Bulletin OEPP/EPPO Bulletin (2012) **42** (3), 367–381 ISSN 0250-8052. DOI: 10.1111/epp.2610

European and Mediterranean Plant Protection Organization
Organisation Européenne et Méditerranéenne pour la Protection des Plantes

PP 1/152(4)

Efficacy evaluation of plant protection products
Evaluation biologique des produits phytosanitaires

Design and analysis of efficacy evaluation trials

Specific scope

This standard is intended for use in association with EPPO Standards of set PP 1 *Standards for the efficacy evaluation of plant protection products* and provides detailed advice on the design and analysis of efficacy evaluation trials.

Specific approval and amendment

First approved in 1989–09.
First revision approved in 1998–09.
Second revision approved in 2006–09.
Revision mainly to reflect zonal assessment approved in 2012–09.

Introduction

This standard is intended to provide general background information on the design and analysis of efficacy evaluation trials. The EPPO Standards for the efficacy evaluation of plant protection products provide more detailed instructions on such trials for individual host/pest combinations. The set-up of a trial is first considered (experimental design, plot size and layout, role and location of untreated controls). The nature of observations to be made is then reviewed (types of variables, modes of observation). Finally, suggestions are made on the statistical analysis of the results of a trial and of a trial series (estimates of effects, choice of the statistical test, transformation of variables). Appendix 1 gives examples of scales used in the EPPO standards.

What follows is intended to give an outline of good statistical practice in the analysis of data. It is not, and cannot be, a prescription for all analyses, and cannot cover all situations. Practitioners should never underestimate the need for professional statistical advice. It is important for practitioners to understand the advice they receive. It is often better for them to perform a simple analysis that they can report and defend with confidence, than to accept advice that leads to an analysis that they may understand only partially. The bibliography at the end of these standards may be helpful. It gives several good texts that attempt to reveal the principles of good statistical practice, rather than to provide a series of statistical recipes to be followed blindly.

1. Experimental design

1.1 Experimental scope and objectives

Before the design of a trial is considered, its scope and objectives should be defined clearly, because these constrain the available choices of design. In practice, an iterative process is often used: scope and objectives are gradually adjusted to fit the experimental resources available. It is vital that the scope and objectives are updated to reflect decisions made during this process.

The scope of the trial reflects the range of practical outcomes that may result from the trial and which are relevant to its objectives. Part of the scope relates to the population which the trial is sampling. Another part determines the range of environmental conditions, crops, treatment chemicals, application methods and target pests which the trial is intended to test. The scope defines the context in which the experimental units and observations are studied.

The objectives of the trial should be in the form of questions about the treatments to which answers are desired. Typical answers will be ‘yes’ or ‘no’, a ranking of treatments or an estimate of a value.

The scope and objectives should form part of the trial protocol, as described in EPPO Standard PP 1/181 *Conduct and reporting of efficacy evaluation trials, including good experimental practice*. The planned experimental methods, design and analysis described below should also form part of the protocol.

2	7	3	7	8	3	5	4
1	2	6	2	2	3	4	6
8	4	5	4	6	8	1	5
1	5	7	8	1	7	3	6

Fig. 1 A fully randomized design. Each treatment (labelled 1–8) is replicated 4 times; individual treatment labels are assigned completely randomly to the 32 plots.

1.2 Types of design

EPPO Standards for the efficacy evaluation of plant protection products envisage trials in which the experimental treatments are the '*test product(s), reference product(s) and untreated control, arranged in a suitable statistical design*'. It is also envisaged that the products may be tested at different doses and/or application times. This applies particularly to the use of a higher dose in selectivity trials and dose-response studies in general.

Mono-factorial designs are appropriate for trials if the test product(s), reference product(s) and untreated control can be considered as different levels of a single factor, and if there are no other factors that require study. However, if, for example, the effect of each product in an efficacy trial is to be studied at different doses, then a factorial design may be used with, in general, all possible combinations of treatments from both factors represented. In this way, important interactions between the factors may be revealed and estimated.

