

Review of Defra's Research Programme on Environmental Stewardship

Co-ordinating reviewer's report and comments

Prof S. J. Ormerod

Cardiff School of Biosciences, Cardiff University, PO Box 915, Cardiff CF10 3AX
Ormerod@cardiff.ac.uk

Summary

Defra's 'Environmental Stewardship' (ES) research programme provides key evidence required to support Environmental Stewardship in England. The programme was last reviewed in 2004, but further assessment is now timely in view of changes in project management, altered strategic priorities and evolving Defra policy linked to recent agenda-setting documents.

For this assessment of the programme's science quality, 26 general ecological reviewers and six statisticians appraised a large sample of 39 projects of total value £9.6 million on criteria reflecting reporting, methods, data, analysis, conclusions and overall quality.

Among 97 reviews, 87% categorised projects as either 'sound and robust' or 'sound' science, the top two categories available. This result is near-identical to the previous assessment in 2004, when 91% of reviews were in these two grades.

Recommendations are made to address minor issues of statistical analysis, scientific design and reporting; to improve the small number of weaker projects; and to address other issues raised by reviewers with respect to the organisation of on-line reports.

Overall, however, it is concluded that Defra are not only receiving overwhelmingly effective science in this research programme, but also that high quality has been maintained through recent changes in project management.

Context and overview

Defra's 'Environmental Stewardship' (ES) research programme is intended to provide key evidence required to support Environmental Stewardship in England. This, in turn, is a policy response designed to offset the adverse effects of long-term agricultural intensification on semi-natural habitats, farmland biological diversity, landscape quality and historic landscape features. The drivers of adverse change are multiple, but include increased agricultural specialisation, losses of spring-sown crops and weedy stubbles, increased use of pesticides and fertilizers, more efficient crop harvesting and storage, intensification of grassland management, loss of hedges, margins and other habitats, and drainage of wet grassland.

Launched by Defra in 2005, ES encompasses several voluntary agri-environment schemes in which participants are paid at rates based on income forgone to manage land according to prescriptions thought likely to conserve the characteristics features and wildlife of the English countryside. The four key elements are i) Entry Level Stewardship (ELS), open to all farmers and land managers who

want to deliver a basic level of environmental management; ii) Organic Level Stewardship (OELS), open to all farmers registered as organic but not receiving aid under any other organic aid scheme; iii) Uplands ELS (UELS), open to upland farmers with land in Severely Disadvantaged Areas and iv) Higher Level Stewardship (HLS), open to farmers who want to deliver higher levels of environmental management. The Rural Development Programme for England (RDPE) provided £2.9bn for ES over the period 2007-13, or about £400m p.a. Of this, the budget for the ES research programme is around £2m per annum, or 0.5% of the total expenditure.

It is important that the quality of the research underpinning ES prescriptions is reviewed periodically to ensure that the scientific evidence provided is of the highest possible quality. As Defra's 'Farmland Conservation and Biodiversity' programme, the science quality of this research area was last reviewed in 2004 (Ormerod 2005). This is distinct from a Review of Progress of the whole of ES in 2008, carried out largely by the Central Science Laboratory (now the Food and Environment Research Agency), to provide assurance in relation to delivery, explore ways of securing better value for money and take account of new policy priorities, notably climate change.

Three key developments now make a further review of science quality of the ES research programme timely.

