OECD Workshop

on the

Evaluation of Agri-environmental Policies

20-22 June 2011
The Johann Heinrich von Thünen Institute,
Bundesallee 50, 38116 Braunschweig, Germany

Monday, 20 June

Session 1: 09h30-11h15

Keynote Address

Evaluating Policies for Delivering Agri-environmental Public Goods

Alison Burrell

Introduction

Over the last two decades, policies for stimulating the provision of environmental public goods in rural areas have claimed a growing share of government money allocated to agriculture in OECD countries. In parallel with this trend, there has been a sharp rise in the number of formal evaluations of these policies commissioned by administrations, often within a legally defined timetable that stipulates regular policy appraisals[1]. At the same time, the number of more *ad hoc* policy evaluations conducted by academics and NGOs has also increased markedly.

This Workshop comes at a good moment for examining and comparing the approaches taken to evaluating agri-environmental policies in different countries, and for exploring in more depth the issues that arise in the specific agri-environmental context.

This paper aims to provide a conceptual framework for the Workshop discussion. At the request of the organisers, it begins by restating the formal definition of a public good and discussing how it has been applied to the agricultural sector in developed countries. The definition of an agri-environmental public good used in this paper is also given.

The perspective taken here for discussing the main methodological and analytical challenges arising when evaluating agri-environmental policies is a comparative one. However, instead of comparing evaluation methods between countries or types of intervention, it contrasts three stylised types of evaluation activity – what I call the administrative, the scientific and the economic approaches to evaluation. Each has its role and purpose, and each gives rise to specific methodological problems. The aim here is not to establish a hierarchy among these three approaches, but to explore how each of them might perform its role more efficiently and usefully. Concluding remarks are intended to stimulate discussion on future improvement strategies.

Definition of a public good and implications for its provision in a private market

A public good has the following characteristics: it is non-rival[2] and non-excludable[3], and is valued by individuals. For some public goods, the presence of the first two properties may depend on the context of their provision and use. For example, the property of non-rivalry is lost when the good is so heavily consumed that congestion or over-use begin to reduce its availability to others. Some public goods can be made excludable albeit at a very high cost.

A public good will not be provided by an individual for his own private use unless his private benefit exceeds the cost to him of providing it. Many of the 'classic' public goods involve massive

---

[1] For example, for some years the EU's rural development legislation has required independent *ex ante*, mid-term and *ex post* evaluations of all its co-funded rural development programmes. Evaluations are performed at the level of individual programmes. At a more synthetic level, Oréade-Brèche (2005) provides an overviewing appraisal of EU agri-environmental measures for a 10-year period. Policies for assisting agriculture in the EU's less favoured areas, which also have strong environmental objectives, were assessed at EU level in 2006 (see IEEP, 2006).
[2] A good is non-rival when its consumption by one individual does not reduce the quantity available for consumption by other individuals.
[3] A good is non-excludable if, when it is produced, no one can be prevented from using it.

infrastructure and prohibitive operational costs. Examples are a national highway system or a television network. Although the public good services they offer are valued by each of us as individuals, no single individual will provide them for himself. And if he were able to do so, their non-excludability would create a multitude of free-riders who would also benefit from them without sharing the cost. Goods with these characteristics, sometimes also called 'collective goods', are typically funded by government out of tax revenue and provided free of charge at point of use.

An alternative method of collective provision for smaller-scale public goods occurs when a number of individuals agree among themselves to finance the provision of goods or services with public-good characteristics for the exclusive use of the group as a whole. Such an arrangement implies that excludability vis-à-vis non-members of the group can be enforced. Goods provided in this way are called 'club goods' and are often cited to demonstrate that reliance on government for public good provision is not always necessary.

As well as these stand-alone public goods, which are not provided at all without a collective decision to do so, public goods may also occur as positive externalities of other economic activities, where their provision is a by-product of the decision to produce a private good. In this case, the public good will be produced without anyone having taken the decision to produce it or having to pay for its provision. However, it will be provided in quantities that are sub-optimal from the social point of view, since the benefit of the externality that is enjoyed free of charge by others is not taken into account by the economic decision-maker.

Table 1 summarises these two cases.

Table 1: Conditions for private and optimal provision of a public good

Fully separable (stand-alone) public good i	Public good i occurring as a positive externality linked to production of good j
Not produced as a private good at all unless $PB^i \geq PC^i$	Will be produced as long as $PB^j \geq PC^j$, but at a socially sub-optimal level
Provision as a collective good is socially optimal as long as $\Sigma_k PB^i_k \geq PC^i$	Socially optimal level of provision occurs when $PB^j + \Sigma_k PB^i_k = PC^j$
PB= private benefit, PC= private cost. It is assumed that K individuals (*k*=1,…,K) derive individual benefit from the public good. In both cases, the condition for socially optimal provision balances the social benefit (sum of the benefit of all individuals) against the cost of provision. The condition for the stand-alone public good assumes that it is of a fixed size, and compares total social benefit with cost of provision. The condition for determining the optimal provision of the public-good externality (and hence also of the provision of the private good j) should be interpreted as a *marginal* condition: provision should be expanded up to the point where the marginal social benefit (which is the marginal private benefit to the producer of producing j + the sum of the marginal benefits derived by third parties from externality i) equals the marginal cost of producing the private good j.	

Public goods in agriculture

It has been known for generations that agriculture produces public goods. Nineteenth century poets extolled the countryside as a place for peaceful reflection and restoration of the soul. For decades, a favourite textbook example of a reciprocal positive externality has been an agricultural one: that of the beekeeper whose bees pollinate (free of charge) the fruit trees of the neighbouring fruit grower, whilst the fruit grower's orchard provides (also free of charge) the nectar needed for the production of honey. The value of the benefit of the externality conferred is not taken into account in the decision-making of the honey producer or fruit grower, but the benefit of the externality received is crucial for the viability of both private production activities.

At the end of the 1990s, the policy debate surrounding the multifunctionality of agriculture gave prominence to the idea of public goods as positive externalities of agricultural commodity production. The proposition was that the existence of these socially valued but non-remunerated joint products of agricultural production activity justified supporting the price of farm outputs above the minimum cost of getting them onto the market (whether from domestic or imported production), and that to reduce this support would – by precipitating a decline in agricultural commodity production – also reduce the provision of the valued public goods. Conversely, if society wished to have more of these positive externalities, the rate of support to commodity production should be increased. This view makes sense only if the public good externalities are joint, non-separable outputs with agricultural commodities such that a subsidy for commodity output results in both more output and more public good.

