Environ. Biosafety Res. 8 (2009) 65–78
© ISBR, EDP Sciences, 2009
DOI: 10.1051/ebr/2009009

Available online at:
www.ebr-journal.org

Commentary

Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms

Joe N. Perry[1]*, Cajo J.F. ter Braak[2], Philip M. Dixon[3], Jian J. Duan[4], Rosie S. Hails[5], Alexandra Huesken[6], Marc Lavielle[7], Michelle Marvier[8], Michele Scardi[9], Kerstin Schmidt[10], Bela Tothmeresz[11], Frank Schaarschmidt[12] and Hilko van der Voet[13]

[1] Oaklands Barn, Lug's Lane, Broome, Norfolk NR35 2HT, UK
[2] Biometris, Plant Research International, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, The Netherlands
[3] Department of Statistics, 120 Snedecor Hall, Ames IA, 50011-1210, USA
[4] USDA ARS Beneficial Insects Introduction Laboratory, Newark, Delaware 19713, USA
[5] Centre for Ecology and Hydrology, Mansfield Rd, Oxford OX1 3SR, UK
[6] Julius Kuehn Institute, Federal Research Centre for Cultivated Plants (JKI), Institute for Biosafety of Genetically Modified Plants, Messeweg 11/12, 38104 Braunschweig, Germany
[7] INRIA Saclay, Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France
[8] Dept. of Biology & Environmental Studies Institute, Santa Clara University, Santa Clara, CA 95053, USA
[9] Department of Biology, "Tor Vergata" University, Via della Ricerca Scientifica, 00133 Rome, Italy
[10] BioOK GmbH, Schnickmannstrasse 4, 18055 Rostock, Germany
[11] Department of Ecology, University of Debrecen; Debrecen, P.O. Box 71, 4010, Hungary
[12] Leibniz Universität Hannover, Fakultät Naturwissenschaften, Institut für Biostatistik, Herrenhaeuser Str. 2, 30419 Hannover, Germany
[13] Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, The Netherlands

Previous European guidance for environmental risk assessment of genetically modified plants emphasized the concepts of statistical power but provided no explicit requirements for the provision of statistical power analyses. Similarly, whilst the need for good experimental designs was stressed, no minimum guidelines were set for replication or sample sizes. Furthermore, although substantial equivalence was stressed as central to risk assessment, no means of quantification of this concept was given. This paper suggests several ways in which existing guidance might be revised to address these problems. One approach explored is the 'bioequivalence' test, which has the advantage that the error of most concern to the consumer may be set relatively easily. Also, since the burden of proof is placed on the experimenter, the test promotes high-quality, well-replicated experiments with sufficient statistical power. Other recommendations cover the specification of effect sizes, the choice of appropriate comparators, the use of positive controls, meta-analyses, multivariate analysis and diversity indices. Specific guidance is suggested for experimental designs of field trials and their statistical analyses. A checklist for experimental design is proposed to accompany all environmental risk assessments.

Keywords: environmental risk assessment / statistical analysis / experimental design / equivalence test / genetically modified plant / statistical power

INTRODUCTION

The need for review

Traditionally, the major statistical test done for comparative assessment within risk assessments for GM plants is a 'difference test' between a GM plant and an appropriate (usually near-isogenic) comparator. If the GM plant is shown to be different from the comparator this is a 'proof of difference'. Such a difference may constitute a hazard (potential risk), which is then subject to further safety evaluation. For this reason the difference test is sometimes referred to as a 'proof of hazard'. From the statistical aspect, there are two major reasons why the current European Food Safety Authority (EFSA) guidance (EFSA, 2006) may require review.

* Corresponding author: joe.perry@bbsrc.ac.uk

First, whilst the guidance rightly emphasized the concepts of statistical power detailed below, it provided no explicit requirements for the provision of statistical power analyses. Whilst the need for good experimental designs was correctly stressed, no minimum guidelines were set for replication or sample sizes. A more proscriptive approach may have avoided problems of poorly replicated and ill-thought experimental designs or the use of experimental data from experiments designed initially for different purposes (EFSA, 2009c).

Second, the guidance rightly emphasizes the concepts of familiarity (a history of safe use) and substantial equivalence (existing food/feed organisms with a history of safe use acting as comparators), but provides no means of quantification of these concepts or explicit means for them to enter into the recommended statistical analyses. An opportunity to formalize these concepts is available through the use of the so-called 'bioequivalence' approach, first used in pharmaceutical studies (Schuirmann, 1987; Tempelman, 2004).

For the consumer, the error of most concern in a difference test is of falsely inferring that no hazard exists where there may be one. Because the traditional statistical null hypothesis employed is one of equality, this error is relatively difficult to estimate accurately and/or set to a desired magnitude. This had led to obvious problems with credibility and transparency. This disadvantage is overcome by the equivalence test, sometimes referred to as a 'proof of safety', since here the null hypothesis is one of inequality, and the error of most concern to the consumer may be set relatively easily. The advantage of equivalence testing is therefore that the onus is placed back on to those who wish to demonstrate the safety of GMOs to do high-quality, well-replicated experiments with sufficient statistical power. Such tests have a place in both environmental risk assessment, the subject of this paper, and in food-feed risk assessment. However, they were introduced first in updated guidance for the latter (EFSA, 2009a); the next section explains their role in more detail.

Equivalence testing in food-feed risk assessment

For food-feed risk assessment, any differences found are placed into biological context, by defining equivalence as the absence of differences other than those expected naturally through variation between crop varieties. To perform an equivalence test, several quantities are required, for each endpoint tested. All statistical tests require a test statistic, a value calculated from data, and used to decide whether to reject the null hypothesis. In the case of food-feed risk assessment the test statistic is the difference between the mean of the GM and the mean of several commercial varieties with a history of safe use; the mean value of the comparator does not enter the calculations in any form. Furthermore, the null hypothesis for the equivalence test requires the specification of an upper and a lower so-called equivalence limit. In the updated Guidance (EFSA, 2009a) these equivalence limits are set based upon estimates of the natural variation between the commercial varieties from the same compositional field trials used to derive the difference tests between the GM and its comparator. It is essential that the commercial varieties are integrated into those field trials as fully randomized and replicated treatments, because the data on commercial varieties must be obtained in identical conditions to that of the GM and its comparator. This eliminates uncontrollable confounding effects and conforms to the need for randomization as a fundamental principle of good experimental design.

The analysis proceeds through the simultaneous application of both tests, of difference and of equivalence; the whole process is illustrated in Figure 1A. It is recommended to consider a logarithmic transformation because experience suggests it is appropriate, not only to stabilize variance, but also for the more important reason that, for many endpoints, the treatment effects combine multiplicatively rather than additively. The particular approach taken is one termed 'average equivalence' in the drug testing literature (Wellek, 2002). In this, tests are based on endpoint values averaged across sites and seasons. However, because of concerns that negative effects may only show at single locations, allowance is also made for the need to assess differences in treatment effects between sites, the so-called 'site × treatment' interaction.