The principal randomized designs which are likely to be used are: completely randomized and randomized complete block. These are illustrated below on the basis of a mono-factorial example with 8 treatments, i.e. 5 different test products, 2 reference products and an untreated control; each treatment is replicated 4 times.

1.2.1 Completely randomized design

The treatments in a completely randomized design (Fig. 1) are assigned at random to the experimental unit. This design is potentially the most powerful statistically (in the sense that there is a maximum chance of detecting a significant difference if it exists), because it allows retention of the maximum number of degrees of freedom for the residual variance. However, it is suitable only if the trial area is known to offer a homogeneous environment. If there is considerable heterogeneity between different parts of the trial area, residual variance may become unacceptably high, and it is better to use a design that accounts for this, such as a randomized complete block.

1.2.2 Randomized complete block design

A block is a group of plots within which the environment relevant to the observations to be made is homogeneous. In this design, the blocks are laid out deliberately so that plots within them are as uniform as possible before application of treatments. Usually, each treatment appears once and once only, within each block. The treatments are distributed randomly to the plots within the blocks, which act as replicates. The arrangement of treatments in each block should be randomized separately for each block. In the following examples (Figs 2–4), there are 4 blocks and 8 treatments. The layout of the blocks aims to control the heterogeneity of the site (e.g. slope, direction of work at sowing or planting, exposure, degree of infestation etc.), plants (size, age, vigour) or of the conditions occurring during the experiment (application of treatments, assessments). The layout of the blocks therefore requires some preliminary knowledge of the trial area. The arrangement of plots within blocks may be influenced by plot shape: long narrow plots are often arranged side-by-side, whereas, square plots may be laid out in other ways. However, blocks do not have to be placed side by side. If there is good preliminary knowledge of a field, this may be utilized by scattering blocks across the field, to account for previously observed heterogeneity (Figs 5 and 6). Although it is quite possible that in a randomized layout, treatments within a replicate may appear in treatment order, this is to be avoided wherever possible in the interests of unbiased evaluations. If there is extremely good preliminary knowledge, and it can be confidently assumed that conditions will remain the same for the experiment to be done, complex heterogeneity may be allowed for, and it is not even necessary for plots of the same block to be adjacent. For example, blocks may be broken up to account for a known patchy infestation of nematodes. In Fig. 6, plots within block 1 have been deliberately placed at points of visibly low infestation and plots within block 2 at points of visibly high infestation.

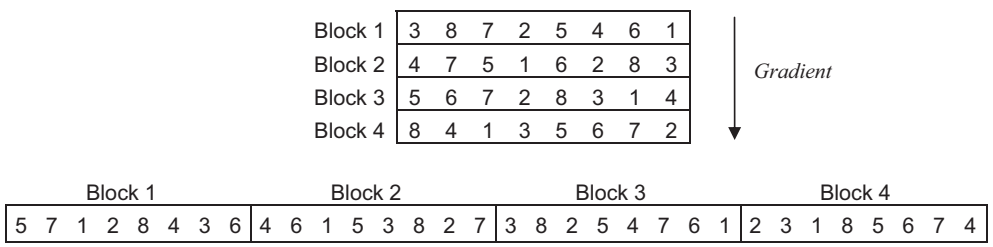

Fig. 2 Possible arrangement of blocks and plots in randomized blocks in field trials. An environmental gradient down the field is accounted for, either by arranging blocks down the gradient, or by placing blocks side by side. In each case, plots within blocks placed across the gradient are affected equally by the environmental variable.

Block 1	Block 2	Block 3	Block 4
3 1	8 1	8 2	3 7
6 4	2 6	6 5	1 6
8 5	7 5	3 1	5 8
7 2	3 4	7 4	4 2

Block 1	Block 2	Block 3	Block 4
1 3 2 4	2 7 8 5	4 1 6 3	7 3 5 1
8 5 6 7	4 6 1 3	5 7 8 2	4 8 2 6

Fig. 3 Possible arrangement of blocks and plots in randomized blocks in field trials. An alternative form of randomized block design for the situation when there is no obvious environmental gradient, but where heterogeneity is to be suspected because the maximum distance between plots within a block is relatively large. Here, the 8 plots are arranged relatively close together in a 4 × 2 rectangle, and the blocks are placed side by side.