The multifunctionalists' position in favour of farm price support was progressively abandoned for two reasons. *First,* it turned out that there are actually very few examples of positive externalities whose production is technically interdependent with that of agricultural commodities. In fact, it is much easier to demonstrate jointness between commodity production and negative externalities[4]. For example, at current production levels in developed countries, an increase in commodity output is far more likely to result in more nitrate pollution and soil erosion than in more biodiversity[5]. *Second*, market price support is a subsidy coupled to production, and hence it is in the WTO amber box of restricted-use measures. It was not likely that other WTO members would agree to make an exception for market price support purporting to remunerate the provision of public good externalities.

These two reasons largely explain the move away from seeing agricultural public goods as externalities of commodity production. In fact, they underpin the idea, already operational in the agri-environmental measures adopted in some OECD countries in the early 1990s, of treating agriculture's public goods as separable 'stand-alone' outputs whose provision can be directly

---

[4] See, for example, Hodge (2008).
[5] As with most generalizations, this one also is not universally true. Many OECD countries have policies in place for maintaining extensive, low-input farming in marginal areas where the abandonment of agricultural activity may well increase soil erosion and reduce wildlife populations. Indeed, the heterogeneity and site-specificity of the link between agricultural activity and environmental values is another reason why blunt policy instruments such as price support are inappropriate for stimulating public good provision in agriculture.

stimulated by targeted policies. The absence of technical interdependence between agricultural public goods and commodity production allows policy interventions to encourage public good provision but *without* creating incentives to increase commodity production. This offers scope for agri-environmental measures that will not be contested as trade-distorting.

There has been much discussion about what exactly qualifies as an agricultural public good[6]. On the short list are items such as wildlife conservation and biodiversity (including preservation of rare of farm animal breeds), maintenance of landscape quality and character, protection of natural resources (soil, water and air quality), flood control and carbon storage. All these items can uncontroversially be characterised as public goods and all of them are associated with benefits to the rural environment. More debate on what counts as an agricultural public good has surrounded various items on the long list, such as the maintenance of rural populations and rural economic activity, national food security and animal welfare. Since these more controversial items fall outside the focus of this paper anyway (they are not *environmental* public goods), it is not necessary to rehearse the debate here. Rather, it is more useful to state explicitly the definition of agriculture's environmental public goods adopted in this paper.

The agri-environmental public goods that are targeted by the policies whose evaluation is the subject of this paper are *non-rival*, *non-excludable* outputs or services that are *valued by society*. Precisely because of their non-rivalry and non-excludability, and despite society's positive willingness to pay for them, *markets do not exist* for them. Their provision enhances and preserves the natural rural environment, local ecosystems and the natural resource base linked to farming. Another important defining characteristic for the purposes of this paper is that their supply can be increased through the actions of farmers – that is, farmers typically have both the *property rights* over the use of the resources needed to supply these agri-environmental public goods and the *requisite skills* for managing their provision alongside their farming activity.

Although this definition covers the quite heterogeneous short list of public goods presented above, it is nevertheless quite narrow. For example, it rules out what might be seen more as *consequences* of farming activity (such as keeping population in rural areas or contributing to national food security), which do not depend on the actions – other than producing food - of any single farmer, but rather on the total amount of farming activity taking place or the number of farmers engaged in it. Under the definition used here, the provision of public goods depends on decisions taken by individual farmers to do things other than producing food, or to produce food in a certain way.

Policies for delivering agri-environmental public goods

Interventions that aim to improve the environmental footprint of farming can be classified according to the mechanism used, and by whether they merely ensure the basic environmental standards expected by society or encourage farmers to go beyond those standards.

---

[6] For recent discussion, see Cooper *et al.* (2009), McVittie *et al.* (2009).

Regulations that prescribe minimum levels of environmental performance or ban certain environmentally harmful practices, and that apply to all farmers, set the parameters for the basic standard of environmental preservation that society expects as a norm. In the following discussion, these measures will be treated as part of the counterfactual against which the impact of policies for delivering environmental public goods should be measured. Regulatory norms stand apart from the other types of intervention identified below in that no public money is devoted to incentivising farmers to adopt or comply with them. Instead, compliance is achieved by the threat of legal sanctions, in the spirit of the polluter pays principle. Therefore, they are not included among the policies whose evaluation methodology is examined in this paper. This does not mean that the environmental effectiveness of these regulations should not be monitored on a regular basis.

Cross compliance measures require farmers to refrain from or to adopt particular farming practices, for the most part motivated by environmental considerations, as a condition of their receipt of income support payments. Debate has raged in the past over how best to describe these payments. Are they really incentive payments targeting environmental benefits, to which farmers respond or not depending on whether they wish to make more room for public good provision in their farming activities? Or does the main objective of these payments remain blanket farm income support that is *also*, and almost incidentally, used to obtain some leverage over the choice of farming methods and bring producers up to a common level of good agricultural practice?

Fortunately, this issue does not have to be settled here. What is clear is that cross compliance requirements cannot be considered as part of the benchmark defining the minimum environmental performance level expected from everyone by society. This is because, *first*, the payment system to which cross compliance is attached does not always apply to all farmers or farming activities, and *second*, there is the (theoretical) possibility that eligible farmers can opt out of the payment system and thus avoid compliance. However, for farmers within a cross compliance scheme, these requirements have to be considered as part of the counterfactual against which additional incentive measures for further enhancing environmental performance should be measured.

Agri-environmental payments are measures designed to target the provision of agricultural public goods beyond the level defined by the regulatory benchmark norms. Generally, these publicly funded schemes are voluntary for farmers and involve them opting into a contract of some years' duration. For compliance with the WTO definition of a green box measure, the payment should be set at a level that merely compensates the farmer for extra costs incurred, including income foregone. In practice, the payment might be determined according to the real costs on each farm, the estimated cost on the 'average' farm, or (quite rarely) the expected environmental benefit. Each of these options has implications for the amount of farm-specific information needed in order to set up the contract, as well as for the profile of the typical farmer-volunteer – that is, whether volunteers tend to be those whose participation would secure the greatest environmental benefits or those whose opportunity cost of participation would be lowest. There may be poor uptake of the measure and hence low density of coverage, which can reduce the overall environmental impact) if the payment is set too low, or risk of over-compensation if it is set too high.