Further technical details with worked examples are given in a report of the EFSA Statistics Working Group (EFSA, 2009b).

Statistical power and equivalence tests

Of crucial importance in any risk assessment framework is the probability that a given test will detect effects of a defined magnitude, known as the statistical power of the test. For transparency and public confidence it is important in GMO risk assessment that the consumer risk be both well defined and low.

Equivalence testing contrasts with much other biological experimentation: in the former the risk assessor tests a null hypothesis of inequality between the GMO and its control, which must be actively disproved if the experimenter is to conclude that the GMO is equivalent to the comparators concerned. By contrast, in the statistical test and in most biological experiments the null hypothesis is one of equality (no difference).

In any difference test of a null hypothesis there are two possible types of errors, which are

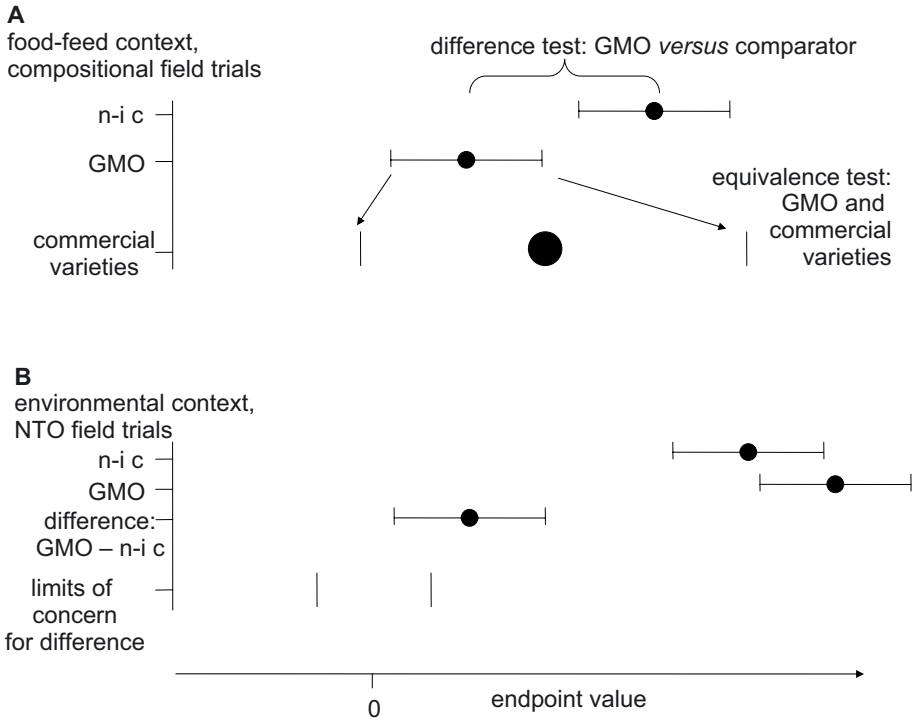

Figure 1. Examples showing difference between testing procedures for: A, compositional field trials for food-feed risk assessment, and B, NTO field trials for environmental risk assessment. **(A)** For food-feed risk assessment, two separate tests are performed: a difference test between the means (shown as filled circles with appropriate confidence intervals shown as bars) of the GM and the near-isogenic comparator (n-i c); and an equivalence test between the mean GM and the mean of the commercial varieties with a history of safe use. Limits for the equivalence test are derived from the natural variation between the commercial varieties as estimated from the trials, and shown as vertical lines. To reject the null hypothesis of no difference for the difference test, a confidence interval for the difference between the means of GM and n-i c must not include zero. To reject the null hypothesis of non-equivalence for the equivalence test, the confidence limit for the GM must lie entirely within the equivalence limits (as in the example shown). **(B)** For environmental risk assessment, two separate tests may again be performed, usually after a logarithmic transformation to achieve a multiplicative scale. Here, example mean values are shown for the GM and the n-i c, as well as, on the same axis, the difference between these quantities. To reject the null hypothesis of no difference for the difference test, the confidence interval for the difference between the means of GM and n-i c must not include the value zero shown on the endpoint value axis (as is the case in the example shown). Limits of concern for the difference between GM and n-i c are set prior to the field trials and shown as vertical lines. To reject the null hypothesis of non-equivalence for the equivalence test, the confidence limit for the difference between GM and n-i c must lie entirely within these limits of concern (as is not the case in the example shown). Similar considerations apply to laboratory experiments as to field trials.

mutually exclusive. A so-called 'type I' error occurs if the null hypothesis is erroneously rejected when it is actually true. A 'type II' error occurs when the null hypothesis is not rejected even though it is actually untrue. It is relatively simple for scientists to set the type I error rate for an experiment, but it is much more difficult to estimate the type II error rate accurately, and technically impossible to set it *a priori* to a desired value.

Traditionally, in many experimental disciplines the type I error rate, α, is set to $\alpha = 0.05$. The so-called '5% level' is conventionally considered as acceptable for safety tests (Marvier, 2002). However, in risk assessment, it is the type II error of the difference test that is the most serious and relevant one (Hill and Sendashonga, 2003). Poorly designed experiments, or those with inadequate replication, even though using a 5% type I error rate, may have such large type II error that they lack the ability to discriminate in a difference test between the GMO and its comparator. The mathematical complement of type II error is termed 'statistical power', which is defined as unity minus the type II error. Recall from above that for equivalence tests the error of most concern to the consumer is the type I error, which may be set at 5% relatively easily. Also, the framing of the null-hypothesis as one of non-equivalence, rather than equivalence, is fully in line with a precautionary approach, in which

the experiment must have sufficient statistical power to reject the null hypothesis in favor of the alternative, which is the only way a conclusion of no safety concern may be drawn. It is for this reason that equivalence testing was said (see above) to place the onus on to those who wish to demonstrate the safety of GMOs to do high-quality, well-replicated experiments, which should achieve the required degree of statistical power for both the equivalence and difference tests. For some information concerning the power of equivalence tests see Niazi (2007), Yata (2008) and http://www.tau.ac.il/cc/pages/docs/sas8/analyst/chap12/sect4.htm.