Block 1	2 7 3 1 8 5 4 6	8 5 2 7 6 3 4 1	Block 2
Block 3	3 6 8 7 1 4 5 2	6 3 5 2 7 4 8 1	Block 4

Fig. 4 Another example of an arrangement for blocks and plots when, as in Fig. 3, heterogeneity is suspected but there is no obvious environmental gradient. Here, the 8 plots are again arranged relatively close together in a 4 × 2 rectangle, but the blocks themselves are arranged in a 2 × 2 grid.

Of course, the choice of design and the dimensions and orientation of the blocks used, if any, depend on the heterogeneity perceived in the trial area (e.g. for soil, slope, exposure, pest infestation, cultivar, etc.). Such variables are never entirely uniform, and a randomized block design in a moderately heterogeneous area will usually give more useful information on product performance than a fully randomized trial in an area thought to be homogeneous, but which subsequently transpires not to be. Block layout will also depend on plot size and shape (Figs 5 and 6). In general, smaller blocks are more effective in reducing heterogeneity. In trials with a large number of treatments other designs should be considered (e.g. lattice designs, incomplete block designs).

Randomized block trials carried out in different regions with distinct environmental conditions and/or in different years may in appropriate cases be considered as a trial series. In the statistical analysis it is then necessary to separate the additional between-sites variance from the variance between blocks, and also to estimate a site × treatment interaction, which may be of particular interest. Note that, in each separate trial, the treatments should be randomized anew within each block.

1.2.3 Split plot design

When a multifactorial trial is carried out, then the usual design is a randomized complete block, with each treatment combination occurring once in each block. However, sometimes one of the factors cannot be randomized fully to the plots in a block. For example, suppose a trial had 2 factors: product (with 4 levels, labelled 1–4) and cultivation equipment (with 3 levels, labelled A, B, C) and that plots were relatively small. Then the size of the machinery to apply the cultivation treatment may preclude full randomization over the 12 plots in each block. In that case, a split-plot design is recommended, where, in each block, subplots are associated together in groups of 4 to form 3 whole plots per block, the factor cultivation is randomized to these whole plots, and the factor product is randomized, separately, to subplots within whole plots (Fig. 7). With a split-plot design, a slightly more complex analysis of variance is required, in which there are 2 strata, each having a separate error mean square, against which to test the effect of the different factors and their interaction.

1.2.4 Systematic designs

Non-randomized, systematic designs are almost never suitable for efficacy evaluation purposes [they may be suitable in some very special cases (e.g. varietal trials on herbicide selectivity)]. In general, they are only suitable for demonstration trials.

1.3 Power of the trial

In planning experiments, it is important to consider the power required for any statistical tests that are to be performed. The power is the probability of detecting a given difference between treatments if this difference exists. The power depends on a number of parameters, among which are:

- The precision of the results (residual variation);
- Number of replicates, including any replication over sites.

A design should be chosen which gives a good chance of detecting, with statistical significance, a difference which is of practical importance for the comparison in which one is interested. One may also have the related requirement that confidence intervals on treatment estimates should be no more than some predetermined width. Before the trial is

Block 1	5 3 8 2 4 7 1 6	
Block 2	6 8 5 7 3 1 4 2	
Block 3	4 7 1 6 2 8 3 5	
Block 4	3 4 6 8 7 5 1 2	

Fig. 5 Possible arrangement of blocks and plots in randomized blocks in field trials. Blocks scattered across the field, according to previously observed heterogeneity.