Schemes for creating markets in environmental externalities are relatively new initiatives that follow the general cap-and-trade principle. Two examples are water quality trading schemes[7] (mainly in the US, but also represented in Australia, Canada and New Zealand), and the scheme for trading 'mineral rights' (effectively, nitrogen and phosphorous surpluses or deficits relative to an allowable farm-level surplus) that has operated for over a decade in the Netherlands[8]. Of course, it is the collective 'cap' and its binding allocation to individuals that provides the environmental benefit. However, the trade allows the total economic cost of applying the cap to the polluting economic activities to be minimized, or in other words permits the greatest environmental benefit for a given cost to the polluting industries. Government participation in these programmes involves, at the very least, putting in place the overall regulatory framework required for imposing the cap. Typically, some financial assistance in setting up the trading scheme is involved, and maybe on-going assistance with running costs. How to evaluate such interventions is not explicitly discussed in this paper. However, these programmes should also be regularly evaluated. Clearly, many of the issues touched on in this paper are relevant to their evaluation as well.

Definitions of terms

Since markets do not exist for public goods, they have no observable price. Hence, their value in money terms has to be estimated. But the *valuation of a public good* is only one element of the *evaluation of a policy* designed to deliver it, and one that is often not performed in practice. Policy evaluation assesses a sequence of changes and their consequences occurring in different dimensions, at different points in time, and at different degrees of 'remoteness' from the initial policy stimulus, in order to establish what effect it has had.

Our discussion of this sequence adopts the terminology used by the European Commission's framework for evaluating rural development policies (CMEF)[9]. Other evaluation approaches adopt analogous classifications, but the terminology may differ.

The consequences of the policy intervention can be classified as:

- *outputs* (actions financed and accomplished with the money allocated to the intervention) (e.g. number of farm visits, enrolment of participants);

- *results* (direct consequences for programme participants) (e.g. changes in farm practices, participation in training);

- *impacts* (longer-lasting effects of the intervention with direct or indirect relevance to the programme's long-term objectives and their attainment) (e.g. improved water quality, reappearance of skylarks).

---

[7] See Selman *et al.* (2009).
[8] See Komen and Peerlings (1998), Peerlings and Polman (2008).
[9] Common Monitoring and Evaluation Framework (European Commission, 2006).

This classification can be used in the context of both *ex ante* and *ex post evaluation*. Most of the discussion in this paper relates to *ex post* evaluation.

Ex post evaluation involves evidence-based judgements and recommendations. A major task of the evaluator is therefore to assemble the available evidence on the consequences of the policy. The selection of relevant evidence should be guided by a conceptual 'model' of the various causal pathways through which the intervention is expected to operate, the key links in the expected chain of effects and the outcomes expected. This conceptual roadmap (the *intervention logic*) is implicit in the design of the policy and would have already been elaborated and tested for internal consistency during an *ex ante* evaluation of the intervention. Its role *ex post* is to guide the evaluator in his choice of what evidence to review and what criteria to use for assessing the success of the programme.

Three stylised approaches to evaluation

The large and growing volume of evaluation studies can be classified into three main categories according to why and by/for whom the evaluation is performed, which in turn tends to determine the approach taken, the type of information provided and the challenges faced. This classification is presented in Table 2. Three stylised groups of stakeholders/practitioners are also involved.

Administrators and officials of public bodies directly responsible for a policy need fairly rapid feedback on its progress so as to correct mistakes at an early stage. They also need information on its first-round results in order to accommodate the relatively short time-horizon of the political process. Their needs are closely driven by their institutional setting.

Scientists (ecologists, environmentalists, biologists, water scientists, landscape ecologists and so on) are interested in uncovering the causal dynamics of the natural processes that the intervention is intended to affect, and the causal links between policy and outcome. An environmental policy intervention can be seen as an experiment that provides an excellent opportunity for them to gain scientific knowledge about underlying relationships in the natural world, and about the speed and sensitivity of responses inherent in these relationships. At the same time, the information they acquire through this investigation is extremely valuable for improving the policies themselves at a later stage. Given the biological lags involved and the slow cumulative nature of some of the responses expected to agri-environmental interventions, it makes little sense to evaluate the full impact until an appropriate number of years have elapsed since the start of the policy. Of course, this does not fit well with the needs and the timetable of the administrators.

Finally, *economists* explore the social justification for a policy by examining the trade-off between its benefits and costs in a society-wide context. The validity of any conclusion they can draw depends on the prior confirmation of causality. That is, it only makes sense to compare benefits with costs if those benefits have indeed been produced by the policy intervention. As well as the expected direct and indirect consequences of the policy, the economist must also value in money terms any unintended side-effects. An 'economic' evaluation needs input from both the other types of

evaluation in Table 2. When done rigorously, it involves the longest delay with respect to the start of the policy. Unfortunately, this does not shield economists from pressures to evaluate policies far more rapidly.

Table 2: Three evaluation stereotypes, according to methodological approach

	Type of evaluation		
	'Administrative'	**'Scientific'**	**'Economic'**
Purpose	• account for the use of public money • verify that the targets as specified by the policy objectives were achieved • streamline the delivery of policy	• improve understanding of natural systems and causal mechanisms • provide advice on instrument choice and ways of improving the impact of policy mechanisms	• examine social justification of policies • analyse behavioural responses to policies • advise on improving overall policy design (including incentives, value for money)
Timing	Short term (1-4 years after the start of the policy)	Longer term (5 or more years after the start of the policy)	Ideally, requires input from a scientific evaluation, so >5 years, but pressure to provide full economic evaluations more rapidly
Focus	• monitoring of outputs, results • ideally would also assess impacts, but often not feasible given time frame	• existence and magnitude of impacts • causality between policy and impacts observed	• social value of impacts and their relationship to full social cost • conclusions depend on prior confirmation of causality
Approach	• use of indicators measuring changes in key variables over the evaluation period	• detailed measurement of environmental changes • scientifically rigorous attempts to establish causality	• various approaches to valuing environmental benefits • cost-benefit analysis • simpler methods for comparing costs and benefits
Issues and challenges	• need to collect baseline data before the policy is implemented • additionality • time horizon may be too short to verify desired impacts • can results be good predictors of impacts?	• availability of baseline data, research design • high resource cost of field data collection • what is the counterfactual/ control? • uncertain dynamics in the causal chain • scaling up from surveys to whole population, from individual farms to region	• valuation of benefits in money terms • importance of measuring policy transaction costs • need to identify and value unintended consequences • need to assume time horizon for future cost and benefit flows • choice of method for comparing costs & benefits

While waiting for the scientific evidence on whether environmental benefits have been achieved, economists often study society's willingness-to-pay (WTP) for the expected environmental benefits

on the hypothesis that these benefits are indeed produced. It is reassuring for the administrators if WTP turns out to be greater than what has been allocated or spent from public funds. But this hardly takes us beyond an *ex ante* evaluation, and is certainly not a substitute for an evidence-based appraisal of whether the policy actually changed anything in the real world.