Contrasts between statistical approaches for environmental risk assessment and food-feed risk assessment

In order to apply the concept of equivalence testing to environmental risk assessment (ERA) it may help to identify some major statistical issues concerning the requirements for ERA and for food-feed risk assessment. First, the particular GM trait concerned is often designed specifically to have effects that necessarily impinge on ecosystems but would not be intended to affect food-feed parameters. For example, GM herbicide-tolerance and GM insect-resistance systems, by intention, cause some, albeit small, environmental impact, but are not intended to affect, say, allergenicity. Therefore, the concept of a history of safe use from food safety relates less easily to ERA, in which environmental harm is measured. Here, it is more fruitful to base arguments on the likely effect of a GMO, and then to contextualize whether that effect is sufficient to cause significant environmental harm. To retain the undoubted benefits of the equivalence approach, outlined above, the test must therefore be adapted. Second, for ERA, it makes little practical sense for the equivalence limits to be based on the natural variation of extraneous varieties. Instead, it is more appropriate to define them directly as 'limits of concern', by which is meant the minimum relevant ecological effect that is deemed biologically significant, and is deemed of sufficient magnitude to cause harm. Experience suggests that the setting of limits of concern is more feasible than in food-feed risk assessment. Third, if commercial varieties do not form an essential part of the process, it is natural to use the same test statistic for the equivalence test as is used for the difference test, *i.e.* the difference between the GM and its (usually near-isogenic) comparator (Fig. 1B, and see section on comparators below and EFSA, 2009b). Fourth, there is even more reason to recommend a logarithmic transformation for endpoints in ERAs than for endpoints in food-feed assessments, because the data are often counts, and the treatment effects are usually multiplicative in nature. Hence, for example, a toxin may cause sub-lethal effects in a certain percentage of a non-target organism population; after a logarithmic transformation the effects become additive on the new scale, facilitating analysis. Fifth, compositional field trials in food-feed applications represent a fairly restricted range of designs compared to the much wider variety of possible studies in ERA. Sixth, experience shows that coefficients of variation are often smaller for food-feed endpoints than for ERA endpoints, and that experimental designs for food-feed tend to have relatively high efficiency, compared to those for ERA that have a greater requirement for the confidence given by a power analysis. Seventh, by their nature, limits of concern may well be non-symmetric, because there might be less reason for concern if the effect of a GM plant was to increase the population of an non-target organism (NTO) rather than if it decreased it; by assumption this is not usually the case for equivalence limits estimated in food-feed risk assessment. Here it is noted briefly that a common medical equivalence test adopts a one-sided approach termed 'non-inferiority' (*e.g.* Laster and Johnston, 2003).

All the above considerations apply equally to field trials as well as to laboratory experiments.

An attempt to summarize some of the above issues and to contrast the ERA approach and that for food-feed risk assessment is shown in Figure 1.

STATISTICAL RECOMMENDATIONS FOR ENVIRONMENTAL RISK ASSESSMENT

Choice of comparators

As outlined above, for ERA, the basic comparison of interest focuses on the difference between the GM and the appropriate (usually near-isogenic) comparator. However, whilst typically consumers would be reassured by a conclusion of equality from statistical tests in food-feed risk assessment, such a conclusion is often not credible in NTO ERA. Furthermore, equality is clearly neither reasonable nor possible when genetically modified insect-resistant (GMIR) systems (*e.g. Bt* maize crops) are compared with a near-isogenic variety managed without the insecticides that would be typically applied to conventional, non-GM crops. Therefore, it is sensible also to consider extra comparators that help to place differences between the GM and its comparator into context. For example, Marvier et al. (2007) reported that a meta-analysis of 42 field experiments that indicated that non-target invertebrates were generally more abundant in *Bt* cotton and *Bt* maize fields than in non-transgenic fields managed conventionally with insecticides, but in comparison with insecticide-free control fields, certain NTO taxa were less abundant in *Bt* fields. The use of extra comparators that help to contextualize such differences

between the GM and its major comparator is recommended.

However, it is recognized that conventional management is difficult to define for some events, such as a GM plant with a composition-modification trait. Also, that conventional management must be site- and year-specific. Care is undoubtedly required in the definition of treatments in studies of GMIR systems, particularly with regard to typical management, because the scientific threshold for action may differ from that of the market, and at different sites there may be different typical practices, particularly if sites are in different regions. The use of detailed record-keeping and published agronomic audits by trained personnel may help to give confidence that management practices are appropriate (Champion et al., 2003).

In particular, for the case of herbicide-tolerant GM plants, three test materials should be compared: the GM plant exposed to the intended herbicide, the conventional counterpart treated with conventional herbicide(s) and the GM plant treated with the same conventional herbicide(s). Such comparisons allow the identification within the assessment of which elements of the altered agricultural practices influence the studied endpoints (the GM plant *per se* or the GM plant in concert with the intended herbicide). The appropriate conventional counterpart for stacked events should be selected in accordance with the principles defined in other sections of this document. In addition, single parental GM lines or GM lines containing previously stacked events that have been fully risk assessed may also be included as additional comparators.

It is essential that if extra comparators are employed, these should be fully integrated within the experimental design, randomized and replicated in the same way as the GM and near-isogenic comparator. Commercial varieties might be useful to provide information concerning variability, but their inclusion in field trials should certainly not be mandatory, particularly because the need for adequate plot sizes puts major constraints on the number of treatments that can be randomized as integral components of a trial. If included, commercial reference varieties should be treated conventionally, not untreated. Of course, detailed information should be provided to justify the choice of any additional comparators and the management employed.

Specifying effect sizes for environmental risk assessment

The size of the effect that it is desired to detect should be stated explicitly for each endpoint sampled. At first sight, it might be thought that it may not always be easy to specify the effect size that it is believed is a threshold for environmental harm for those who wish to demonstrate the safety of GMOs. Of course, the effect of a given percentage decline in abundance in some life stage due, say, to a GMIR system, may not lead to an equivalent decline in the population as a whole. Hence, there may be a need for upscaling to the landscape and higher scales, in order to assess and place into context smaller-scale effects found at the plot or field level. Also, it is reasonable to assume that the effect size may vary by taxa, crop type or functional group.

However, several reasons point to why the specification of effect size is considered an achievable task. First, a major part of the risk assessment dossier is risk characterization; such characterization cannot be done without relating effect to potential harm. Second, it is no more than good scientific practice when planning an experiment to have a good idea of the size of effect that the experiment is designed to detect. Third, in any event, a power analysis, which is deemed essential (see below), must always involve the specification of the magnitude of the effect size that the experiment is designed to detect. Fourth, the approach is routine in other risk assessment administrations, notably the United States of America (for example, see US Environmental Protection Agency (2006) or FIFRA SAP (2000)). There is considerable flexibility in the choice; the effect size may or may not be asymmetric and be placed on either the natural scale, the multiplicative scale, or some other scale, on a case-by-case basis.