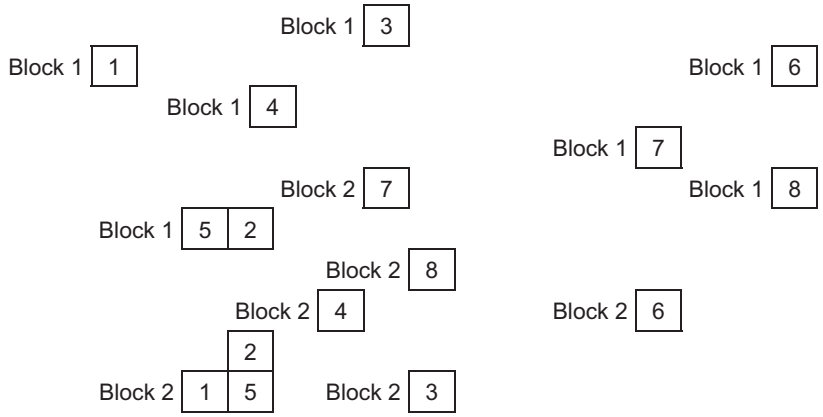

Fig. 6 Possible arrangement of blocks and plots in randomized blocks in field trials. Blocks scattered across the field, according to complex, previously observed heterogeneity.

	Block 1					Block 2			
	1A	2A	3A	4A	Whole plot 1	2B	4B	3B	1B
	3C	4C	1C	2C	Whole plot 2	2A	3A	4A	1A
	2B	3B	1B	4B	Whole plot 3	1C	3C	2C	4C

Fig. 7 An example of a split-plot design. The 2 treatment factors are: product (1, 2, 3, 4, randomized to subplots within whole plots) and cultivation method (A, B, C, randomized to whole plots within each of the 2 blocks)

started, the choice should be made between the performance of a single trial or of a trial series.

According to EPPO Standard PP 1/226 *Number of efficacy trials* the performance of a plant protection product should be demonstrated by conducting a number of trials in different sites, regions and years under distinct environmental conditions. Therefore to study the performance of a plant protection product a trial series may also be planned, conducted and analyzed (see also 3.4.1 for a definition of a trial series).

In general, there may be results from previous experiments to indicate the likely variability of observations. If such data exists, it is possible to make some judgement as to the design and size of the experiment needed to give the required power. Sometimes it is possible from theoretical considerations to determine the numbers required. For example, with binomial data, an upper limit can be put on the variability of proportions. Various computer-based or graphical systems are available to assist in determining the number of replicates needed. These use the magnitude of the difference required to be estimated, or the level of significance required for that difference, and the precision expected. Some simple general rules are indicated in the next section.

1.4 Number of treatments and replicates in relation to degrees of freedom

For a useful statistical analysis to be made, the number of residual degrees of freedom (df) should be sufficiently large. In a trial with 8 treatments and 4 replicates with a randomized block design, there are 21 residual df. These are calculated as: total df (32−1 = 31) minus treatment df (8−1 = 7) minus blocks df (4−1 = 3), i.e. 31−7−3 = 21. In a trial with 3 treatments and 4 replicates repeated at 4 sites, there are 24 residual df. These are calculated as: total df (48−1 = 47) minus treatment df (3−1 = 2) minus sites df (4−1 = 3) minus interaction treatment by site df ((3−1)*(4−1) = 6) minus replicate df over sites ((4−1) *4 = 12), i.e. 47−2−3−6−12 = 24.

Residual df should be increased by increasing the replication, the treatments or the number of sites. The desired number of residual df depends on the degree of precision (power) required of the trial. Expert statistical advice should be sought if in doubt. In general, experience with trials/trial series on efficacy evaluation has shown that one should not lay out trials/trial series with <12 residual df. If for any relevant reasons it is advisable to use only 3 replicates and 3 treatments, then the trial may be executed on at least 4 sites (resulting in 16 residual df) to get the minimum residual degrees of freedom of 15 required for a useful statistical analysis.

The choice of the experimental design also has an influence on the number of residual df. The fully randomized design gives the maximum number. The randomized block design uses some of these df to allow for the heterogeneity of the environment (such as that along one gradient). The split-plot design uses df to allow for the possible sources of more than one component of variation. The experimental designer should try to leave the maximum number of df to estimate the residual variation, while choosing an optimal design to minimize that variation, by allowing for all the known sources of heterogeneity (see EPPO Standard PP 1/181).

The relationship between the number of replicates and the residual degrees of freedom for differing number of treatments and sites can be extracted from Table 1.