Inevitably, the various practitioners and users of agri-envronmental evaluations do not always share each others' point of view or fully appreciate each other's priorities. The delay that is usually necessary for a rigorous scientific or economic evaluation does not fit well with the administrator's timetable. Scientists and economists may be dismissive of some of the cruder assumptions made in a more rapid evaluation conducted for administrative purposes, failing to realise that maintaining the momentum of the policy process is essential to change, and that policy-making routinely involves more uncertainty and a higher margin of error than scientific research.

However, despite some occasional friction, these groups tend to share some common goals. Moreover, within *each* of the three stereotypic evaluation approaches shown in Table 2, there are currently a number of issues and areas where issues still have to be resolved and challenges have not yet been met. The following sections look at the most important of these challenges for each type of evaluation.

Current challenges and issues

'Administrative' evaluation

The main constraint faced by the administrative evaluation is its limited time frame[10]. This poses several problems for the evaluation of agri-environmental schemes. *First*, it affects the type of data that can be used. Berriet-Solliec *et al*. (2011) point out the hierarchical nature of different types of data, in terms of reliability and quality, which they consider begins at the lowest level with opinion surveys (where the respondents may be participants, experts or official authorities), followed by historical or geographical comparisons, and several levels higher, culminates in the results of randomised trials. It seems inevitable that the sooner the evaluation is performed, the lower the availability of 'high reliability' data.

Second, many of the environmental benefits targeted require a number of years to emerge. For example, targeted increases in bird populations are reported from various studies in the UK for periods taking 7-10 years, whereas water and soil quality improvements resulting from policy interventions may need some decades (Boatman *et al.*, 2008). Kleijn (2006) refers to evidence that 'on intensively used farmland, the restoration of species-rich communities following the re-instatement of more extensive management may take considerably longer than six years'. These long

---

[10] Boatman *et al*. (2008, p.121) describe how in the UK during 1992-2003 ('a period of intense experimentation with these schemes') the evaluation focus shifted from longer-term reviews to shorter-term monitoring and 'early, indicative performance measures'. In their view, 'this has reduced our ability to actually measure scheme impacts for UK agri-environmental schemes'. However, it may also have allowed lessons to be learned more quickly and hence benefited scheme design and delivery. They also note that, because of this shift, more weight is given to expert judgement and 'other predictive methodologies' and less to repeat surveys as sources of 'evidence'. These observations appear to be applicable also beyond the UK.

time-lags mean that administrative evaluations have to rely extensively on indicators capturing outputs and results, since evidence of longer-term environmental benefits tends to fall outside the time frame.

The important issue of additionality[11] arises first in this chain of linked effects when interpreting evidence on the results (first-round direct effects) of the policy. It has a major bearing on the cost effectiveness of the programme: when additionality is low, then a large part of the payment costs was not necessary to achieve the results. Such a finding cannot be neutralized or offset by subsequent information on impacts. Even when the links in the next stage of the causal sequence – those between results and longer-term benefits – *are* verified, they cannot be attributed to the policy if the changes in land use choices, farm practices, attitudes and so on that led to the increase in benefits would have happened in any case without the policy stimulus.

In their survey of over 200 evaluation studies, Boatman *et al*. (2008) report various estimates of additionality where the source of evidence is mainly farmer (participant) interviews or geographic comparisons. The rates of measured additionality vary enormously between projects. It is interesting to note that farmer interviews produced both high and low estimates, suggesting that this source of evidence does not necessarily yield biased results. More generally, uncertainty regarding additionality is just one problem caused by the lack of a counterfactual scenario showing what would have happened in the absence of the policy (this is discussed further below).

The 'administrative' evaluation relies on evidence supplied in the form of indicators measuring the current situation of the variables targeted in terms of results and possibly also impacts, and their changes since the start of the policy. Oñate *et al*. (2000) define three key operational attributes for agri-environmental indicators: *reliability* (extent of scientific support for the relationship between the indicator variable and the targeted environmental benefit), *relevance* (degree of pertinence of the indicator variable and its associated benefit to the policy goal), and '*realisability*' (data availability).

A pre-requisite for the meaningful interpretation of indicators is the availability of baseline data measuring the levels of those same indicators prior to the start of the programme. In their assessment of an agri-environmental programme for maintaining and enhancing dry-cereal landscapes in north western Spain, Oñate *et al.* (2000) found that availability of reliable and relevant data from existing sources was severely limited by problems of scale and content, and that a great deal of new data had to be collected at farm level. In these circumstances, the construction of a baseline retrospectively is not easy and may not produce reliable results. Ten years later, I wonder how much this situation has changed. Until collection of appropriate (in terms of content and level of observation) baseline data is included as a *necessary initial step* in the setting up of all agri-environmental programmes, the full value of indicators as evidence will not be realized.

The implications of a short time-frame for the type of evidence that can be considered prompt the question: Can results be good predictors of impacts? Clearly, a minimum requirement for this is the

---

[11] Additionality refers to the extent to which the observed changes (e.g. behavioural changes of farmers enrolled in the scheme) are due to the programme, or whether they would have occurred in any case.

availability of robust prior evidence on the causal links between on-farm changes and environmental impacts, and this is often lacking. Kleijn (2006) states that for most EU schemes targeting biodiversity, knowledge on the causal links between various management prescriptions and species abundance does not exist. Programme design is 'largely inspired by traditional agricultural practices that used to be associated with high levels of biodiversity'. When this is the case for a targeted environmental benefit, the scope for predicting from results to impacts may be severely limited.