Risk managers need to understand clearly that there is a difference between statistical significance and biological significance (Perry, 1986). In particular, for a given effect size, the p-value of a difference test is not constant, but decreases quickly as sample size increases. Hence, if there is any non-zero effect a difference test will always detect it and be statistically significant if the sample size is large enough. As an example, if the difference between two groups of size N is $d = \sigma/4$ (where σ is the standard deviation in both groups), then the difference test will not be significant for a sample size of $N = 50$ (for which $p = 0.21$) but will be highly significant for a sample size of $N = 500$ (for which $p = 0.001$).

In conclusion, for each study, it is recommended that the size of the effect that it is desired to detect should be stated explicitly for each endpoint sampled. The effect size should be linked to the minimum relevant ecological effect that is deemed biologically significant. Again, full justification for the choice of scale and effect size should be given.

Statistical power for environmental risk assessment

For the difference test, statistical power is the probability of detecting an effect of a given size, when such a real effect exists. Power is often quoted as a percentage. The risk assessor must ensure that an evaluation has sufficient power to provide reasonable evidence. A level of 80%

is usually considered to be an acceptable level for statistical power (ICH, 1998; Marvier, 2002), although it must be appreciated that for ecological studies this may be an aspiration that can only be achieved in well-resourced and extensive experiments (Perry et al., 2003). Statistical power depends, amongst other things, upon the chosen experimental design, the magnitude of the variety difference, the baseline variability of the experimental units, the critical probability level of the test and the replication of the experiment. In general, other things being equal, a decrease in this critical probability level *i.e.* α, the type I error rate will generally lead to a decrease of power.

A power analysis, executed when the study is being planned and prior to its start (Perry et al., 2003), may be used to estimate power, to choose appropriate replication, and to give confidence that the experiment will detect any significant effect that is present. As Marvier (2002) expresses it: "such details of risk assessment studies could greatly increase the public's ability to evaluate industry's claims of safety". However, there have rightly been many criticisms of the calculation of statistical power from the experimental data obtained (so-called retrospective or *post-hoc* power analysis). Firstly, power depends on several parameters, notably the baseline variability of the experimental unit, here denoted σ^2. Whereas the magnitude of the variability of estimates of treatment means is proportional to σ^2 the magnitude of the variability of estimates of variability greatly exceeds that, often by several orders of magnitude, being proportional to σ^4. For the range of experiments envisaged here, estimates of variability, and hence of power, may be expected to be imprecise. Schuirmann (1987) showed how, for technical statistical reasons concerned with the lack of convexity of what is termed the rejection region, retrospective power analysis may lead to completely wrong inferences, incompatible with common sense. Further problems associated with such a strategy were identified, for example by Hoenig and Heisley (2001) and by Walters (2008). Andow (2003) argued that it should not be used in publications in this journal. Hence, it is recommended that *post-hoc* or retrospective power analysis should never be used in the risk assessment of GMOs.

The provision of a power analysis is recommended for the difference test for each experiment done to support an ERA, to aid transparency and public confidence that the consumer risk is well defined and low. This should ideally be based on the provision of 80% power for a defined treatment effect size with a 5% type I error rate. The power analyses should use only information verifiable as available prior to the study, and, in particular, should not use data from the study itself. They should be done at the planning stage of the study.

Experimental design for environmental risk assessment

The following section contains eight sub-sections designed to give guidance to experimenters on how to design efficient experiments for NTO risk assessment. It may be read in conjunction with the checklist that forms an appendix to this document.

(i) General principles for all studies – sample sizes

It is recommended to give a quantitative justification of the sample sizes used for each experiment (replication, number of experimental units, number of each type of blocking factor such as cages, cohorts, sites, years). It might help in the planning of studies to estimate the expected width of the confidence interval(s) of the mean difference(s) between the GM and its comparator(s), although this is not deemed essential. It is recommended that a checklist be provided and completed (see Appendix to this document) to guide those who wish to demonstrate the safety of GMOs through a series of questions intended to foster sound experimental designs.

(ii) General principles for all studies – use of positive controls

It is recommended to consider the inclusion of a positive control in each study, for two possible reasons. First, to demonstrate *post-hoc* that the study was capable of detecting the desired effects, for example that there was sufficient population density of NTOs present in the experimental area. The form of the positive control should be decided on a case-by-case basis and may vary with the GM trait; examples might be: current management, reference substance, broad-spectrum insecticide (Marvier et al., 2007). It may be necessary to sample both before and after applications, and possibly to repeat applications. Whilst positive controls might be external to the experiment, for example on a single unrandomized plot, it must be realized that in that case, no statistical tests can be applied, so it would not be possible to distinguish whether any difference recorded between the positive control and a treatment was merely random error or a true effect. Hence, it is recommended that the following treatments should be fully randomized and replicated with the experimental design: the GMO; its near-isogenic comparator untreated with insecticide; and its near-isogenic comparator treated conventionally with insecticide (where appropriate). The second reason for including positive control(s) is that they may allow any differences that are detected between the GMO and its near-isogenic comparator to be placed into

a proper agronomic context (*e.g.* Marvier et al., 2007). This is particularly the case for GMHT (genetically-modified herbicide tolerant) or GMIR systems, where an *untreated* near-isogenic comparator would not provide a realistic comparison, because in usual agronomic practice this would be treated if necessary with, respectively, herbicides or insecticides. Indeed, meta-analyses (Marvier et al., 2007) of GMIRs have shown that, on average, abundances of NTOs are ranked according to treatment as near-isogenic treated conventionally (smallest) through GMIR to near-isogenic untreated (largest). Although baselines for biodiversity in arable ecosystems differ between regions, it seems sensible for ERA to assess risk of harm due to GMOs relative to existing such conventional baselines rather than to agronomically unrealistic paradigms.

(iii) Field trials – size of plots

The sizes and scales of experimental units vary widely, from paired fields and grower fields at the larger scale, to small plots of a few square meters at the small scale. Larger plots produce more realistic and representative data, particularly for relatively mobile NTOs such as pollinators, and may suffer less from treatment interference between plots. However, they require more land and thus limit possibilities for adequate replication. The limitations of land may be severe if GM plants are grown, and this also restricts the number of treatments that can be studied. In particular, it is a further disincentive to employ several commercial varieties as extra comparators. Furthermore, for grower fields, there may be a need to (i) pay compensation, hence increasing the cost of the study, and (ii) to exercise care in the definition of management. Grower fields or paired fields are probably of most use for post-commercial studies, particularly for what are termed 'tier 4 studies' (large-scale monitoring or mitigation studies). For pre-commercial experimentation, smaller plots, where variation could be controlled and defined treatments imposed more easily, are more appropriate. It is recommended to separate plots within sites, often by strips of bare soil of specified width, and to sample towards the center of plots. Attempts to determine optimum plot size (*e.g.* Prasifka et al., 2005; Winder et al., 1999) require considerable resources and the ephemeral nature of aggregation patterns means that inferences may be subject to major temporal heterogeneities. A related problem is how to determine the optimal intensity of sampling within a plot; again, for most cases, the resources required (Perry, 1989) may not make its determination prior to studies cost-effective, although Clark et al. (2007) made progress using a components of variance approach.