Table 1 Residual degrees of freedom in relation to number of sites, treatments and replicates in a site

Sites	1 Site						4 Sites						6 Sites					
Replicates	3	4	5	6	7	8	3	4	5	6	7	8	3	4	5	6	7	8
Treatments																		
3	4	6	8	10	12	14	16	24	32	40	48	56	24	36	48	60	72	84
4	6	9	12	15	18	21	24	36	48	60	72	84	36	54	72	90	108	126
5	8	12	16	20	24	28	32	48	64	80	96	112	48	72	96	120	144	168
6	10	15	20	25	30	35	40	60	80	100	120	140	60	90	120	150	180	210
7	12	18	24	30	36	42	48	72	96	120	144	168	72	108	144	180	216	252
8	14	21	28	35	42	49	56	84	112	140	168	196	84	126	168	210	252	294

1.5 Experimental units/plots: size, shape, need for borders

The experimental unit is that part of the trial material to which a single treatment is applied and on which observations are made. Sufficient units are necessary for the planned treatments and replications. In practice, trial material is limited and compromises may often be necessary. Examples of experimental units are: an area of crop (plot), a container of one or more plants, a single plant, a part of a plant (e.g. leaf, stem, branch) and a baiting point in a field. The experimental units should be chosen to be representative of the population the trial is testing and to be as uniform as possible. Lack of uniformity can sometimes be mitigated with replicate blocks.

In general, plots should be rectangular and of the same size in one trial and of similar size for a single trial series. Accuracy increases with plot size, but only up to a certain limit, for variability in soil and infestation conditions also tends to increase. Long thin rectangular plots are suitable for mechanical harvesting. Nearly square plots reduce the risk of interference between plots. For observations of spatially aggregated pests, such as some weeds and soil-borne diseases, a greater number of smaller plots are better than fewer larger plots.

Plot size is given in specific EPPO standards for particular crop/pest combinations. In cases where interference between plots is liable to occur, the plots will be larger (gross plot) and the observations will be limited to the central area (net plot). The difference between the net plot and the gross plot is called a discard area. In general, the EPPO standards suggest net plot sizes, and the gross plot size is usually left to the experimental designer, who should determine the discard areas necessary by considering all the potential sources of interference between plots in each trial or trial series. One common source of interference is spread of the product (for example spray or vapour drift, or lateral movement on/in soil) outside the plot to contaminate adjacent plots. This can be particularly important for sprays applied to tall crops. However, with greater discard areas, the experimental error can often be minimized.

Another common source of interference is spread of the pest (for example air-borne fungi or highly mobile insects) from untreated plots or from plots where control of the pest is poorer. Such spread can both increase the pest population in plots with more efficacious treatments and decrease it in plots with less good ones. Similarly, if a product is being tested in a crop where integrated control is practised, adverse effects on predators and parasites may be masked by their migration between plots.

A further source of interference is competition for light and nutrients. This is particularly relevant if yield is to be measured. If guard areas between plots are different from the plots themselves (e.g. bare paths, a different crop), caution should be exercised when selecting the area for assessment.

According to the application or harvesting equipment used, net plot size may need to be increased above that needed for observations.

Plots may be laid out across or along the direction of work (sowing or planting). The crosswise layout (Fig. 8) has the advantage that, if some mistake is made in the work (cultivation, sowing, etc.), all plots in a block will probably be equally affected. However, then treatment and harvesting become more difficult. The lengthwise layout offers practical advantages for treatment and harvesting, but runs the risk of greater heterogeneity along very long blocks. The hybrid layout may provide a compromise.

1.6 Role and location of untreated controls

1.6.1 Purpose of the untreated control

The main feature of 'untreated controls' is that they have not been subjected to any of the plant protection treatments under study. Untreated controls should, however, receive all the measures which are uniformly applied throughout the trial, in particular cultural measures and applications against pests not being studied. Though the untreated control normally receives no treatment at all against the pest being studied, in certain circumstances it may be useful to modify the untreated control to include certain operations received by the other treatments. For example, where the other treatments receive the products in aqueous solution through the passage of spray machinery over the plot, the untreated control may be modified to include a passage of spray equipment, but with water alone. The idea is to replicate, as far as possible, the operations of the other treatments, with the exception only of the application of the product itself.