1. Research management in ES has been modified since the last review. Although the budget for the work remains in Defra, Natural England (NE) is now a joint manager with the intention that commissioned R&D improves ES delivery on the ground. NE staff act as Managers or Nominated Officers (equivalent to Project Officers) for most units of assessment (e.g. for BD12, BD13, BD14 etc. see Table 1). Additionally, the R & D programme, along with a separate ES Monitoring and Evaluation Programme, are overseen by an ES Evidence Management Group that is co-chaired between Defra and NE. This group provides strategic direction for the programme and recommends R&D priorities to the Defra/NE ES Liaison Group.
2. The strategic priorities of the ES research programme, as articulated by the ES Evidence Plan, are evolving. The current plan (from 2009) identifies four key evidence gaps in ES that are respectively i) the effectiveness of current interventions; ii) the need to inform EU and national agri-environment policy, for example with respect to current and potential contributions to ecosystem services and climate-change mitigation; iii) the need to report on EU and national targets or objectives as outcome indicators for agri-environment schemes, for example whether SSSI condition is responding to ES; and iv) research to improve scheme design and delivery, for example, to estimate the scale of intervention needed to reverse the decline in key widespread farmland specialist organisms, particularly birds, and to agree surrogate indicators to measure the contribution of ES contribution to safeguarding farmland bird populations.
3. The context in which research priorities are set is also evolving. For example, the 2011 iteration of the ES Evidence Plan reflects the publication during last year of several agenda-setting documents including Making Space for Nature ("Lawton review"), the Natural Environment White Paper, the England Biodiversity Strategy and the National Ecosystem Assessment. The context is also affected by the current financial environment as reflected by the announcement of the Government Spending review in 2010. Looking forward, reform of the Commons Agricultural Policy in 2013 will create yet further challenge – likely to be of major significance to the management of biodiversity and ecosystem services on English farmland, and hence the need for underpinning research.

All of the above issues require support from a research programme in which the science is not only robust, well-designed and well-executed, but also relevant and able to deliver at the scale at which ES must operate.

This review was commissioned with the overall aim of evaluating the science involved through a co-ordinated process in which peer reviewers assessed final reports and provided quantitative and qualitative feedback. The questions asked of reviewers meant that the focus was on appraising quality rather than scope, dissemination or translation into practice. While Defra provided funding, secretarial services and support for the review, the review process was entirely independent, for example in the selection of projects, choice of reviewers and in the final reporting. Contractors were shown the report and offered a chance to comment on the contents. They were also offered a chance to provide comments on individual reviews to which reviewers were asked respond – though no material change in project grades resulted.

The sample of projects

A sample of 39 projects formed the review and was chosen to reflect:

1. *All active Units of Assessment that had completed projects (see Table 1)*
2. *A range of project sizes*

Projects had a median size £169k closely similar to that in the entire programme of £149k. Individual projects ranged in size from £5k-1165k, with 38% < 100k and 77% < 300k. Again, this range was near identical to the overall project list suggesting that the sample was representative. The total value of projects reviewed was c £9,612,000, and hence a large proportion of the total spend from 2005-2010.

3. *A range of the institutions and types of organisations delivering Defra research.*

These included projects led by ADAS (4), BBSRC (1), colleges and universities (5), the BTO (7), Butterfly Conservation (1), CABI (1), CEH (6), CSL or the Food and Environment Research Agency (2), Forest Research (1), IGER (5), RSPB (3), Oxford Archaeology (1), Pond Conservation (1) and Consultants (1).

4. *All project types and approaches in the research programme (e.g. experiments, reviews, modelling) roughly in proportion to their occurrence.*

There was some slight (but not significant) bias toward more recent projects to reflect currency and topicality as well as the logistics of contacting contractors who were still operating at the time of review.

The review process and assessment methods

A total of 6 specialist statisticians and 26 reviewers with more general expertise in biodiversity, agriculture or archaeology was selected using indicators of esteem, relevant research and scientific experience indicated, for example, from the ISI Web of Knowledge. Most were from the UK, although a small number were experts active in similar areas elsewhere in Europe. Reviewers were selected also for their independence and to avoid involvement with ongoing or recent Defra-funded research anywhere in the ES programme. Most were asked to review several projects in their field, and for each the reviewers were asked to appraise the work reported on 13 questions about:

Table 1. *The units of assessment included in the current review, and the number of sample projects reviewed in each.*

*Unit of assessment	Scope	N projects reviewed
BD 12	Uplands	3
BD 13	Wetlands	6
BD 14	Grasslands	12
BD 15	Lowland Heath	1
BD 16	Arable Farmland	9
BD 17	Landscape, History and Amenity	1
BD 20	General Conservation	1
BD 21	Hedgerows	1
BD 23	Horizon-scanning and Cross-cutting Issues	3
BD 50	Horizon-scanning and Cross-cutting Issues	1
BD 53	Landscape, Historic & Amenity	1
Total projects		39

*Note: No projects from BD18 (Development of Incentive Schemes), BD19 (Countryside Survey 2000), BD22 (Resource protection), BD51 (Management and Restoration of Semi-Natural Habitats) or BD52 (Improving the Wildlife Value of Farmland) were included either because none were commissioned, all were completed before 2004 or because they had not reported by the cut-off date for this evaluation of March 31 2010.

