, peer effects and other externalities can be taken into account in a "partial population design" (Moffitt, 2001) that introduces random variation in the exposure of non-treated units to interactions with treated units; applications include Miguel and Kremer (2004) in the context of health in Kenya, or Avvisatti et al. (2014) in French public schools. General equilibrium effects can also be estimated in designs that consider local markets as the units of randomization (see Crépon et al., 2013, for an active labor market program in France).

Differential attrition[7] threatens the validity of RCTs as survey response behavior potentially reintroduces selectivity: the subsamples of the treatment and control groups for which a follow-up can be completed may not be comparable anymore, even though the initial two samples were comparable due to randomization. Selective attrition is a pervasive issue. It is often neglected or only corrected on the basis of observable covariates in panel surveys. By contrast, evaluators conducting RCTs have paid increasing attention to it both as a threat to representativeness and to causal identification. More and more RCTs use exhaustive administrative data that do not suffer from the problem; in addition, robust statistical methods have been developed to estimate or bound the effects under minimum assumptions in the presence of differential attrition (Lee, 2009; Behaghel et al., 2015).

Feasibility and acceptability

Another possible reason why RCTs have not been used to evaluate the CAP may be that they are simply not feasible, or not acceptable from a political or ethical perspective. Technically, RCTs are feasible under two minimum conditions: assignment to treatment[8] can be randomized, and both the "control" and "treatment" groups that are thus created can be followed. As illustrated in the next section, randomization can be conducted in ways that minimize disruption and may not even ever be perceived by individuals participating in the experiment. Also, the follow-up can rely upon administrative data that are collected for other purposes. In such cases, there is little doubt on the technical feasibility of RCTs. But it clearly depends on the question at stake, and some policies will probably never be evaluated with an RCT (think of major macroeconomic or fiscal reforms). Our take of the literature is, however, that many evaluation questions for which RCTs were deemed not feasible have been actually

[7] Differential attrition is characterized by the fact that outcome measures are missing in different proportions in the treatment and control groups.

[8] Note that random assignment to treatment does not mean that all members of the treatment group enroll in the program, and that no member of the control group enrolled (which would define perfect compliance to the assignment). It is enough to randomly send an encouragement to enter the program ("encouragement design"); this random encouragement is then used as an instrumental variable to identify the impact of program enrollment.

successfully addressed by RCTs. Take for instance the evaluation of the structure of peer effects in education. An active researcher in that field, Caroline Hoxby, had been putting forward quasi-experimental results, arguing that experiments that would artificially generate significant random differences in class composition would not be politically feasible. A few years later, at least two experiments of that sort had been run: one in Kenya testing the effectiveness of tracking (Duflo, Dupas and Kremer, 2011); one in the US, testing the optimal composition of squadrons at the naval academy (Carrell et al., 2013). Concerning measurement and follow-up, innovative ways of measuring outcomes have been found -- see for instance the measurement of corruption in Olken (2007). In short, what RCTs can technically do must be judged empirically, not a priori. This does not mean that RCTs come at no cost: the sheer fact that access to a program needs to be manipulated for evaluation purposes may complicate the program operations. This may require long-run investments by evaluators to gain trust within implementing partners' organizations. A related concern is that only a subset of programs run by selected implementing partners may be evaluated with RCTs (Ravallion, 2018): RCTs would solve the issue of selection into treatment, but would do so only for selected programs, which would limit their external validity. While this concern is real, it is unclear if it is a concern against RCTs, or a call to run more of them. Feasibility is endogenous: methodological innovations increase the range of policies that can be evaluated with RCTs, and their diffusion increases the readiness of implementing organizations and authorities to rely on them, pushing to their use in more and more varied contexts.

A full discussion of the ethics of RCTs is again beyond the scope of this paper. Ravallion (2018) insists that ethical concerns should be taken more seriously, building upon guidelines elaborated for medical trials, such as the principle of equipoise, by which experimenting (and depriving the control group from treatment) is only ethical as long as we are sufficiently ignorant about treatment impact. What however "sufficient ignorance" entails is debatable. From a positive perspective, Banerjee et al. (2017) precisely argue that RCTs have developed as a way to respond to "adversarial audiences" that would not accept "reasonable" priors. Applying this discussion to the CAP, a tension may appear between two agendas: a "learning" agenda that would use RCTs to answer questions for which uncertainty is high, and a "cost-cutting" agenda that would use RCTs as a disciplinary device through which any intervention would have to prove its (cost-) effectiveness, somehow imposing the burden of the proof to farmers involved in the experiments. Healthy scepticism fosters good science, but should not be used to deny existing evidence.

Overall, our take of this quick review is that RCTs have proven a useful learning tools in many policy areas, and experience has accumulated on how to use them best, so that RCTs could usefully enter the CAP evaluation toolkit, with due methodological and ethical care. We try to outline how this may occur by considering specific examples in the next sections.

3. Inducing behavioral changes: nudges and information-based experiments

Farmers are involved in decisions with complex trade-off dynamics and sometimes lack relevant information when considering engaging in new agricultural practices. The complexity and changing nature of some of the environmental regulations and policies, as well as the heterogeneity in conditions and constraints among many farmers in the EU, may well imply that some farmers are not necessarily aware of all the information and practices relevant for their particular situation. This potentially is a bigger constraint among smaller (potentially part-time) farmers or less-connected farmers in newer member states. Whether such information constraints actually exist, and if so for whom, is clearly an empirical question. A policy experiment that provides detailed information regarding eligibility for certain benefits, or advantages of shifting practices, to a subset of farmers selected randomly from a specific targeted group, and afterwards comparing their outcomes with those from the same group that were not selected, could help evaluate such questions.