In those cases where substantial scientific evidence exists on the causal links between results (changes in practice) and impacts (long-term environmental benefits), mathematical models can be used to simulate potential benefits[12]. However, a caveat is needed here: what has been observed in the general case may not hold for a given area with particular local characteristics Site- and area-specific factors are of major importance for determining the effects of management practices on many targeted benefits[13]. Therefore, unless the model is sophisticated enough to allow the 'average' causal relationship to be modified through the inclusion of other conditioning variables describing site/regional characteristics, model predictions may be quite misleading.

A final question then is the following: when changes in the appropriate indicators of results are available, measured correctly from the relevant baseline data, together with an intervention logic based on proven replicated evidence on longer-term causal links, is this sufficient to predict that environmental benefits will be enhanced? The answer is already given above: any prediction made on this basis may well be valid for the 'average situation' yet be quite unrepresentative for the area or type of farm where the intervention has been undertaken[14].

The above discussion supports the recommendation of Boatman *et al.* (2008, p.122): 'We suggest that, whilst the short-term indicative evaluation methods are undoubtedly useful for the policy review cycle as well as directing ongoing learning and experimentation, they should be used in conjunction with, and not as replacements for, direct and longer-term evaluation of scheme impacts on the ground'.

'Scientific' evaluation

The focus of this type of evaluation is to verify the existence of intervention impacts (environmental benefits), measure their extent and establish causality between the intervention and increases in benefits. Kleijn and Sutherland (2003) reviewed 62 *ex post* evaluations investigating the impacts of EU agri-environmental programmes that targeted biodiversity. They concluded (p.947) that 'in the majority of studies, the research design was inadequate to assess reliably the effectiveness of the schemes. Thirty-one per cent did not contain a statistical analysis. Where an experimental approach

---

[12] See, for example, Boatman *et al.* (2007).
[13] For example, even 'the location of a buffer strip in relation to the pollutant source or pathway is of paramount importance in determining its performance in protecting water quality' (Boatman *et al.,* 2008, p.106)
[14] Pearce (2005) argues strongly against using 'results' as proxies for 'impacts' in agri-environmental evaluations. For him, results are to be treated as a *means to an end*, not the end itself. He writes (p.76) that 'Failure to make this distinction has led to some confusion in the literature'.

was used, designs were usually weak and biased towards giving a favourable result.' A common reason for bias in this type of study is the use of an inappropriate counterfactual.

The counterfactual is the situation that would have prevailed without the policy. The rigorous identification and measurement of policy impacts requires that it is done by comparing performance within the programme against the counterfactual. For policies like statutory regulations that are compulsory for everyone without exception, the counterfactual is a wholly hypothetical state of the world that has to be reconstructed using a model and with the aid of a number of assumptions. By contrast, for voluntary measures where the uptake is partial, some farmers remain outside the scheme. At first sight it might seem that their behavioural choices and the environmental benefits they generate could provide the counterfactual for the farms in the programme. However, this choice creates what is known as *selection bias*. Without proof to the contrary, farmers opting into the scheme did so because they had certain characteristics relevant to the scheme that are not shared by farms remaining outside. Perhaps they already had higher environmental performance on their farms, or were situated in areas more conducive to realising environmental benefits. The difference between their performance and that of non-scheme farms confounds two effects: the impact of the scheme *and* the underlying differences that determined their choice to opt in.

Another simple but invalid choice of counterfactual is the pre-policy situation of those farms opting in. But here too the difference between their situation at the start of the programme and x years later confounds two types of impact: that of the policy and the combined effect of all the other changing factors that may have influenced environmental outcomes on their farm over time (other policies, (micro-)climate change, unusual weather effects and so on). The problem here is that, in this kind of non-experimental context, it is not possible to impose *ceteris paribus* (that is, hold other factors constant) over the time period in order to isolate the effects of the policy.

Selection bias might be greatly reduced by collecting baseline data for farms in and outside schemes and comparing the *changes* between the two groups over a number of years, or comparing changes in biodiversity between areas with and without schemes, or by pairing with-scheme and control sites having similar environmental conditions for comparison, or by using statistical methods like propensity score matching[15] to create an artificial control group.

Kleijn and Sutherland (2003) go further than any of these strategies, however, in proposing the following ideal experimental design: the designation of control sites that are similar to scheme sites in every respect except for the changes obligated by the programme, ensuring 'sufficient replication' of both types of site, and the collection of baseline data for both. Boatman *et al.* (2008) challenge the feasibility of this proposal, not least because of the difficulty of finding suitable control sites that are neither coercively excluded from the programme nor atypical in respect of characteristics that are relevant for the programme.

---

[15] Propensity score matching is recommended when selection bias is present in non-experimental data where there are few 'natural pairings' between the with-scheme and control groups, and/or natural pairings are difficult to identify because of a large number of potentially relevant pre-treatment characteristics. However, it is not without its own technical problems.

Another problem affecting the scientific measurement of environmental impacts of agri-environmental measures is that the *causal dynamics* between initial management change and (full) environmental impact is often unknown. Initial responses may be delayed for several years and response paths are unlikely to be linear. This raises the question of *when* and *how often* to monitor environmental impacts on farms within an agri-environmental programme. Various studies recommend that not only should baseline data be collected before the scheme starts but also that repeated surveys should be carried out over the life of the scheme (and maybe beyond). Regular monitoring would of course help to establish the time profile of the environmental impact, which would contribute to knowledge and may lead to better policy design in the future. However, it needs to be recognized that frequent on-farm monitoring, not to mention the technical and practical difficulties inherent in measuring certain environmental changes in the field, place a heavy data-collection burden on research budgets.

To control cost, data collection is usually carried out for a sample of farms. This raises the issue of scaling up from the sample to the whole group in order to measure the total effect of the programme. Unless the sample has been selected randomly or using a stratified sampling technique, the appropriate set of raising factors (weights) for aggregating benefits over all farms in the scheme is not a straightforward issue. In any case, it may well not be appropriate to apply standard weights like farm area when aggregating environmental impacts.

Finally, according to the definition adopted in this paper, the decision on whether and how to produce agri-environmental public goods is taken by the individual farmer. This implies that policies to promote them can act through incentives for each farmer to modify his management and resource use decisions. It also implies that evaluation of policy impacts should look at the link between individual decisions and the resulting environmental benefits. However, there is a also collective element in the provision of many agri-environmental public goods. Critical mass or synergies between the decisions of individual farmers, the extent and manner in which they are coordinated or the spatial design of the intervention within an area – such factors may also determine the effectiveness of individual farmer decisions.