1. *The reporting and methods*

- whether all the objectives were met
- whether the hypotheses were flawed
- whether there were any weaknesses in the study design
- whether any assumptions were sound and clear
- whether the use of any animals was justified.

2. *The data and analysis*

- whether uncertainties and statistical measures of dispersion were recognisable
- whether statistical analysis was sound and appropriate
- whether figures and tables were adequate and used to draw appropriate inference

3. *The conclusions*

- whether alternative explanations for the results had been evaluated
- whether all relevant subject matter and evidence was included
- whether judgement was clearly separated from evidence in supporting conclusions
- whether conclusions and recommendations followed logically from the data and evidence

4. *The overall grade for the project*

- ranging from 'sound and robust' to 'evidence of 'poor quality'.

Reviewer responses to each question were ranked into three or four grades from 1 (= the most desirable) to 4 (= the least desirable). Both the grading process and criteria were near-identical to that used in the last review of the research programme in 2004 so that the results are directly comparable. Some analyses were carried out on mean grades calculated across all criteria for each project.

In practice, the question about the use of animals was of limited relevance because very few animals were handled during the projects and none required licensing by the Home Office. In one instance where small mammals might have been captured, the investigators found a novel means of collecting data using other methods that circumvented this need (Pocock & Jennings 2006)

The review outcome

i) Overall quality

The 39 projects generated 75 and 27 reviews, respectively, from the general reviewers and specialist statisticians. Of 97 reviews providing an overall grade, 32% scored projects as 'sound and robust', while a further 55% provided 'sound science'. Thus, 87% of reviews scored as being in the top two categories (Table 1). Moreover, among the 13% of reviews that placed projects in the lower two categories, there was only project where two reviewers agreed that overall quality was sub-standard, and none where all three reviewers agreed. This is in marked contrast to 17 projects where both general reviewers agreed on grades that were either 'sound' or 'sound and robust'.

There was no relationship between project cost and overall project grades showing that quality is being maintained across the entire project range from small to large size.

ii) Individual quality criteria

On individual criteria, 75-100% of responses from both types of reviewer placed projects in the top or second-top category on all except one case (Table 2). This exception was where around 40% of specialist statistical reviews indicated that measures of dispersion and uncertainty could have been clearer in those projects where there had been detailed statistical analysis. Aspects of study design and clarity of assumptions otherwise received the lowest grades from both statistical and general reviewers.

iii) Consistency among reviewers

Where two reviewers examined projects, non-specialist reviewers were, in general, in broad rather than perfect agreement about the quality of projects ($r = 0.46$, $P < 0.005$, $n = 36$; based on mean grade for all criteria). This was less true for statistical reviewers who tended to look mostly at project statistical approaches so that their overall grades departed slightly from the non-statisticians ($r = 0.34$, $P = 0.08$, $n = 29$). With respect to the mechanics of the review, therefore, the statisticians added a distinct component, while the use of two non-specialists was useful in broadening views on each project. There was no systematic variation among reviewers in the grades proposed per project; in other words none was systematically harder or more lenient than others, showing that consistent standards were applied in the review process.

iv) Consistency and variation among organisations providing research

When classified as 'University', 'Private', 'NGO' or 'Government Institute', there were no significant variations in overall mean review grade among sectors from general reviewers ($F_{3,35} = 1.04$, N.S.).

For specialist statisticians, variations approached formal significance ($F_{3,23} = 2.74$, $P = 0.066$) due to marginally improved performance by the NGO sector in general.

Where multiple projects (> 1) had been carried out by the same contractors, there was no systematic variation in mean grades among providers ($F_{6,38} = 0.75$ NS) suggesting that strengths and weaknesses arose on a project basis rather than an institutional basis – for example where project design, data collection and scale had been particularly challenging. For statistical reviewers, there was some minor variation among contractors due largely to particular strength in one organisation rather than systematic failings in others ($F_{5,21} = 3.66$, $P < 0.05$)

Areas for improvement

A review in which 87% of reports graded projects as “sound” or “robust” leaves only modest scope for improvement. This overall outcome is particularly positive given that research in this programme must address the challenges of research at whole ecosystem scales, often over prolonged timescales, and with complex processes and confounding effects that present major difficulties of experimental design, replication and manipulation. These are exactly the difficulties that this programme overcomes to deliver policy-relevant science, often in an exemplary fashion and with considerable methodological innovation.