(iv) Field trials – optimal design

It is recommended to choose the precise form of design on a case-by-case basis. However, a randomized block design is usually appropriate.

(v) Field trials – additional factors and split-plot designs

When there are factors that may interact with the main factor of interest, which is always the contrast between the GM and its (usually near-isogenic) comparator, then these may need to be built into the experimental design. For reasons of practicality, the introduction of the extra factor may need to be done in a split-plot design. For example, if the extra factor is a sprayed application of a chemical, then the equipment used may restrict the plot size to be larger than some minimum width, perhaps corresponding to the spray boom. As an example, consider an herbicide-tolerant oilseed rape system, suppose the NTOs studied are Collembola, and it is thought that timing of herbicide application may have a major effect on the difference in observed abundance between GMHT rape and its near-isogenic comparator (n-i c). Suppose further that at each site there are four replicate blocks. A suitable experimental design would be to randomize all four combinations of the main factor, GMHT *versus* n-i c, and an extra treatment factor with two levels, early application *versus* late application, over each of the four plots in each of the four blocks (Fig. 2A). However, this may not be possible if the plots are too small for the spraying equipment. Then, a split-plot design can be used. In this design, the extra factor is first randomized onto two main plots within each of the blocks (Fig. 2B), and then the main factor of interest, GM herbicide-tolerant *versus* n-i c, is randomized onto two sub-plots within each of the eight main plots (Fig. 2C). It is important to realize that for reasons of maximizing power, the main factor of interest should always be randomized to sub-plots (giving replication of 8 in this instance), whilst the factor with which it is believed to interact should always be randomized to main plots (giving replication of 4). The ANOVA for the completely randomized design in Figure 2A is sketched in Table 1A; for the split-plot design in Figure 2C see Table 1B.

(vi) Field trials – use of same plots in successive years

Designs often use part of a field in one year and a different part of the same field in the next year; this is satisfactory. More rarely, exactly the same plots are used over more than one year at a particular site. Considerable care is required for this latter case, because of possible residual effects in year 2 from the treatment applied

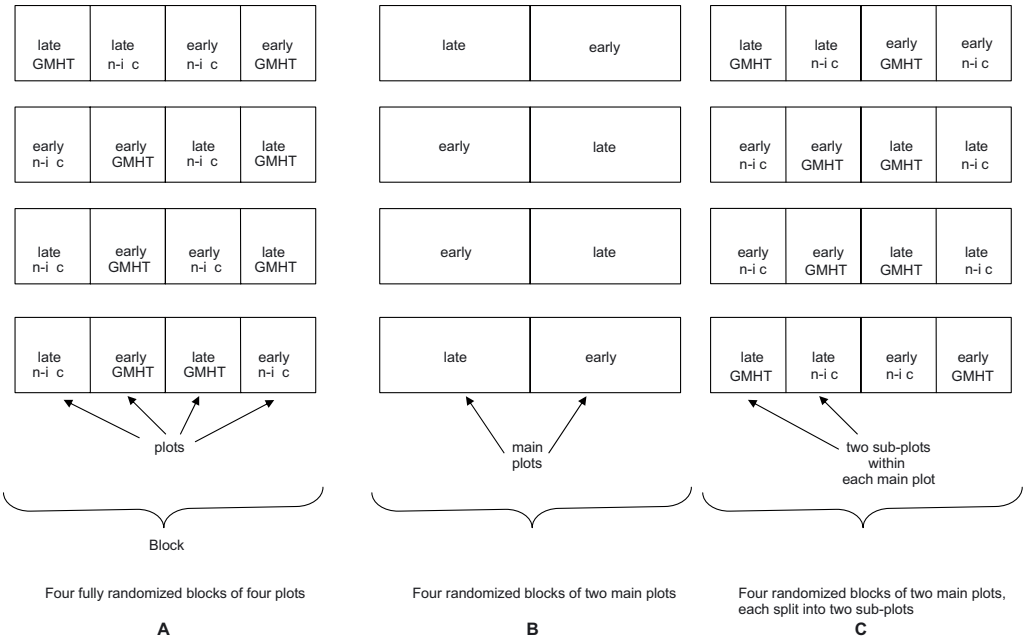

Figure 2. Example of design for NTO study with two factors. The major factor with two levels is the contrast: GM herbicide-tolerant oilseed rape *versus* its near-isogenic comparator (n-i c). An extra factor, also with two levels, is the contrast defined by the timing of herbicide application: early *versus* late. A: Each treatment combination is completely randomized to the four plots within each of four replicate blocks. B: Example of similar design where spray applications cannot be applied to plots smaller than twice the width of those in A. In the first stage of allocation of treatments to plots, the timing factor is randomized to two main plots within each of the four replicate blocks. In the second stage, C, the full design is shown, after the GM effect factor has been randomized to two sub-plots within each of the eight main plots. See also Table 1.

in year 1. Indeed, it is recommended that in general such designs should be avoided, unless the biology of the NTO requires impacts over more than one growing season to be tested. In that case the design is suitable to estimate the residual effects that require study, since that element is explicitly incorporated into the design. Of course, a proper statistical analysis of such designs should account for possible correlations between repeated measurements from the same experimental units.

(vii) Field trials – replication

The question of how many replicates to use within each site differs in kind from the question of how many sites and how many years to employ. The principle applied for food-feed assessments is that each field trial at a site should have sufficient replication to be able to yield a stand-alone analysis if required, even though the main analysis averages over the complete set of field trials at all sites and years. It is recommended that this principle should be retained for ERA assessments. Replication within sites is motivated by the need for sufficient power and efficiency to achieve such stand-alone analyses. Under the case-by-case approach, it is not possible to recommend a minimum number of replicates per field. However, this will be determined by the recommended mandatory power analysis (see above). Notwithstanding this, it is most unlikely that less than three replicates per site would provide an adequate design.

By contrast, there is an additional need to replicate over sites and years, namely to achieve representativeness across geography and climate. The choice of sites of the trials must represent as fully as possible the range of receiving environments where the crop will be grown, reflecting relevant meteorological, soil and agronomic conditions; the choice must be justified explicitly. Environmental conditions may vary greatly between sites, even locally, and in general years vary even more than sites. It is recommended that each field trial be replicated over at least two years, within each of which there should be replication over at least three sites, and that explicit justification be given for any deviation from any of these recommendations for replication.

The use of data from different continents could be informative, so long as the sites within these continents are each representative of where the crop will be grown, reflecting relevant meteorological, soil and

Table 1. ANOVA structure for (a) the fully randomized and (b) the split-plot designs.