Even when farmers have access to all the relevant information, the complex and dynamic trade-offs they need to consider for many of their decisions may foster the use of heuristics and a susceptibility to behavioral inertia and the status quo in agricultural practices (Dessart et al., in this issue). Insights from behavioral science suggests that "nudges" (Thaler and Sunstein, 2008) can then be powerful tools to shift decision making. A nudge is not a cash payment. It may consist, for instance of the provision of information about the social norm (about what others think or do). It can also simply consist of the way in which a choice is presented. Empirical evidence shows that nudges can be effective in changing consumer behavior, in particular pro-environmental behaviors (see Schubert (2017) for a review). The question then arises: is it possible to nudge farmers to change their practices, and possibly adopt more pro-environmental (or "greener") practices?

There are indeed some reasons to believe that agri-environment scheme (AES)[9] uptake could be improved with the use of nudges. Some studies suggest that nudges can affect farmers' intentions to (re-)enroll in AES (Kuhfuss et al., 2016; Chen et al., 2009). There is also evidence of a "social identity" mechanism behind individual decisions (Goldstein et al., 2008), by which an individual adheres to the descriptive norms of the group of which he considers himself a member. This effect may well apply to

[9] Agri-environment schemes encourage farmers to protect and enhance the environment on their farmland by paying them for the provision of environmental services.

farmers, especially among members of a cooperative. Thus, there is reason to believe that this type of psychological lever could be useful in improving the design of agri-environmental policies.

In recent years, behavioral economics research using lab experiments has grown rapidly and has provided evidence of a large number of tools that can be used to encourage green practices. Because these results cannot always be directly extrapolated to the field, lab-in-the field experiments[10] are sometimes used to validate results derived from the lab. RCTs could be considered as the next logical step in the chain, and the insights from the lab-in-the field regarding context can indeed help to subsequently design a RCT. The RCT in turn allows study whether and to what extent the nudge affects actual real-life decisions.

Despite a recent craze for nudges in both the academic and the public sphere (in UK and US in particular[11]), there are almost no experimental studies providing evidence of the impact of nudges on farmers' decisions in the real world. One exception is a study by Messer et al. (2015), who ran an RCT in which farmers from Texas, Delaware, and Maryland competed in an auction of conservation contracts that required them to adopt practices that reduced nutrient run-off. The authors show that changing the default option can result in larger bids, and that providing information about what others think of the required practice increases the likelihood that a producer participates in the auction. Wallander et al. (2017) and Chabé-Ferret et al. (2018) also use social norms in field experiments with the aim of inducing greater farmer participation in AESs, and they make use of RCTs to test the effectiveness of these nudges. Both studies failed to detect a statistically significant impact of these interventions on farmer participation in the agri-environmental schemes, demonstrating the importance of testing the theoretical attractive instruments in real life conditions.

Many alternative nudges remain to be tested in the field and could help improve the design of agri-environmental policies. Ferraro et al. (2017), for example, suggest using a default option in order to increase farmer participation in the US Conservation Reserve Program. The authors also suggest presenting payments in agri-environmental programs using a loss frame, specifically by stating the maximum payment the participant could earn and how much he would lose for every practice not adopted. Both of these suggestions could also work in the European context. Previous evidence thus suggests that insights from behavioral sciences, combined with rigorous empirical tests of the concepts, could indeed be used to improve the design of the second pillar of the CAP, and notably the way it is presented to farmers.

[10] Lab-in-the field experiments differ from other laboratory experiments by recruiting participants from the field for which a measure is experimented (e.g., farmers rather than students when considering AES) and by framing the experiment to represent the field context ("contextualized" experiment). Suter and Vossler (2014) have run ambient tax experiments with student subjects and dairy farmers and found that the two groups may produce different results.

[11] Several governments in developed countries have constituted 'behavioral insights teams' within their civil services. In 2014, the U.S. Department of Agriculture (USDA) created the Center for Behavioral and Experimental Agri-Environmental Research (CBEAR).

For a variety of information interventions or nudges, the evaluation of the scheme's effectiveness using a RCT seems quite feasible, since large samples can be easily reached and as long as the design assures that the Stable Unit Treatment Value Assumption (SUTVA) is likely to hold. One difficulty of such experiments, however, is that small (albeit valuable) effects may be hard to detect, as the record of the "What Works Centres" in UK shows: experimenters must therefore make sure that they have sufficient statistical power. One limitation is that the interpretation of the effects is not always straightforward in the presence of a variety of plausible "behavioral" models and heterogeneous types of behavior (e.g. Duflo, Kremer and Robinson, 2011). But even if the impact of a nudge is small, it may remain quite cost-effective, as these types of interventions tend to require very little in terms of implementation costs. This is all the more true in countries such as France, where the online application for agricultural subsidies has been mandatory since 2016, and the (marginal) cost of providing information or nudges through pop-up windows in the Telepac website would thus be small. Take, by way of example, potential measures to reduce contamination by pesticides from agriculture, a source of water quality degradation in several countries in the European Union. In France, during the 2007–2014 CAP programming period, AESs have been used to fulfill the objectives of the European Union water framework directive and have been implemented in catchment areas where water quality improvement has been identified as a priority from 2007. As has been the case for many other measures, the participation of farmers has been low (Kuhfuss and Subervie, 2018). One reason for this may be that many farmers actually lack all of the relevant information when considering reducing the use of pesticides. Since 2008, the Dephy farm network of the French Ecophyto program[12] has been experimenting with techniques that reduce the use of pesticides and herbicides without affecting crop productivity and profitability and recent analyses suggest that they have been successful in many cases (Lechenet et al., 2017). The dissemination of these innovative practices is a major challenge faced by the program. But as long as this information can be embedded in customized messages,[13] dissemination to a very large number of randomly selected farmers via the Telepac application should be quite feasible.