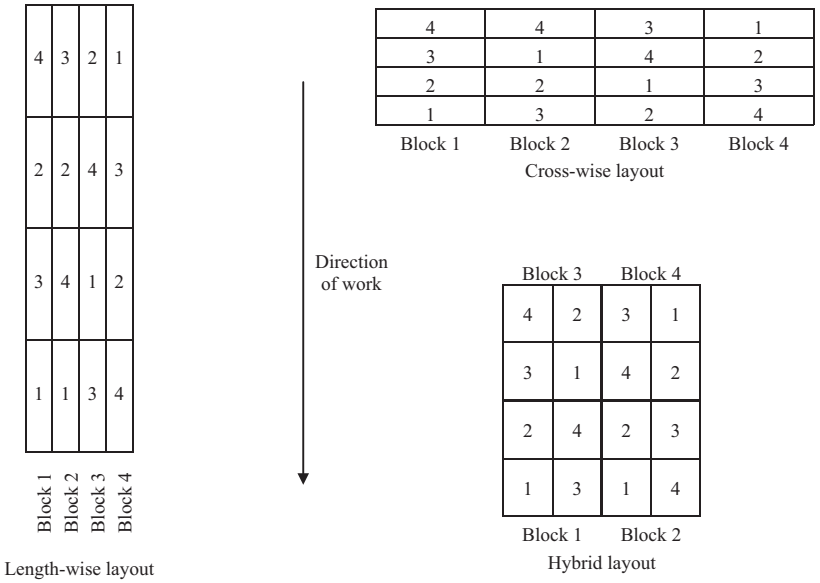

Fig. 8 Similar randomized block designs, but with different layout of plots relative to the direction of work.

The main purpose of the untreated control is to demonstrate the presence of an adequate pest infestation. For example, unless an untreated control has confirmed the presence of an adequate pest infestation, efficacy cannot be demonstrated and results are then not meaningful. This confirmation may be qualitative (presence of dominant species, type of flora, weeds, etc.) or quantitative (compliance with minimum and maximum thresholds, spatial distribution). Under exceptional circumstances, an untreated control may not be possible (e.g. for quarantine pests).

Depending on the objective and the type of experiment, untreated controls play a useful role, and possibly several roles at the same time. Among them are:

- Showing the efficacy of a new product and the reference product. The primary proof of the efficacy of a new or reference product is always obtained from a comparison with the untreated control;
- Assistance in making observations. A visual estimation of damage or infestation may sometimes be done in relative terms, by comparison with a control;
- Use of the technique of the 'adjacent control' to measure and take account of spatial distribution in the plots;
- Observation of the development of the pest (emergence, flight, spore release, etc.), in particular as a basis for determining dates for application or observation;
- Provision of a reserve of inoculum in order to ensure that inoculum level does not fall too far or become too heterogeneous (in extreme cases, this may be practically equivalent to artificial infestation);
- Assistance in interpreting the results of trials. For example, a significant difference between 2 treatments may not have the same importance depending on the level of infestation.
- Making the results of the analysis more accessible for users by expressing them in a different form, or by allowing their graphical representation (e.g. transformation of mortality into efficacy rate);
- Allowing for additional observations, in particular quantitative or qualitative yield, which it may be interesting to link with the other results of the trial.
- Finally, and exceptionally, formation of a comparison term for the treatments under study if no reference product is available. This may occur, for example, when the type of product, or its use, is new or when all potential reference products have been withdrawn from use. This role is then similar to the role of the reference product, although its interpretation is very different. Controls may then be compared with the various treatments using formal statistical significance tests, in the same way as the reference product is compared with them in usual trials.

1.6.2 Types of arrangements of untreated controls

Four types of arrangements of the control are possible.

Included controls: the controls are considered like any other treatment, the control plots are the same shape and size as the other plots, and the controls are randomized in the trial. The included control is the most usual way to carry out trials and all other versions are used exceptionally (mostly in herbicide testing).