A		**B**	
Source	*Degrees of freedom*	*Source (main plots)*	*Degrees of freedom*
Blocks	3	Blocks	3
Timing (Early *versus* Late)	1	Timing (Early *versus* Late)	1
GM (GMHT[1] *versus* n-i c[2])	1	Error (main plots)	3
Interaction between GM factor and timing factor	1	Total (main plots only)	7
Error	9	*Source (sub plots)*	*Degrees of freedom*
Total	15	Main plots	7
		GM (GMHT *versus* n-i c)	1
		Interaction between GM factor and timing factor	1
		Error (sub plots)	6
		Total (sub plots)	15

[1] GM herbicide-tolerant.
[2] Near-isogenic comparator.

agronomic conditions. However, within Europe it is recommended that strong and explicitly-argued justification be provided for not having field trials in EU Member States.

(viii) Laboratory experiments

Experimental designs are recommended to follow the general principles set out above, and in addition to conform to accepted international standards and protocols such as those published, for example, by OECD or similar organizations specializing in ecotoxicology.

Statistical analysis for environmental risk assessment

The following section contains four specific and one general sub-sections designed to give guidance to experimenters on the statistical analysis of data from experiments for NTO risk assessment.

(i) Statistical protocols

It is recommended to provide a statistical analysis protocol for each study which should include full information on the questions the study addressed, what the measurement endpoints were and why they were included, what form the analysis took, why that form of analysis was chosen, the criteria for identifying outliers, reasons for any transformations used, justification for distributional assumptions, assumptions made, and any other relevant information. For field trials, the protocol should include a clear and explicit statement concerning the minimum levels of abundance acceptable for each taxa sampled, below which results would lack credibility, with full justification for the values chosen.

(ii) Equivalence testing for environmental risk assessment

At least one pair of limits of concern should be set for each endpoint; the values set should reflect, and in most cases be identical to the desired effect size which will have been defined explicitly. As much relevant information as possible should be employed to set such 'limits of concern', including information from historical databases. However, it is recommended that further pairs of limits may be set if desired; an equivalence test should then be performed for each pair of limits.

It would be helpful and facilitate clarity if the presentation of the results followed the principles of graphical presentation adopted for food-feed assessment detailed

in EFSA (2009b). There, both the difference test and the equivalence test are implemented using the well-known correspondence between hypothesis testing and the construction of confidence intervals. The graph combines both tests, difference and equivalence, in a user-friendly form of presentation. Presentation in the form of confidence intervals gives extra information compared with a mere statement of whether a test is or is not significant. In the case of equivalence testing, the approach used follows the two one-sided tests (TOST) methodology (*e.g.* Schuirmann, 1987) by rejecting the null hypothesis only when the entire confidence interval falls between the equivalence limits. The choice of the 90% confidence interval corresponds to the customary 95% level for statistical testing. The graph shows the line of zero difference between the GM and its major comparator and, for each endpoint: the lower and upper limits of concern, the mean difference between the GM and its comparator, and its confidence interval, as in Figure 1B. Note that the line of zero difference on the logarithmic scale corresponds to a multiplicative factor of unity on the natural scale. The horizontal axis may then be labeled with values that specify the change on the natural scale. In the case of logarithmic transformation, changes of $2\times$ and $^1/_2\times$ will appear equally spaced on either side of the line of zero difference.

For studies that use extra comparators, the analysis should encompass separate difference tests (between the GM and each of its different comparators) and separate equivalence tests (between the GM and each of its different comparators), and these should be reported similarly.

As a simplified numerical example to facilitate understanding, consider an imaginary experiment involving a GMIR maize plant and a non-target lepidopteran as the NTO selected for study, with the endpoint as abundance as measured by observation along a transect walk, and with three treatments to be compared: the GMIR maize variety; its near-isogenic comparator (denoted n-i c) untreated with insecticides; and a positive control comprising the n-i c managed conventionally with insecticides (denoted conv). All abundances, c, are to be transformed to $\log_{10}(c+1)$ for analysis. Suppose the mean abundances on the logarithmic scale (with geometric means on natural scale in brackets) for the GMIR, n-i c and conv are, respectively, 0.600 (3.98), 1.050 (11.2) and 0.50 (3.16), and suppose for the sake of simplicity that replication was equal and therefore all differences between pairs of means have confidence intervals with the same width of 0.110. In contrast to the graph recommended for food-feed risk assessment, where all values were referred to the same zero line defined by the conventional counterpart, here it would seem to be more sensible for all values to be referred to the same zero line defined by the GMIR. Further, suppose that prior to the experiment it was decided that an increase in abundance of this lepidopteran of greater than 65% would constitute a biologically significant effect and define the upper limit of concern, whilst a similar magnitude decrease in abundance would define the lower limit of concern. Then, assuming all abundances, c, are transformed to $\log_{10}(c + 1)$ for analysis, the upper and lower limits of concern are ±0.217. Hence, if all values were referred to the baseline, the mean of the logarithmically transformed abundances of the GMIR, of 0.600, the graph in Figure 3 would be obtained. The graph shows the difference between untreated n-i c and GMIR centered on 0.45 (indicating an almost three-fold decrease in abundance of GMIR relative to its untreated n-i c) and that between conv and GMIR centered on –0.10. Note that the line of zero difference on the logarithmic scale corresponds to a multiplicative factor of unity on the natural scale. Regarding inferences, clearly the lack of overlap of the confidence interval for the difference between n-i c and GMIR with the zero line implies this difference test is significant; similarly the difference between conv and GMIR is not significant. Even allowing for the uncertainty expressed through its confidence interval the almost three-fold difference between n-i c and GMIR can be seen to greatly exceed the 1.65-fold effect size defined by the upper limit of concern, so the conclusion from the relevant equivalence test is that the magnitude of the difference between n-i c and GMIR is of ecological significance. By contrast, the entire confidence interval for the difference between conv and GMIR lies within the upper and lower limits of concern, so the conclusion is that any difference does not represent a significant ecological effect. The importance of contextualizing such inferences on the risk assessment of GMIR plants through the use of a positive control has been stressed by Marvier et al. (2007).

The use of a statistical mixed model is an important feature of analysis for food-feed assessments because of the need to estimate the natural variation of the commercial varieties. However, for ERA it is recommended that equivalence limits are set explicitly and the use of commercial varieties is not mandatory. Hence it is not recommended that mixed models be required forms of analysis, although their use should be encouraged if they are appropriate.

(iii) Testing for interactions

It is recommended that the presence of a treatment × site interaction should always be tested. If data from sites in different continents are included in the same set of field trials, then the continent × treatment interaction should be tested where possible. This will reveal if the level of risk varies significantly across sites and/or continents.