Evaluating the impact of such intervention would also require collecting data on farmers' pesticide use after they have received information about the innovative techniques. This could be done with no additional cost through the Farm Practices Survey run every three years by the French Department of

[12] Since 2010, the French Ecophyto plan for the reduction of pesticide use has been providing free technical assistance to 3,000 volunteer pilot farms, a group called the Dephy network, with the aim of converting these farms to eco-friendly farming systems. More information is available on EcophytoPIC, the French website for integrated pest management (http://www.ecophytopic.fr/).

[13] Interaction between the control and treatment groups can lead to a number of challenges that can threaten the internal validity a randomized evaluation. In our example, the control group may benefit from the treatment if treated farmers share the information they receive with untreated farmers. Sending customized messages (of no apparent interest to other farmers) may limit such spillovers.

Statistics of the Ministry of Agriculture (SSP), which collects data on phytosanitary treatments in French farms from the four main categories of crop production.[14] Many other similar information-based experiments could be envisaged, in France and presumably in other European countries.

4. Incentives to coordinate: experimenting with collective bonuses

As illustrated in the previous section, a number of different AESs have been put in place at the level of individual farms. However, the recent literature has emphasized the potential gains that could be realized by adopting an approach that is characterized by a larger, landscape-wide scale.[15] Accordingly, AES compensation payments are now allowed by EU regulation to be paid to groups of farmers (Regulation N° 1305/2013, article 28, cited in Westerink et al., 2017). Such payments, often referred to as "collective bonuses", raise important design questions. Contract theory highlights the various potentially counterproductive mechanisms at play. An obvious rationale for setting incentives at the landscape level is that many of the outcomes for which the schemes are designed (e.g., wildlife conservation, water quality and storage) are also applicable at the landscape level. While actions at the individual farm level could certainly contribute to these objectives, their impact on landscape-level results would arguably be weaker, thus yielding imperfectly aligned incentives. Employing "multi-tasking" models (Holmstrom and Milgrom, 1991) may also induce farmers to focus on better-incentivized tasks at the expense of other tasks that could be more important with respect to the environmental objectives identified. Collective bonuses can also create collective action problems, such as free-riding. The free-rider problem can, however, be mitigated to some extent by institutional arrangements or peer pressure. In short, the appropriate design of collective bonuses is a difficult theoretical question involving numerous plausible mechanisms whose effects may be context dependent. Empirical evidence is needed in order to weigh the relative strengths of these mechanisms and to assess their net effect. Relatedly, given the evidence in several instances that incentives may

[14] One caveat with using the Farm Practices Survey survey is that it is based on a representative sample of plots (and not farms). Thus, one agricultural practice referenced at the level of a surveyed plot may not reflect the average practice of the farm.

[15] A number of lab experiments have shed some light on the likely interactions between heterogeneous agents of the same group in a context of non-point source pollution (Cason and Gangadharan, 2013; Miao et al., 2016; Poe et al., 2004; Spraggon, 2004) or in the management of shared groundwater resources (Suter et al., 2012).

have perverse effects, transparent and convincing evidence that collective bonuses have the expected positive effects in the contexts of interest will be key to their political viability.

Although to our knowledge RCTs have not yet been used to evaluate agri-environmental schemes at the landscape level,[16] two examples suggest that they may prove a useful and feasible tool in this regard. In Uganda, Jayachandran et al. (2017) analyze the impact of financial incentives for forest owners to maintain the integrity of their forestland, thus providing experimental evidence of the effectiveness and cost-effectiveness of Payments for Ecosystem Services. Even though the incentives in this study are set at the farm level, where each individual is given the opportunity to enroll to receive payments if she refrains from clearing trees, the randomization of treatment assignments takes place at the village level such that the impacts that are measured encompass any "leakages" (externalities) and collective dynamics within villages. The authors also use random variation in the proximity between treatment and control villages to account for potential spillovers across villages. As a result, outcomes in this study are measured at the landscape level. Though there is no evidence of spillovers and limited evidence of leakages, program enrollment is low (32 percent), seemingly due to limited program awareness, leading the authors to question whether the program could be better marketed. Collective incentives could indeed be a way to do so, as they could induce farmers to advertise the program to other farmers in their village. In other words, everything in this experimental design is in place to analyze alternative collective schemes, and the results suggest that such schemes are indeed worth trying.

The second example comes from the incentive literature in the economics of education. Quite naturally, given the collective nature of the education production function at the classroom and school level, researchers have been experimenting with various forms of individual and collective bonuses for teachers and students as incentives to increase student performance. Empirical evaluations in this literature underscore the complexity of the mechanisms at play, and specifically their dynamic nature. For instance, Muralidharan and Sundararaman (2011) compare the effectiveness of teacher- and school-level incentives in India: while the two compare well in the first year, teacher-level incentives outperform school-level incentives in the second year, suggesting collective action problems might have come into play. Other systematic investigations such as those conducted by Roland Fryer or John List suggest that the literature in the economics of education remains on the steep segment of the learning curve regarding how to best harness the power of incentives in this area (e.g., Fryer et al., 2012). Much could certainly be learned from similar efforts with respect to AESs.