Imbricated controls: the control plots are arranged systematically in the trial. Plot size and shape need not be the same as for other plots in the trial. The observations made in these plots are of a different nature and should not be included in the statistical analysis. The purpose of the arrangement is to ensure a more homogeneous distribution of the effect of an adjacent untreated area than is possible

with the included randomized design. Various arrangements are possible; the plots may be placed between blocks or between treated plots within blocks (Fig. 9).

Excluded controls: control plots are selected outside the trial area and not adjacent to it, in an area with conditions closely similar to those of the trial. Replication is not essential but may be useful if the area is not homogeneous. The observations made in these plots should not be included in the statistical analysis.

Adjacent controls: each plot is divided equally into 2 subplots, and one of these (at random) is left untreated. Observations are made in the same way in both sub-plots. The observations made in these plots should not be included in the statistical analysis unless due allowance is taken of the fact that the design is a form of split-plot. In a split-plot design, the variability within plots may differ from that between plots; consequently the analysis of variance should include 2 strata of error. Specialist statistical advice may be necessary to interpret the results.

1.6.3 Choice of the type of arrangement of untreated control

The choice of the arrangement of the untreated control depends on its role(s) in the trial. Although the included control has very often been used in the past in efficacy evaluation trials, and is still frequently used in practice, it is not necessarily the most suitable. The following decision scheme gives guidance.

(a) If the control is used in a statistical test, then the 'included control' is essential. If not, another type of control can be used. In either case the heterogeneity of the plots should be considered;
(b) If heterogeneity is high, the 'adjacent control' is suitable. If heterogeneity is low or moderate, the interference of the control plots with the adjacent plots should then be considered;
(c) If the control plots are not liable to interfere with adjacent plots, then the 'imbricated control' is suitable;
(d) If control plots are liable to interfere with adjacent plots, then the 'excluded control' should be used.

1.7 Selection of the sample size in a plot

The main purpose of taking several samples inside a plot is to reduce the variability of the estimated plot mean to a suitable level for the assessed variable. The sample size should be large enough to achieve this purpose. The sample size required depends greatly on the nature of the observation and the variability within the plot. EPPO standards on the assessment of specific pests, weeds and diseases give advice on sample sizes. In practice, sample sizes of 10–50 elements are usually enough to accomplish the goal of correct estimation of the mean value in a plot, depending on the inherent variability. Note that, if the treatments are applied to plots, then increasing the sample size within plots only gives a strictly limited return of efficiency, because between-treatment comparisons should be made at between-plot scale.

Sampling should always be random and should adequately cover the area of the plot and the experimental