There is, nevertheless, an opportunity to consider i) how to address minor shortcomings and ii) how weaker projects could be improved.

Improving minor shortcomings

On the basis of reviewer grading, the three minor shortcomings most in need of attention were:

i) Measures of dispersion and uncertainty

Statisticians scored performance on this criterion less positively than general reviewers (58 v 78% of reviews in top two grades) indicating that this is a mostly a statistical issue. As a matter of practice, investigators on projects involving data analysis need to involve competent ecological statisticians at all stages from the design stages onwards. Report authors need to take full responsibility for ensuring that error, uncertainty and statistical power (where appropriate) are anticipated during design, incorporated explicitly into analysis, depicted clearly in data displays, and recognised in interpretation.

ii) Aspects of study design

In contrast to issues of dispersion and uncertainty, problems with study design were noted more by general reviewers (22% of reports) than statisticians (17% of reports). However, only for 5 of the 39 projects did two or more reviewers express design concerns.

Shortcomings were related invariably to inadequate replication or statistical power, insufficiently large experimental plots and risks of edge effects. These are among some of the most widespread challenges in ecological experimentation, but hint at the need in this programme to treat statistical power as a serious issue in experiments and also to reduce experimental complexity by reducing numbers of treatments in order to increase numbers of independent replicates per treatment. Trading-off realism and complexity in experiments against ideal design will require Investigators and programme managers to ensure an appropriate balance between pragmatism and statistical robustness.

One other failing noted in some reports was the simpler problem of poor description of experimental design or site selection. Basic good practice should ensure that reports receive comment from independent readers prior to completion in order to minimise such problems.

iii) Making assumptions clear

Issues about the clarity of assumptions were raised in roughly a quarter of all reviewers' reports, and at the same rate from both general reviewers and statisticians. As with design concerns, however, two or more reviewers agreed there was a problem in only a small number of projects (4).

Issues about assumptions ranged from the generic (e.g. assumptions made that were unsupported by evidence; correlations assumed to reflect cause and effect; assumptions about the representativeness and appropriateness of data for statistical procedures) to the specific (e.g. the use of untested models). In some cases, investigators had assumed that experimental protocols mimicked real-world agricultural practice and then extrapolated perhaps inappropriately.

These minor concerns over assumptions arise at all stages from project design and inception through to write-up. Solutions might include i) fuller independent scrutiny and consultation from project inception onwards, for example with programme managers and ii) requesting that assumptions are made explicit in reports. However, some of these issues are simple errors of basic good science practice, for example with respect to analysis and reporting.

Improving weaker projects

Only five projects (12%) had mean grades across all assessment criteria that were > 2.5 as scored by general reviewers, indicating that standards were consistently outside the top two categories. In these cases, performance was always poor on several measures simultaneously, usually some combination of unclear objectives, poor hypotheses, weak design, unclear or unjustified assumptions, unsound analyses, tendency to rely on judgement rather than evidence, poor recognition of dispersion and weak communication of alternative explanations for the patterns observed. Such multiple sub-standard scores suggest generic issues at the project level. Solutions are likely to be in improved management within projects along with closer scrutiny from project officers at Defra and Natural England.

Just one project had an average score of > 2.5 according to statistical reviewers, implying general satisfaction across projects despite the problem noted above with respect to dispersion and uncertainty.

Conclusions and recommendations

It is important to recall that this review has addressed aspects of science quality supporting Environmental Stewardship rather than the implementation and uptake of research into the programme itself, or of the scope of research undertaken relative to the changing context in which Stewardship is funded. In addition, the review has operated mostly on a project-by-project basis rather than as an assessment of the aggregate gain and added value, for example, of accumulated long-term data.