[16] Banerjee (2018) uses a laboratory experiment to examine the role of various mechanisms in incentivizing spatially coordinated land uses under the Agglomeration Bonus scheme. Using lab and artefactual field experiments, Fooks et al. (2016) evaluate the relative effectiveness of collective bonuses and spatial targeting (which provides the government with a means to select contiguous parcels over non-contiguous parcels when enrolment is budget-constrained) for achieving optimal contiguity of parcels in the landscape.

RCTs can easily be designed to evaluate the impact of collective bonuses. The number of farmers involved needs to be large, so that randomization takes place by clusters (e.g., randomization units at the landscape levels). Importantly, administrative data should be used as much as possible, and the monitoring of compliance with AESs should be strictly identical in the treatment and control groups: as many outcomes would be measured at the landscape level, there should be no attrition issues. Political feasibility may be more of a problem. Under which conditions would experiments with collective bonuses be acceptable in the field? A first condition is the acceptability of the bonuses themselves. If the scheme that one wants to test is faced with strong opposition, control and treatment groups will be affected by the conflicting views, and may distort their behavior to prove the policy wrong or right. In the context of educational policy, a proposed evaluation of collective "project grants" at the class level thus had to be abandoned: the proposed scheme had been the subject of heated debates in the medias, so that the conditions for a valid RCT were not met (Behaghel and Gurgand, 2010). Clearly, however, this had less to do with RCTs as such than with the intervention itself, and the way it was framed in the media. The pilot helped clarify what schemes would be acceptable and feasible, but it came too late. The role of piloting interventions before reaching the stage of a full RCT should not be underestimated. This pilot stage can and should build on a variety of evaluation tools: qualitative methods (focus groups, interviews, and observation) and lab-in-the-field experiments with stakeholders, for instance.

The second condition for acceptability is to answer fairness concerns that arise if, for example, the expected gains of treated individuals are higher than the gains of control individuals. It is however possible to design experiments in which the expected gains of the different experimental groups are similar -- see for instance Fryer et al. (2012), where incentives are framed differently in two treatment arms, but are financially equivalent. As suggested by Morawetz (2014), control farmers may also receive as a lump sum the same payment as the expected or average payments made to farmers in the treatment group.

5. Experimenting to understand design trade-offs for direct payments

The first pillar of the CAP involves direct payments which have become a core component of the CAP and are likely to be sustained in the foreseeable future as a key principle of the program. Prima facie, it may therefore appear that there is no useful role for RCTs in the evaluation of these policies. Yet beyond the normative questions regarding the principle of the payments per se, many questions exist regarding the specific design of the different interventions within the first pillar, over which member countries do exercise some degree of freedom. Limiting efficiency losses resulting from transfers matters for the sustainability of the CAP over the long term.

Incidentally, RCTs have proved to be a very useful tool in shedding light on the advantages and disadvantages of direct payments to farmers and other rural households in developing countries. The evaluation of a conditional cash transfer program in Mexico (*PROGRESA*, now named Prospera) was one of the first large-scale RCTs conducted in the developing world, and the experimental evidence of its impacts was crucial in assuring continued support through changing political climates. In fact, the program rapidly expanded nationwide in Mexico and triggered similar programs across Latin America, where they now reach 25 percent of the total population. Such programs have also spread to Africa, Asia, and even Europe and the United States over the last 20 years (Fiszbein and Schady, 2009; Robles, Rubio and Stampini, 2015).

The initial evaluation of the program in Mexico, which tested the effectiveness of one particular design, was followed by a large literature experimentally comparing alternative design variations. These studies helped policymakers understand, for instance, whether making transfers conditional, labeling them, or making them entirely unconditional impacted their effectiveness (Baird, Ozler and McIntosh, 2011; Benhassine et al, 2015). They also tested other design features, such as whether making payments conditional on actions or achievements was more effective (Barrera-Osorio et al, 2011). Such studies have now been carried out in many different settings (e.g. Malawi, Morocco, Burkina Faso, Bosnia, Colombia), and frequently in close collaboration with government ministries. The agencies implementing these studies often have valid questions regarding how to optimally design a transfer program for a specific objective or context.

Once it is acknowledged that the answers to such questions are not obvious beforehand, experimentation becomes an ethical policy procedure. While such RCTs do not call into question the benefits themselves, nor the justification for the policy goals they aim to achieve, they are well suited to provide empirical evidence on potential trade-offs between different objectives that alternative design options may imply. Learning from these experiences suggests that RCTs have potential as useful instruments in comparing alternative direct payment schemes under the first pillar of the CAP, and in investigating how different designs may imply different trade-offs between the possible objectives therein. A potential application could be to use a RCT to compare different methods of monitoring farmer's compliance with greening. Green direct payments account for 30% of EU countries' direct payment budgets. Farmers receiving an area-based payment have to make use of various straightforward, non-contractual practices that benefit the environment and the climate. Whether they use such practices then of course need to be monitored and paying agencies are required to inspect at least 5% of declarations, via field visits and photointerpretation of High Resolution satellite imagery (European Commission, 2013). Recently, using much cheaper remote sensing techniques has been suggested as a possible alternative (Sitokonstantinou, 2018). Clearly the method of monitoring affects both the costs and the potential effectiveness and trade-offs may include the accuracy and frequency of the measure. This in turn can affect farmers' compliance behavior. Indeed, the problem is quite similar to that raised for cash transfer programs in developing countries, where some evidence suggest that hard monitoring may not be needed to change behavior. Hence experimenting with the compliance monitoring method could help answer an important question. Such experimentation appears quite feasible, as files of different farmers could be assigned to different monitoring methods.