For example, maintaining a given landscape usually requires the participation of a number of farmers in the same small area, as does improving water quality in a watershed. An individual farmer's contribution to preserving certain wildlife populations depends in part on design features such as the provision of wildlife corridors over wider tracts of neighbouring land. The value of schemes to improve public access across farmland can depend on whether paths are coordinated to link with and extend existing local path networks (Boatman *et al*. 2008). Scientific evaluations based on individual farm data need to find a way of incorporating such synergistic effects: are they simply an exogenous contextual element to be controlled for, or are they endogenous aspects of the policy that should be included in what is to be evaluated?

'Economic' evaluation

An economic evaluation introduces monetary values into the evaluation framework. The most comprehensive type of economic evaluation is a cost-benefit analysis (CBA), in which all present and future costs and benefits arising from the intervention are valued in money terms[16]. These monetary flows, which are distributed over time, are converted to their equivalent value at a single point in time and aggregated[17]. If the net value of the aggregate is positive, then the policy is judged to make a positive contribution to social welfare. The information obtained from 'scientific' evaluations about the size and flow of benefits over time is clearly an important input into this calculation.

This ambitious methodology poses various challenges. *First*, agri-environmental benefits, which by their nature are not marketable and hence have no easily observed market price, have to be valued. Economists have been busy with this issue for decades. Currently, the two most frequently used approaches are *stated preference* (SP) methodologies (survey respondents are asked directly for the value they attach to particular non-marketed benefits)[18] and *revealed preference* (RP) techniques. Among the latter are approaches where actual choices are analysed in order to infer the value of associated unpriced benefits (e.g. the travel cost method, hedonic pricing) and choice experiments. Choice experiments ask participants to choose between (usually hypothetical) options in which different levels of non-marketed attributes are 'bundled'. When the resulting data are processed using complex econometric techniques, the implicit value attached to each attribute can be extracted. Both SP and RP approaches have strengths and weaknesses, as well as their own partisans and detractors. These methodologies are still being refined in order to improve general performance and flexibility in particular situations.

Second, a social CBA requires all costs and benefits arising from the policy to be included. This involves not only direct programme costs (agri-environmental payments themselves), but also so-called policy-related transaction costs (PRCTs) (e.g. administrative costs, legal costs of setting up agreements with farmers and non-remunerated farmer costs like the opportunity cost of extra time spent on administrative and recording tasks)[19]. Studies have found that these other cost components can be high relative to size of the payment, and may depend quite strongly on the design and delivery of the programme.

Third, all unintended consequences of the policies - whether beneficial or harmful – need to be identified and valued. For example, if landscape rehabilitation due to an agri-environmental programme stimulates farm tourism in the area, which generates multiplier effects in the local

---

[16] For a recent example, see Boatman *et al.* (2010).
[17] They may be aggregated in unweighted form, or with distributional weights attached to particular items that reflect society's preferences between social groups who may benefit from the policy or have to bear its costs.
[18] See, for example, Pruckner (1995), Loomis *et al.* (2000) for studies based on stated preferences (contingent valuation methodology), and Hanley *et al.* (1998), Scarpa *et al.* (2007) for studies using choice experiments.
[19] For some relevant literature, see OECD (2006), Falconer and Saunders (1999), Nilsson (2007), Mettepenningen *et al*. (2009).

economy, this should also be included as a benefit since it will not be covered by the valuation of the landscape improvement *per se*[20].

Fourth, an assumption has to be made about what happens to the flow of benefits at the end of the agri-environmental contract (Pearce, 2005). The most optimistic assumption is that the farm continues to be managed as when it was under the contract; the most pessimistic is that farm practices and land use choices quickly revert to their pre-contract status. For contracts of long-duration and if a high discount rate[21] is used, the choice hardly makes any difference. For a 10-year contract and a discount rate of, say 2%, it will affect the calculation.

The difficulty of imputing a monetary value to non-marketed benefits increases the appeal of two other techniques for economic valuation, namely cost effectiveness analysis (CEA) and multi-criterion analysis (MCA) (Pearce, 2005). In both cases, economic information is used for costs, which are compared with benefits measured in some physical unit. For example, CEA applied to a programme targeting biodiversity maintenance might measure the increase in hectares of High Nature Value (HNV) farmland, or the increase in nesting pairs of a threatened bird species, per euro of cost. This criterion only makes sense for single-impact programmes, or where multiple impacts can be aggregated into a meaningful total.

MCA is used for multi-objective programmes. Experts give scores to a programme according to how far they think it has met each of its objectives. These scores are then summed over objectives, using as weights the experts' preferences or priorities to reflect the relative importance of each objective. The resulting composite score can then be compared with cost. With both CEA and MCA, the question remains open as to whether all costs should be taken into account, as in CBA, or only budget costs. In practice, usually only the latter are considered

Pearce (2005) argues strongly for the superiority of a full cost-benefit analysis. His main arguments are that (a) CBA uses citizens' valuations (expressed through their stated or revealed preferences) to value public good outputs, rather than the preferences of experts, and is therefore more 'accountable' than MCA; (b) CBA indicates whether the project provides net benefit to society and should be implemented or not, whereas CEA and MCA can only rank a set of alternatives, on the assumption that *one* of them will be implemented. In fact, none of them might pass the CBA test. Proponents of MCA react to (a) by questioning whether non-experts are sufficiently well informed on agri-environmental issues for their preferences to be as valid as those of experts. As for (b), this argument carries much more weight as regards *ex ante* evaluations, where no policy commitment has been made yet.

This issue might look like a squabble among economists. In fact, it has real operational importance, since these alternatives have different interpretations and answer different questions. Users need to realise that they are not interchangeable options.

---

[20] Cost-benefit studies of publicly funded projects usually also take into account the marginal cost of obtaining tax revenue. Empirical studies show that the cost to society of €1 of tax-payer money is greater than €1.
[21] Rate at which society trades off a benefit occurring in the present against one received in the future. An annual discount rate of 5% means that €100 today is valued as equivalent to €105 in a year's time. The lower the discount rate is, the more willing society is to defer some benefits to the future.