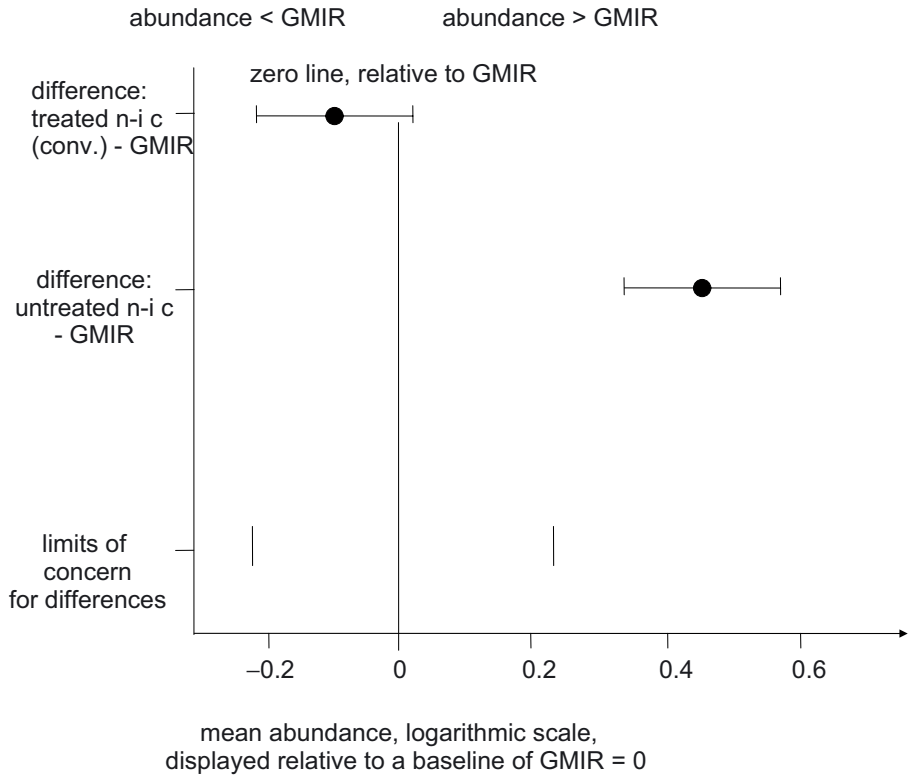

Figure 3. Imaginary worked example of graphical data analysis with equivalence tests for an ERA of an NTO. Three treatments are compared: a GMIR variety; its near-isogenic comparator (n-i c) and a positive control comprising a non-GM commercial variety managed with conventional insecticides (conv). Mean abundances (filled circles) and confidence intervals (bars) are shown on a logarithmic scale. Limits of concern for multiplicative differences for the equivalence test are shown as vertical lines. All values are referred to a zero baseline defined by the mean of the GMIR.

(iv) Multiple comparisons

With many endpoints, or many species in a study, there will be a multiplicity of tests. It is accepted that setting the size of the difference test at the 10% level will lead to a large proportion (about 10%) of tests being found to be significant by chance alone. Of course, a large proportion of significant differences is not considered a sufficient reason for safety concern *per se*, unless the proportion is significantly larger than the 10% expected. Safety concerns may also be raised if the differences follow some systematic pattern such that endpoints of a certain type form a cluster that are all statistically significant. When many species have been included in a study the results of all species for which sufficient records have been obtained should be reported, not just those deemed to be of particular biological or statistical interest.

The principle behind the statistical tests of difference and equivalence is to provide information with quantified uncertainty that may be used by biologists in risk characterization of those endpoints for which differences or lack of equivalence are found and put into context. There is a well-known distinction between statistical and biological significance, and it should not be the function of statistical analysis to lead to a decision-theoretic approach, in which data is fed into very proscriptive protocols that yield decisions on safety. Indeed, because of the many difficulties of formalizing the complete safety assessment, a full statistical decision-theoretic approach is not feasible. This would, in any case, be counter to the principle of the case-by-case approach. Rather, the judgments on biological significance are left to experts.

As for food-feed assessment, the following approach is recommended: the independent univariate evaluation of single endpoints, a joint graphical presentation, and the simple reporting of the frequency of significant results in the set of investigated endpoints.

(v) General issues

Meta-analysis is a statistical technique to assess simultaneously the results of several studies. The use of

meta-analysis (*e.g.* Marvier et al., 2007) is recommended, particularly to quantify studies that may not all have the power to be individually significant, in the statistical or biological sense. It can also serve to provide an overview of broad patterns when individual studies may appear to contradict each other.

Diversity indices are not recommended for general risk assessment in pre-commercialization studies. Diversity indices might be useful but are not essential for post-commercial studies. Whilst it may be desirable to measure diversity, the idea that it can be represented as a single number is overly naive for environmental impact studies. Studies of diversity are complex and require considerable technical expertise; they are most suited to tier-4 studies of assemblages and communities at large scales.

A multivariate approach, commonplace in community ecology, is not essential but may be very useful, particularly for: (a) post-commercial studies of the composition of assemblages and their community dynamics and (b) in data screening, for the identification of the correlation structure between endpoints, and to provide a holistic approach to summarizing complex data structures. For general ecological research multivariate methods are standard, but for small-scale risk assessment studies that demand the study of specific species, other methods are recommended to comprise the primary method of analysis. Where parametric assumptions are of dubious validity for multivariate tests, permutation tests provide a useful alternative. Principal Response Curves (van den Brink and ter Braak, 1999) and explanatory ordination methods are recommended as methods if multivariate approaches are employed. MANOVA should usually be avoided, since the distributional assumptions underlying the method can rarely be justified empirically.

ACKNOWLEDGEMENTS

This paper is based on discussions at the Fourth Biosafenet Seminar, entitled "The Statistical Design and Analysis of Field Trials for Assessing the Risk Associated with GM Plants, with a Focus on Non-Target Organisms", which was held at the ICGEB Biosafety Outstation in Ca' Tron di Roncade (Italy) on 12–15 January 2009. We thank the participants for their helpful comments. The Biosafenet project is supported by the European Commission (contract # 043025).

Received February 22, 2009; accepted June 25, 2009.