6. Targeting under adverse selection: experimenting with innovative agri-environmental contract designs

Voluntary agri-environmental contracts that offer fixed payment schemes suffer from two major problems. First, farmers who face the lowest costs for complying with environmental requirements are more likely to enter the program; in cases where the program would pay some farmers for doing nothing differently from what they would have done in the absence of payment, adverse selection may induce large windfall effects (Chabé-Ferret and Subervie, 2013). Second, those with the highest costs of participating are less likely to enter the program, though may precisely be those whose engagement would have the greatest contribution to the program's effectiveness (Kuhfuss et al., 2014).

One option for reducing the efficiency losses due to adverse selection is to shift from fixed payment schemes to auction mechanisms. Procurement auctions have been in place in other countries for many years (e.g. the US Conservation Reserve Program, established in the 1980s, as well as several pilot programs in Australia). However, the context in which conservation auctions are implemented may impact their effectiveness (Lundberg et al. 2018; Ferraro 2008). It is therefore impossible to anticipate the additional gains from auctions (compared to fixed payment contracts) in the European context; here again, RCTs could provide a useful way to address this question.

For several decades, researchers have used experimental auctions to estimate consumer willingness-to-pay for new products (Corrigan et al., 2009). In a demand-revealing auction mechanism similar to the Vickrey (1961) and Becker-DeGroot-Marschak (BDM) (1964) auctions, bids indeed provide a direct measure of auction participants' willingness-to-pay for the good being sold. There is a substantial literature dealing with the implementation of such designs in university experimental economics labs (Berry, Fischer and Guiteras, 2015 and references therein). More recently, a number of studies in developing countries have also demonstrated that these insights can have broader relevance for addressing design trade-offs in actual policy applications. Hoffmann, Barrett and Just (2009) used a BDM design to measure the gap between willingness-to-pay and willingness-to-accept for bed nets in Uganda. Berry, Fischer and Guiteras (2015) also use a BDM mechanism to estimate the willingness to pay for water filters in Ghana, while Guiteras and Jack (2017) use the same mechanism to investigate how workers respond to different contractual arrangements in the context of informal day labor markets in rural Malawi. Finally, and more closely related to the topic we address here, Jack (2013) uses a Vickrey auction to explicitly take into account landholders' willingness-to-accept an afforestation contract in Malawi. She moreover runs a RCT to demonstrate that landholders who received a tree planting contract as a result of bidding in the auction kept significantly more trees alive over a three-year period than did landholders who received the contract through a lottery. Much can certainly be learned from a systematic analysis of the demand for AESs in European countries using a similar approach, especially taking into account the heterogeneity that is likely to be present with respect to this demand.

In fact, recently in France, environmental auctions have been implemented in a pilot program run by the Water Agency Artois-Picardie (Kuhfuss et al., 2012). This pilot program was implemented in a way akin to the afforestation program evaluated by Jack (2013) in Malawi: auction participants were given a thorough explanation of the auction rules, then sealed bids were collected and ranked to determine both the auction clearing price based on the available budget and the rejected bids. Unfortunately, contrary to the Malawian program, the French pilot has not been implemented through random assignment of participants; therefore, it was not possible to assess how far it had outperformed (or not) the usual system of AESs.[17] This initiative suggests, however, that there is a demand for

[17] In an auction setting à la Vickrey (1961), it is possible to estimate the effectiveness of the contract using a regression discontinuity design (RDD), which is a quasi-experimental method that allows estimating the causal effect of an intervention by assigning a threshold above or below which the intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomization is unfeasible. This method however requires that the number of participants around the threshold is very large, which is rarely the case.

experimentation by the practitioners and public authorities, and also that it is *de facto* feasible to pilot such innovative approaches with some farmers and not others, so that there is only one step left to reach the stage of evidence-based evaluation of innovative schemes.

7. Concluding comments

The lessons from this prospective study can be summarized as follows. First, many insights from laboratory and field experiments in the behavioral sciences could be used to improve the design of the second pillar of the CAP, and RCTs would be useful in evaluating how well these approaches can contribute to achieving higher participation rates in real world conditions. Second, the value of collective incentives is a challenging theoretical question due to the many mechanisms that are potentially at play in determining the ultimate impacts of these incentives. RCTs may therefore be of particular use in weighing the relative strengths of plausible mechanisms in specific field settings. Importantly, the acceptability of experiments involving (collective) bonuses needs to be ensured: this requires preliminary work and evaluation using other methods on the scheme and its framing; it also requires paying attention to fairness concerns and providing control farmers with appropriate alternatives. Third, ensuring that direct payments to farmers do not induce large efficiency losses is an important consideration in fostering the sustainability of the CAP; this question, too, can be usefully addressed by RCTs. Finally, the potential for adverse selection is high in current schemes where farmers receive the same payment regardless of the opportunity costs of conservation they face. Here again, RCTs can be used for both eliciting individual willingness-to-accept an agri-environmental contract and for separately identifying the effects of farmer selection and payment on the provision of environmental services.

In this paper, we strongly advocate for experimentation, aiming to demonstrate that RCTs have significant potential for improving the design of the CAP. We do not view RCTs as a way of evaluating the CAP per se, but rather as a tool for evaluating different design alternatives for which there are no obvious ex ante expectations. This may be particularly attractive under CAP2020, since it will give even more leeway to member States to implement their own policy. We moreover argue that RCTs could be used as a tool to test new schemes with complementary interventions in an effort to enable the CAP to reach its target audience and objectives. For several years in France, the legislative framework has authorized social experiments that aim at testing the effectiveness of public policies through pilot programs (Gomel and Serverin, 2013) and in fact, a number of RCTs have already been implemented since then, in particular in the fields of education (Behaghel et al. 2017) and employment (Behaghel et al. 2014).