Concluding remarks, future directions

A review of the state of today's evaluation methodology and the way it is applied pinpoints various areas where some new thinking and new procedures could be adopted, and received ideas can be challenged.

Links between ex ante and ex post evaluations

The *ex ante* evaluation of a policy is the moment when, logically, its objectives should be clarified, its intervention logic defined and the questions that are pertinent to its evaluation should be formulated. Once this is done, it is possible, still in the *ex ante* perspective, to identify the data needed both to establish the baseline and to monitor the policy after implementation. Once the policy is adopted, this would serve as the blueprint for data collection procedures.

The *ex ante* evaluation would logically form the starting point for the *ex post* appraisal. And yet, in the more typical case, institutional *ex ante* evaluations are not available to *ex post* evaluators and therefore highly relevant information cannot be exploited. *Ex post* evaluators may have to reinvent a statement of the policy objectives or reconstruct the intervention logic – at least, versions that are precise enough to act as an operational basis for an evaluation – by inductive processes. Key mechanisms and responses that were seen as crucial to the operation of the policy when it was evaluated *ex ante*, may have to be re-identified *ex post*.

The fact that the *ex ante* appraisal was conducted under conditions of greater uncertainty than prevail a few years later, and that the version of the policy finally adopted may not correspond in all details to the one that was formally appraised, does not mean that *ex ante* evaluations are worthless to *ex post* evaluators. If the policy as implemented has been modified from the one first appraised, this could easily be recorded in a codicil to the original *ex ante* evaluation, explaining the rationale for the modification and specifying any changes it implies for data collection needs. This would add value rather than neutralise it.

One-size-fits-all approach?

It has often been observed that the model of the stylized formal evaluation framework as applied in the EU for administrative evaluations is not used on other countries. According to Blandford *et al*. (2010), practitioners' perceptions of this framework range from seeing them as 'box-ticking exercises' to appreciating them as useful attempts at improving the efficiency of policies. Whatever one's view, it has to be acknowledged that the policies to be evaluated are themselves far from being as standardized as the methodology for evaluating them. Might this in some cases be counter-productive? In particular, given the slow emergence of many agri-environmental benefits, it seems inappropriate that *ex post* evaluations after a few years should require long-term impacts to be measured as set out in the common methodological framework.

An interesting – but perhaps outlying – case is that of the Conservation Reserve Programme (CRP) as described by Hellerstein (2005). The main objective(s), rules for eligibility and range of likely impacts have all been evolving over the 25 years of this programme's existence. As Hellerstein puts it, 'as we increase our understanding of how the CRP (and other conservation programmes) affect the bio-physical environment, so will our ability to carefully determine the value of these programmes. Given the geographical breadth of these programmes, and their sometimes subtle impacts, acquiring such information is not inexpensive'.

It is hard to imagine how this policy might be evaluated using the stylised CMEF of the EU. Yet, as documented by Hellerstein, a sustained stream of empirical studies and appraisals, some performed in-house at the ERS and others conducted outside, have built up over time a rich picture of what the CRP has delivered in different sites with respect to different, locally prioritized, conservation objectives. And the evidence continues to accumulate, adding to scientific understanding of the interaction between human decisions and rural environmental outcomes. Having said this, one can also ask whether, given increasing budget stringency and a greater clamour for accountability, a more formalized approach to evaluating at least some aspects of such programmes might also have some merit in the US policy context.

Balance of resources and follow-up ?

As already stated, each type of evaluation approach has its role and purpose. Discussion of the inherent superiority of one or another is quite pointless. However, it is useful for administrations and the research community to ask themselves periodically whether the balance of resources deployed over the three approaches is delivering the maximum benefit to the policy process, to the participants in the policies and to the ultimate stakeholders? Moreover, do the different types of exercise, with their different time-frames, receive their appropriate share of the attention and effective follow-up from policy-makers?

The issue of whether or not policy evaluations are followed up or ignored is hot one. If administrative evaluations are not taken up because their message has become blunted by compliance with a rigid methodology, what can be done about it? If scientific evaluations are not acted upon, is it because they come too late, because the conclusions are expressed in scientific jargon and not well communicated, or for other reasons? If evaluations are not being exploited to upgrade policies and policy-making, then this is an area for improvement where everyone should benefit.

Use of economic information

For Pearce (2005), 'Any rational evaluation of agri-environmental policies, of which AESs are the chosen example here, must compare costs and benefits…' As explained above, there are different measures for trying to do this. But is the choice made always an appropriate one?

The cost-benefit study, which establishes whether a project will yield a net social benefit, can make its greatest effective evaluation contribution *ex ante*, since this is the moment where policy makers

have most degrees of freedom for avoiding policy failures. Nonetheless, when performed *ex post,* CBA can also deliver useful information from which much can be learned.

Short of this, evaluations quite often settle for measuring a project's cost effectiveness. This is designed as a comparative measure, for ranking different policy options. The usefulness of CEA for a *single* policy in isolation, as it is frequently used in practice, can be seriously challenged. Is it helpful to the policy-maker to know that in a project for restoring ancient hedgerows as field boundaries, the cost was X hundred euros per metre of hedgerow? Is the figure 'reasonable' or not? What can we compare it with? If the objective of restoring the hedgerows is to enhance landscape amenity and biodiversity, and preserve cultural heritage, which are valued by society, how can such a ratio help us to evaluate it? Maybe the policy-maker will compare the cost per metre with an independent estimate of the social value of hedgerows per metre. But doing this simply brings cost-benefit analysis in by the back door, and in a most non-rigorous way.

Or it might be argued that, in an *ex post* context, CEA could have a role to play in comparing the performance of the same scheme between regions, or over time. Certainly, this strategy would deploy the CE ratio in a comparative context where it belongs. But what conclusions can we draw from a comparison revealing that cost per metre is higher in region A than in region B? Maybe the benefits deriving from hedgerows are also valued more highly in region A (it may be closer to urban centres, and the countryside may be more extensively used for recreation and tourism).

In conclusion, weighting outcomes according to price (i.e. measuring them in money terms) has the great advantage of allowing us to aggregate them. A second advantage is that the aggregates have a very straightforward and easily understood interpretation as costs and monetized benefits. Still, it does not follow that using economic information in evaluation studies always leads to sensible answers to the questions asked. Economists could be more critical of the way they use these measures. There is a strong temptation to see the economists' main challenge in terms of valuing non-marketed items – and this challenge is, indeed, very real. However, another challenge arises when choosing technique for combining these valuations so as to yield valid conclusions that are useable for evaluating policies.