On this basis of quality, these data overall indicate that Defra are receiving overwhelmingly sound science in support of Environmental Stewardship. For comparison, a similar exercise in 2004 examined 19 projects that generated 47 reviews from an identical constituency of reviewers of which 24% scored projects as 'sound and robust', while a further 68% scored them as 'sound science'. In other words, 92% of projects were in the top two categories by comparison with 87% here, a difference that is not formally significant. In addition to receiving good quality science from their investment in this programme, therefore, Defra are also receiving evidence of a standard that is consistent through time. Two corollaries are that i) science quality has been maintained through the changes in project management outlined in the introduction and ii) science quality is in a position to continue to deliver at a high standard going forward as the programme's context evolves.

Recommendations about this programme are similar to those made in 2004.

1. To address statistical and design issues:

- Researchers on quantitative projects should continue to involve specialist data analysts and those skilled in experimental design at all project stages to ensure that design is optimum and analysis of high quality.
- Investigators should give more emphasis to anticipating, addressing and representing dispersion, uncertainty and statistical power.
- Investigators and project managers should consider how best to trade off experimental complexity against power by reducing numbers of treatments in order to increase numbers of independent replicates per treatment.
- There should be on-going dialogue and consultation from project inception onwards between investigators and programme managers on all aspects of experimental design.
- Investigators and project managers should consult with agronomists and practitioners, where appropriate, to ensure that any manipulations represent real practice faithfully.

2. To address issues in reporting:

- Investigators should make their assumptions explicit in reporting their projects, for example in a clear table
- As a basic element of good practice, investigators should continue to seek critical, independent review of their work and reports either from colleagues or, ideally, through international peer-review and publication in high quality journals
- Defra should continue to make clear from the outset the specific standards required for reporting.

3. To improve weaker projects:

- The small number of weaker projects in the programme appeared to require improved management, and the institutions involved should be informed. Clear plans for effective project management should continue to form a key part of the selection of contractors by Defra.

4. Other issues raised by reviewers

- Some reviewers recorded some frustration about the format of project reports held by Defra – which were often disaggregated across several different electronic files.

Report organisation and accessibility could be improved for all readers and users of the important and hard-won information they contain.

All these issues are relatively minor and, as noted throughout, for the most part represent the maintenance of good science practice.

References

Ormerod, S. J. (2005) Review of Defra's Research Programme on Farmland Conservation and Biodiversity: Co-ordinating reviewer's comments.
<http://archive.defra.gov.uk/evidence/science/how/documents/programmes/Review-FarmlandConservation.pdf>

Pocock, M. J. O. and Jennings, N. (2006), Use of hair tubes to survey for shrews: new methods for identification and quantification of abundance. *Mammal Review*, 36: 299–308. doi: 10.1111/j.1365-2907.2006.00092.x

Table 2. Grades given by general and specialist statistical reviewers to each of 39 reports assessed during the review on each criterion. Shaded numbers indicate modal response, while the overall percentage of reviews in the top two grades is indicated in the right-hand column.

Criteria	N	Not scored	1 (Best)	2	3	4 (Worst)	% in top two grades
<i>Scores from general reviewers</i>							
Objectives satisfied?	75		34	29	10	2	84
Hypotheses flawless?	53	21	25	19	7	3	81
Design robust?	67	8	28	24	8	7	78
Assumptions clear?	72	3	23	31	16	2	75
Dispersion and uncertainty clear?	65	10	38	13	10	4	78
Statistical analysis sound?	59	16	36	15	3	5	86
Figures and tables appropriate?	73	2	44	26	3		96
Alternative explanations considered?	70	5	23	33	11	3	80
Evidence used appropriately?	72	3	36	28	6	2	89
Judgment v evidence distinguished?	74	1	48	19	7		91
Data and logic clear?	74	1	31	37	4	2	92
Overall rating	75		27	39	6	3	88
<i>Scores from statistical reviewers</i>							
Objectives satisfied?	23	4	12	8	2	1	87
Hypotheses flawless?	9	18	4	4	1		89
Design robust?	24	3	6	14	3	1	83
Assumptions clear?	19	8	3	11	5		74
Dispersion and uncertainty clear?	27		6	10	11		59
Statistical analysis sound?	27		10	12	4	1	81
Figures and tables appropriate?	27		14	11	2		93
Alternative explanations considered?	15	12	6	6	3		80
Evidence used appropriately?	14	13	8	4	2		86
Judgment v evidence distinguished?	16	11	9	7			100
Data and logic clear?	20	7	8	10	2		90
Overall rating	22	5	4	14	4		82