It cannot be denied that researchers and policymakers would face many real-world challenges when designing and implementing randomized evaluations within the CAP. Fortunately, lots has been learned about how to prevent and/or specifically account for such effects in trial design and implementation, and the accumulated experience and development of best-practice will be useful to address similar challenges to tackle key questions in the European agricultural and environmental context.

Sceptics of RCTs often point to the ethical problems that randomized experiments could create. Such issues can arise when randomized assignment implies creating a group of individuals who will be denied a program that is clearly beneficial, and who otherwise would have benefitted. For many components of the CAP, however, this question appears to be of less concern. For instance, consider agri-environment schemes: it has been more than twenty years since voluntary schemes to support the change of agricultural practices were proposed to French farmers and only a very small proportion of them finally took part in those programs (less than 6 per cent of them subscribed to an agri-environmental measure during the 2007-2014 CAP period[18]). Experimentations to increase uptake of programs that are already available to all does not raise obvious ethical problems.

The lack of randomized experiments in the field of agricultural policies may thus indicate the presence of political economic issues – such as the fear that the tool may do a disservice to its promoters and damage the credibility of CAP itself, rather than potential legal, technical or ethical obstacles. As such, it could be useful to follow the guidance offered by Campbell et al. (1966) in order to avoid such conflicts and misperceptions: real decisions often imply choosing between plan A and plan B rather than putting the global architecture of a policy in question; when such choices can be made following a documented and transparent process, the global architecture may be continuously improved, and the policy reinforced.

[18] This figure does not include measures related to grassland.

References

Athey, S. and Imbens, G. (2017), “The Econometrics of Randomized Experiments”, in Banerjee, A. and Duflo, E. (eds), *Handbook of Economic Field Experiments*, North-Holland, Volume 1, Pages 73-140.

Avvisati F., Gurgand M., Guyon N. and Maurin E. (2014). "Getting Parents Involved : a Field Experiment in Deprived Schools", *Review of Economic Studies*, vol. 81, n°1, 57-83.

Baird, S.; McIntosh, C. and Özler, B. (2011), 'Cash or Condition? Evidence from a Cash Transfer Experiment', *The Quarterly Journal of Economics* 126(4), 1709-1753.

Banerjee, S. (2018), 'Improving Spatial Coordination Rates under the Agglomeration Bonus Scheme: A Laboratory Experiment with a Pecuniary and a Non-Pecuniary Mechanism (NUDGE)', *American Journal of Agricultural Economics* 100(1), 172-197.

Banerjee, A., Chassang, S., Montero, S, and Snowberg, E. (2017) "A Theory of Experimenters", NBER Working Paper No. 23867.

Barrera-Osorio, F., Bertrand, M., Linden, L. L. and Perez-Calle, F. (2011), 'Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia', *American Economic Journal: Applied Economics* 3(2), 167-95.

Becker, G. M.; Degroot, M. H. and Marschak, J. (1964), 'Measuring utility by a single response sequential method', *Behavioral Science* 9(3), 226-232.

Behaghel, L., Crépon, B. and Gurgand, M. (2014) "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment." *American Economic Journal: Applied Economics*, 6 (4): 142-74.

Behaghel L., Crépon B., Gurgand M. and Th. Le Barbanchon, "Please Call Again: Correcting Non-Response Bias in Treatment Effect Models", *Review of Economics and Statistics*, vol. 97, n°5, 1070-1080, December 2015.

Behaghel, L., de Chaisemartin, C. and Gurgand, M. (2017) "Ready for Boarding? The Effects of a Boarding School for Disadvantaged Students." *American Economic Journal: Applied Economics, 9 (1): 140-64.*

Behaghel, L. and Gurgand, M. (2010) Programme expérimental "Bourse aux projets de classe" : bilan de la phase pilote du point de vue de l’évaluateur. Final report, May 2010, https://www.parisschoolofeconomics.eu/IMG/pdf/Pilote-BoursesProjets-PSE-juin2010.pdf.

Benhassine, N.; Devoto, F.; Duflo, E.; Dupas, P. and Pouliquen, V. (2015), 'Turning a Shove into a Nudge? A "Labeled Cash Transfer" for Education', *American Economic Journal: Economic Policy* 7(3), 86-125.

Berry, J.; Fischer, G. and Guiteras, R. P. (2015), 'Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana', CEPR Discussion Paper No. DP10703.

Campbell, D.; Stanley, J. and Gage, N. (1966), Experimental and quasi-experimental designs for research, R. McNally.

Carrell, B., Sacerdote, B, and West, J. (2013) "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, 81(3): 855 - 882.

Cason, T. N. and Gangadharan, L. (2013), 'Empowering neighbors versus imposing regulations: An experimental analysis of pollution reduction schemes', *Journal of Environmental Economics and Management* 65(3), 469 - 484.

Chabé-Ferret, S., Coent, P. L., Lefebvre, C., Préget, R., Subervie, J. and Thoyer, S. (2018), 'Can Nudges Induce Changes in Farmers' Choices of Agricultural Practices? Evidence from a Randomized Controlled Trial with French Winegrowers', CEE-M Working Paper.

Chabé-Ferret, S. and Subervie, J. (2013), 'How much green for the buck? Estimating additional and windfall effects of French agro-environmental schemes by DID-matching', *Journal of Environmental Economics and Management* 65(1), 12 - 27.