References

Berriet-Solliec, M., Pierre Labarthe, Catherine Laurent and Jacques Baudry (2011). Empirical validity of the evaluation of public policies: models of evaluation and quality of evidence. Paper prepared for the 122nd EAAE Seminar "Evidence-Based Agricultural and Rural Policy Making: Methodological And Empirical Challenges of Policy Evaluation" Ancona, February 17-18, 2011.

Blandford, D., Boisvert, R. and Hill, B. (2010). Improving the evaluation of Rural development Policy. Guest Editorial, Special Issue: Evaluating Rural Development Policy, *EuroChoices* 9: 4-9.

Boatman, N., Conyers, S., Parry, H., Pietravalle, S., & Ramwell, C. (2007) *Estimating impacts of ELS on key biodiversity indicators and diffuse pollution of surface waters by nutrients.* Final Report to Defra, project MA01041.

Boatman, N., Ramwell, C., Parry, H., Jones, N., Bishop, J., Gaskell, P. Short, C., Mills, J. and Dwyer, J. (2008). *A review of environmental benefits supplied by agri-environment schemes*. FST20/79/041. Report to the Land Use Policy Group, UK.

Boatman, N., Willis, K., Garrod, G., and Powe, N. (2010). *Estimating the Wildlife and Landscape Benefits of Environmental Stewardship*. Final Report. Report for the Food and Environment Research Agency, Defra, UK.

Cooper, T., Hart, K. and Baldock, D. (2009). The Provision of Public Goods Through Agriculture in the European Union. Report prepared for DG agriculture and Rural Development. IEEP, London.

European Commission (2006). *Handbook on Common Monitoring and Evaluation Framework. Guidance document and Annexes*. Directorate General for Agriculture and Rural Development. Available at http://ec.europa.eu/agriculture/rurdev/eval/index_en.htm

Falconer, K. and Saunders, C. (1999). Transaction costs for SSSIs and policy design. *Land Use Policy* 19: 157-166.

Hanley, N., R.E. Wright and V. Adamowicz (1998). Using choice experiments to value the environment. *Environmental and Resource Economics* 11: 413-428.

Hellerstein, D. (2005). Conservation policy and agriculture in the US: valuing the impacts of the conservation reserve program. Chapter 13 in: OECD, *Evaluating Agri-Environmental Policies. Design, Practice and Results*. Paris: OECD, 231-257.

Hodge, Ian (2008). To what extent are environmental externalities a joint product of agriculture? Overview and policy implications, in OECD, *Multifunctionality in Agriculture: Evaluating the Degree of Jointness, Policy Implications*. Paris: OECD, 85-118.

IEEP (2006). An Evaluation of the Less Favoured Area Measure in the 25 Member States of the European Union. Report prepared by the Institute for European Environmental Policy for DG Agriculture November 2006.

Kleijn, D. & Sutherland, W.J. (2003). How effective are European agri-environment schemes in conserving and promoting biodiversity? *Journal of Applied Ecology* 40: 947-969.

Kleijn, D. (2006). Guidelines for the evaluation of AE schemes. Wageningen University. Available at: //www.ncp.wur.nl/NR/rdonlyres/EA88E46A-1B26-4A04-AF51-4FBB5AC50619/52783/ Guidelines_evaluation_ae_schemes2.pdf

Komen, M.H.C. and J.H.M. Peerlings (1998). Restricting intensive livestock production: Economic effects of mineral policy in The Netherlands. *European Review of Agricultural Economics* 25:110-128.

Loomis, J., P. Kent, L. Strange, K. Fausch, and A. Covich (2000). Measuring the total economic value of restoring ecosystem services in an impaired river basin: results from a contingent valuation survey. *Ecological Economics* 33: 103-117.

McVittie, A., Moran, D. and Thomson, S. (2009). A review of literature on the value of public goods from agriculture and the production impacts of the single farm payment scheme. Report prepared for the Scottish Government's Rural & Environment Research & Analysis Directorate (RERAD/004/09), Land Economy and Environment Research Group, Scottish Agricultural College.

Mettepenningen, E., Verspecht, A. and Van Huylenbroeck, G. (2009). Measuring private transaction costs of European agri-environmental schemes. *Journal of Environmental Planning and Management* 52: 649-667.

Nilsson, F.O.L. (2007). Transaction costs of agri-environmental policy measures: the case of Swedish permanent grasslands. Swedish University of agricultural Sciences. Working Paper Series 2007:6. Sweden: SLU.

OECD (2005). *Evaluating Agri-Environmental Policies. Design, Practice and Results*. Paris: OECD.

OECD, 2006. Policy related transaction costs and policy choice: main report. AGR/CA/APM(2003)15/FINAL. Paris:OECD.

OECD (2010). *Guidelines for Cost-effective Agri-environmental Policy Measures*. Paris: OECD.

Oñate, J.J., Andersen, E., Peco, B. and Primdahl, J. (2000). Agri-environmental schemes and the European agricultural landscapes: the role of indicators as valuing tools for evaluation. *Landscape Ecology* 15: 271–280.

Oréade-Brèche (2005). Evaluation des mesures agro-environnementales : Rapport final. Report for the European Commission.

Pearce, D. (2005). what constitutes a good agri-environmental policy evaluation? Chapter 3 in OECD, *Evaluating Agri-Environmental Policies. Design, Practice and Results*. Paris: OECD, 71-93.

Peerlings, J.H.M. and Nico Polman (2008). Agri-environmental contracting of Dutch dairy farms: the role of manure policies and the occurrence of lock-in. *European Review of Agricultural Economics* 35: 167-191.

Pruckner, G. (1995). Agricultural landscape cultivation in Austria: An application of the CVM. *European Review of Agricultural Economics* 22: 173-190.

Scarpa, R., D. Campbell and W.G. Hutchinson (2007). Benefit Estimates for Landscape Improvements: Sequential Bayesian Design and Respondents' Rationality in a Choice Experiment. *Land Economics* 83: 617-634.

Selman, M., Greenhalgh, S., Branosky, E., Jones, C. and Guiling, J. (2009). *Water Quality Trading Programs: An International Overview*. WRI Issue Brief #1. Washington DC: World Resources Institute.