REFERENCES

Andow DA (2003) Negative and positive data, statistical power, and confidence intervals. *Environ. Biosafety Res.* **2**: 75–80

Champion GT, May MJ, Bennett S, Brooks DR, Clark SJ, Daniels RE, Firbank LG, Haughton AJ, Hawes C, Heard MS, Perry JN, Randle Z, Rothery P, Skellern MP, Scott RJ, Squire GR, Thomas MR (2003) Crop management and agronomic context of the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. *Phil. Trans. R. Soc. Lond. B* **358**: 1801–1818

Clark SJ, Rothery P, Perry JN, Heard MS (2007) Farm Scale Evaluations of herbicide-tolerant crops: assessment of within-field variation and sampling methodology for arable weeds. *Weed Res.* **47**: 157–163

EFSA (2006) Guidance Document of the Scientific Panel on Genetically Modified Organisms for the risk assessment of genetically modified plants and derived food and feed (Question No EFSA-Q-2003-005). *The EFSA Journal* **99**: 1–100

EFSA (2009a) Updated guidance document of the Scientific Panel on Genetically Modified Organisms (GMO) for the risk assessment of genetically modified plants and derived food and feed. http://www.efsa.europa.eu/EFSA/efsa_locale-1178620753812_1211902010430.htm

EFSA (2009b) Scientific Opinion on Statistical considerations for the safety evaluation of GMOs, on request of EFSA. *EFSA Journal* **1250**, 62 p. http://www.efsa.europa.eu

EFSA (2009c) General mandate – aspects of the environmental risk assessment (ERA) and the ERA guidance. Question EFSA-Q-2008-262. http://www.efsa.europa.eu/EFSA/efsa_locale-1178620753812_1178697446987.htm

FIFRA SAP (2000) Report of the Federal Insecticide, Fungicide, and Rodenticide Act. http://www.epa.gov/scipoly/sap/meetings/1999/december/report.pdf

Hill RA, Sendashonga C (2003) General principles for risk assessment of living modified organisms: Lessons from chemical risk assessment. *Environ. Biosafety Res.* **2**: 81–88

Hoenig JM, Heisley DM (2001) The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am. Stat.* **55**: 19–24.

ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials E9. Current Step 4 version dated 5 February 1998

Laster LL, Johnson MF (2003) Non-inferiority trials: the 'at least as good as' criterion. *Stat. Med.* **22**: 187–200

Marvier M (2002) Improving risk assessment for nontarget safety of transgenic crops. *Ecol. Appl.* **12**: 1119–1124

Marvier M, McCreedy C, Regetz J, Kareiva P (2007) A meta-analysis of effects of *Bt* cotton and maize on nontarget invertebrates. *Science* **316**: 1475–1477

Niazi S (2007) Handbook of Bioequivalence Testing. CRC Press, ISBN: 0849303958, 569 p

Perry JN (1986) Multiple-comparison procedures: a dissenting view. *J. Econ. Entomol.* **79**: 1149–1155

Perry JN (1989) Review: population variation in Entomology: 1935–1950. I. Sampling. *The Entomologist* **108**: 184–198

Perry JN, Rothery P, Clark SJ, Heard MS, Hawes C (2003) Design, analysis and power of the Farm-Scale Evaluations of Genetically-Modified Herbicide-Tolerant crops. *J. Appl. Ecol.* **40**: 17–31

Prasifka JR, Hellmich II RL, Dively GP, Lewis LC (2005) Assessing the effects of pest management on non-target arthropods: the influence of plot size and isolation. *Environ. Ent.* **34**: 1181–1192

Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biop.* **15**: 657–680

Tempelman RJ (2004) Experimental design and statistical methods for classical and bioequivalence hypothesis testing with an application to dairy nutrition studies. *J. Anim. Sci.* **82**: E162–E172

US EPA (2006) Guidance for the Data Quality Objectives Process. EPA QA/G-4. http://www.epa.gov/QUALITY/qs-docs/g4-final.pdf

van den Brink PJ, ter Braak CJF (1999) Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress. *Environ. Toxicol. Chem.* **18**: 138–148

Walters SJ (2008). Consultants' forum: should post hoc sample size calculations be done? *Pharm. Stat.* **8**: 163–169

Wellek S (2002) Testing Statistical Hypotheses of Equivalence. CRC Press, ISBN: 1584881607, 284 p

Winder L, Perry JN, Holland JM (1999) The spatial and temporal distribution of the grain aphid *Sitobion avenae* in winter wheat. *Entomol. Exp. Appl.* **93**: 277–290

Yata K (2008) Two-stage equivalence tests that control both size and power. *Seq. Anal.* **27**: 185–200

APPENDIX

Checklist for all experiments concerning Environmental Risk Assessment

List all the **questions** that this experiment is asking, in words:

Now frame the **questions** listed above formally, in the form of null hypotheses of all the tests that will be done using the experimental data:

What is the **experimental design**?
(tick the design that applies):
- completely randomized
- randomized block
- latin square
- split-plot
- incomplete block (balanced)
- incomplete block (unbalanced)
- other (please specify):

Describe the **experimental units** and give details of the **blocking structure**.

Give the experimental unit (*e.g.* plot, cage, individual insect):

Give its dimensions:

Write in words the details of the structure of the experimental units, often termed the blocking structure (*e.g.* 4 main plots per randomized block, each split into 3 sub-plots):

State what factors represent the blocking structure (*e.g.* sub-plots, main plots, blocks):

How many levels does each factor have (*e.g.* sub-plots 2; main plots 3; blocks 5), and what are they:

State how are the blocking factors nested within or crossed amongst each other (*formulae* or *words* may be used). For example, in *Genstat formulae*: blocks/main plots/sub plots, or in words: sub plots nested within main plots nested within blocks):

State briefly why this blocking structure was chosen:

State how many samples are to be taken from each experimental unit:

State whether repeated measurements will be taken from the same experimental unit:

Describe the **treatments** that will be applied in the experiment.

What factors represent the treatment structure (*e.g.* varieties, irrigation):

For each factor:
- state whether it is qualitative, ordered categorical or quantitative
- give the number of levels
- and list the levels

(For example:
varieties – qualitative - 2 levels – GMHT, near isogenic;
irrigation – ordered categorical - 3 levels – none, weekly, daily as required.):

For designs with more than one treatment factor, state whether any, and if so which, interactions are required to be estimated:
(For example: Yes, it is required to estimate the variety × irrigation interaction.)

State how the treatment factors will be randomized to the experimental units specified in the blocking structure in the previous page.

(For example: varieties will be randomized to sub-plots within main plots; irrigation will be randomised to main plots within randomized blocks.)

State which of the factors and their levels represent controls, and describe the control(s) briefly, indicating if any are positive controls:

(For example: variety – level 2 – near isogenic.)

Answer some more detailed **statistical questions**.
State, if applicable and if known, which of the factors in the experiment will be regarded as fixed and which as random:

For each of the questions listed in the first box of this checklist, state:
(i) the size of the effect that the experiment is designed to detect;
(ii) the expected power of the experiment to detect this effect for a 5% size of test.

State the number of years and sites per year the experiment is designed to be replicated over:

What scale are the effects in the experiment assumed to be additive on (*e.g.* natural untransformed scale, probit scale, logarithmic scale, etc.):