Chen, X.; Lupi, F.; He, G. and Liu, J. (2009), 'Linking social norms to efficient conservation investment in payments for ecosystem services', *Proceedings of the National Academy of Sciences of the United States of America* 106(28), 11812--7.

Colen, L., Gomez y Paloma S., Latacz-Lohmann U., Lefebvre M., Preget R., and Thoyer S., 2016, “Economic experiments as a tool for agricultural policy evaluation: Insights from the European CAP”, *Canadian Journal of Agricultural Economics*, Vol 64, No 4, pp 667-694.

Corrigan, J. R.; Depositario, D. P. T.; Nayga, Jr, R. M.; Wu, X. and Laude, T. P. (2009), 'Comparing Open-Ended Choice Experiments and Experimental Auctions: An Application to Golden Rice', *American Journal of Agricultural Economics* 91(3), 837-853.

Crépon B., Duflo E., Gurgand M., R. Rathelot and P. Zamora, (2013). "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment*", Quarterly Journal of Economics*, vol. 128, n°2, 531-580.

Deaton, A. and Cartwright, N. (2018). “Understanding and misunderstanding randomized controlled trials.” *Social Science & Medicine* 210: 2-21.

De Cara, S. D. and Jayet, P.-A. (2011), 'Marginal abatement costs of greenhouse gas emissions from European agriculture, cost effectiveness, and the EU non-ETS burden sharing agreement', *Ecological Economics* 70(9), 1680 - 1690.

Dessart, F.J., Barreiro-Hurlé, J., and van Bavel, R., “A behavioural approach to farmer decision-making: the case of sustainable agriculture”, European Review of Agricultural Economics, in this issue.

Duflo, E., Dupas, P. and Kremer, M. (2011) 'Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya', *American Economic Review*, 101(5).

Duflo, E., Kremer, M. and Robinson, J. (2011) 'Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence' *American Economic Review*, 101 (6): 2350-90.

European Commission (2013), 'Overview of CAP Reform 2014-2020', Agricultural Policy Perspectives Brief, DG Agriculture and Rural Development, Unit for Agricultural Policy Analysis and Perspectives.

European Union. Regulation (EU) No. 1306/2013. Available online: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:347:0549:0607:EN:PDF

Ferraro, P. (2008), 'Asymmetric information and contract design for payments for environmental services', *Ecological Economics* 65(4), 810 - 821.

Ferraro, P., Messer, K. D. and Wu, S. (2017), 'Applying Behavioral Insights to Improve Water Security', *Choices* 32(4).

Fiszbein, A.; Schady, N.; Ferreira, F.; Grosh, M.; Keleher, N.; Olinto, P. and Skoufias, E. (2009), Conditional Cash Transfers: Reducing Present and Future Poverty, The World Bank.

Fooks, J. R.; Higgins, N.; Messer, K. D.; Duke, J. M.; Hellerstein, D. and Lynch, L. (2016), 'Conserving Spatially Explicit Benefits in Ecosystem Service Markets: Experimental Tests of Network Bonuses and Spatial Targeting', *American Journal of Agricultural Economics* 98(2), 468-488.

Fryer, R.G., Levitt, S.D., List, J. and Sadoff, S., (2012). "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment", NBER Working Paper No. 18237.

Gocht, A. and Britz, W. (2011), 'EU-wide farm type supply models in CAPRI -- How to consistently disaggregate sector models into farm type models', *Journal of Policy Modeling* 33(1), 146 - 167.

Goldstein, N. J.; Cialdini, R. B. and Griskevicius, V. (2008), 'A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels', *Journal of Consumer Research* 35(3), 472--482.

Gomel, B. and Serverin, E. (2013), « L'expérimentation sociale aléatoire en France en trois questions », *Travail et Emploi*, 135 | juillet-septembre.

Guiteras, R. P. and Jack, B. K. (2018), 'Productivity in piece-rate labor markets: Evidence from rural Malawi', *Journal of Development Economics* 131, 42 - 61.

Herberich, D, S. Levitt and J. List (2009), 'Can Field Experiments Return Agricultural Economics to the Glory Days?', *American Journal of Agricultural Economics*, 91(5), pp. 1259-1265.

Hoffmann, V.; Barrett, C. B. and Just, D. R. (2009), 'Do Free Goods Stick to Poor Households? Experimental Evidence on Insecticide Treated Bednets', *World Development* 37(3), 607 - 617.

Holmstrom, B, and Milgrom, P., (1991) "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design", *Journal of Law, Economics, & Organization*, Vol. 7, pp. 24-52.

Imbens, G. (2018), "Comments On: Understanding and Misunderstanding Randomized Controlled Trials By Cartwright and Deaton", Stanford Graduate School of Business Working Paper 3648.

Jack, B. K. (2013), 'Private Information and the Allocation of Land Use Subsidies in Malawi', *American Economic Journal: Applied Economics* 5(3), 113-35.

Jayachandran, S., de Laat, J., Lambin E.F., Stanton, C.Y., Audy, R. and Thomas, N.E. 2017. "Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation", *Science*, Vol. 357, Issue 6348, pp. 267-273

Kuhfuss, L., Menu, M., Préget, R. and Thoyer, S. (2012). Une alternative originale pour l'allocation de contrats agro-environnementaux : l'appel à projets de l'Agence de l'eau Artois-Picardie. Pour, 213,(1), 97-105. doi:10.3917/pour.213.0097.

Kuhfuss, L.; Preget, R. and Thoyer, S. (2014), 'Préférences individuelles et incitations collectives : quels contrats agroenvironnementaux pour la réduction des herbicides par les viticulteurs ?', *Revue d'Études en Agriculture et Environnement* 95, 111-143.

Kuhfuss, L.; Préget, R.; Thoyer, S.; Hanley, N.; Coent, P. L. and Désolé, M. (2016), 'Nudges, Social Norms, and Permanence in Agri-environmental Schemes', *Land Economics* 92(4), 641--655.

Kuhfuss, L. and Subervie, J. (2018), 'Do European Agri-environment Measures Help Reduce Herbicide Use? Evidence From Viticulture in France', *Ecological Economics* 149, 202 - 211.

Lee, D. (2009), ' Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,' *The Review of Economic Studies*,76(3), 1071–1102.

Louhichi, K.; Ciaian, P.; Espinosa, M.; Perni, A. and Gomez y Paloma, S. (2018), 'Economic impacts of CAP greening: application of an EU-wide individual farm model for CAP analysis (IFM-CAP)', *European Review of Agricultural Economics* 45(2), 205-238.

Louhichi, K.; Ciaian, P.; Espinosa, M.; Colen, L.; Perni, A. and y Paloma, S. G. (2017), 'Does the crop diversification measure impact EU farmers' decisions? An assessment using an Individual Farm Model for CAP Analysis (IFM-CAP)', *Land Use Policy* 66, 250 - 264.

Lundberg, L.; Persson, U. M.; Alpizar, F. and Lindgren, K. (2018), 'Context Matters: Exploring the Cost-effectiveness of Fixed Payments and Procurement Auctions for PES', *Ecological Economics* 146, 347 - 358.

Manski, C; (2013) *Public Policy in an Uncertain World: Analysis and Decisions*, Harvard University Press.

Messer, K. D.; Ferraro, P. D. and William, A. (2015), 'Behavioral nudges in competitive environments: a field experiment examining defaults and social comparisons in a conservation contract auction', Bioecon Network.

Miao, H.; Fooks, J. R.; Guilfoos, T.; Messer, K. D.; Pradhanang, S. M.; Suter, J. F.; Trandafir, S. and Uchida, E. (2016), 'The impact of information on behavior under an ambient-based policy for regulating nonpoint source pollution', *Water Resources Research* 52(5), 3294-3308.

Miguel, E. and Kremer, M. (2004), Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, 72: 159-217.

Moffitt, R. (2001) "Policy Interventions, Low-Level Equilibria And Social Interactions." 45–82. MIT Press.

Morawetz, U. B. (2014). 'A concept for a randomized evaluation of agri-environment measures.' In *The Common Agricultural Policy in the 21st Century*, edited by E. Schmid and S. Vogel. Vienna, pp. 113–30. Vienna: Universitat fur Bodenkultur Wien.

Muralidharan K. and Sundararaman, V. 2011. "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, University of Chicago Press, vol. 119(1), pages 39 - 77.

Olken, B. (2007). "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115, 2: 200-249.

Poe, G. L.; Schulze, W. D.; Segerson, K.; Suter, J. F. and Vossler, C. A. (2004), 'Exploring the Performance of Ambient-Based Policy Instruments When Nonpoint Source Polluters Can Cooperate', *American Journal of Agricultural Economics* 86(5), 1203--1210.

Ravallion, M. (2018), "Should the Randomistas (Continue to) Rule?", Center for Global Development Working Paper 492.

Robles, M.; Rubio, M. G. and Stampini, M. (2018), 'Have Cash Transfers Succeeded in Reaching the Poor in Latin America and the Caribbean?', *Development Policy Review*.

Roe, B., and Just, D (2009), 'Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments and field data', *American Journal of Agricultural Economics,* 5, 1266-71.

Schubert, C. (2017), 'Green nudges: Do they work? Are they ethical?', *Ecological Economics* 132, 329--342.

Sitokonstantinou, V., Papoutsis, I., Kontoes, C., Lafarga Arnal, A., Armesto Andrés, A.P. and Garraza Zurbano, J.A. (2018) Scalable Parcel-Based Crop Identification Scheme Using Sentinel-2 Data Time-Series for the Monitoring of the Common Agricultural Policy. *Remote Sens*. 10, 911.

Spraggon, J. (2004), 'Testing ambient pollution instruments with heterogeneous agents'*, Journal of Environmental Economics and Management* 48(2), 837 - 856.

Suter, J. F.; Duke, J. M.; Messer, K. D. and Michael, H. A. (2012), 'Behavior in a Spatially Explicit Groundwater Resource: Evidence from the Lab', *American Journal of Agricultural Economics* 94(5), 1094-1112.

Suter, J. F. and Vossler, C. A. (2014), 'Towards an Understanding of the Performance of Ambient Tax Mechanisms in the Field: Evidence from Upstate New York Dairy Farmers', *American Journal of Agricultural Economics* 96(1), 92-107.

Thaler, R. and Sunstein, C. (2008), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press.

Vickrey, W. (1961), 'Counterspeculation, Auctions, and Competitive Sealed Tenders', *The Journal of Finance* 16(1), 8-37.

Wallander, S.; Ferraro, P. and Higgins, N. (2017), 'Addressing participant inattention in federal programs: A field experiment with the conservation reserve program', *American Journal of Agricultural Economics* 99(4), 914--931.

Westerink, J., Jongeneel, Polman, Prager, K., Franks, J. , Dupraz, P., and Mettepenningen, E., 2017. "Collaborative governance arrangements to deliver spatially coordinated agri-environmental management," *Land Use Policy*, 69 : 176-192.