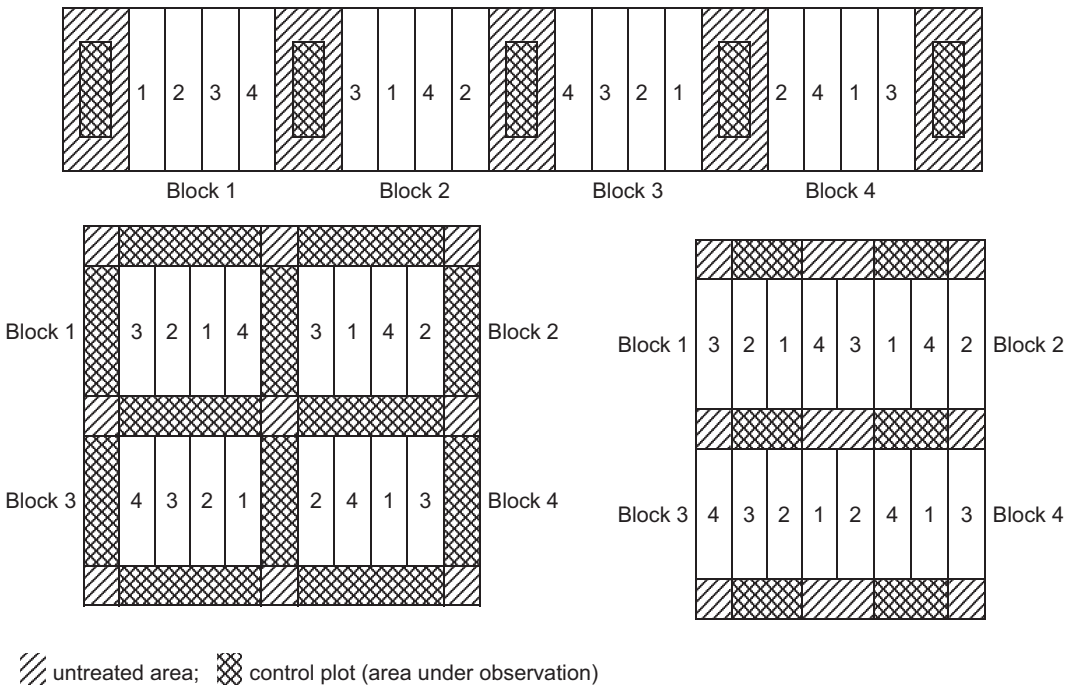

Fig. 9 An example of the use of imbricated control plots for a randomized block trial with 4 blocks and 4 treatments.

material. For practical reasons, sub-sampling may be necessary. A review of sampling methods and references to the literature may be found in Perry (1994).

2. Principles for assessing the effects of plant protection products

In the assessment of the effect of products in an efficacy evaluation trial, 'variables' are assessed by 'modes of observation'.

2.1 Variables

The nature of the variable is important as it generally influences the statistical method chosen for interpreting the results. Several categories of variables can be distinguished.

Binary variables (e.g. yes/no, presence/absence): these may lead to binomial variables, which represent the number of times a state was observed, given a known number of observations (e.g. the number of plants infested, out of 20 randomly selected within a plot).

Nominal variables: variables of equivalent importance which cannot be put into any particular order (non-ordinal), e.g. species of pests, different colours. Values of these variables are generally indicated by words.

Ordinal variables: variables with values that are classes that form a particular order, but which are not measured. They are usually qualitative, and then the classes can only be placed relative to each other (e.g. bad, medium, good; descriptive grades of leaf damage). In other cases, values may be expressed numerically (and could be measured exactly but are not for practical reasons), e.g. categories of weed cover, or categories of aphid infestation.

Quantitative variables: variables which can be measured and are measured in practice, e.g. yield, plant height, number of larvae, percentage of leaf area infected by disease. They may be discrete, if they are expressed only as integers (e.g. results of counts), or they may be continuous (e.g. weights, sizes). Quantitative variables may also result from appropriate mathematical operations. Differences or sums can be obtained (e.g. the difference between a value before and after treatment). Relative values can be calculated, which may be a proportion or a ratio. A proportion represents the quotient 'partial quantity/total quantity' and lies between 0 and 1 (i.e. a limited continuous quantitative variable). In practice, such a proportion is often a relative frequency: 'frequency in one class/total frequency', i.e. is a binomial or multinomial variable. It is often expressed as % (i.e. a value between 0 and 100). Ratios have no upper limit, e.g. (original value−final value)/original value. They may also be expressed as % (in which case values over 100 are possible). A value expressed as a percentage may in fact be a proportion or a ratio, or even a binomial variable, and it is important for statistical purposes to distinguish between these cases.

2.2 Mode of observation of variables

It is imperative to measure variables as accurately as is feasible.

In practice, the following precautions should be taken when assessing a trial:

(a) Establish the scale, key or measurement method beforehand. The selected method should be used in all trials of a trial series;
(b) Do assessments without foreknowledge of the treatment plan;
(c) Work block by block;
(d) Use the same method in all experimental units, e.g. observe all plots from the same direction to avoid differences in lighting.

In order to assess variables, 4 modes of observation are possible: measurement, visual estimation, ranking, and scoring.

2.2.1 Measurement

A measurement determines values objectively. The results of measurement may be a continuous variable (weight, size) or a discrete one (counts). In field experiments, when measurement does not relate to the whole plot, it should be done on a sample whose size and sampling mode were determined beforehand according to the precision required.

2.2.2 Visual estimation

Visual estimation determines values subjectively, but on the same scales and with the same range of values as the measurements referred to above. It usually relates to continuous variables (e.g. weed cover) but may also relate to counts when these are large (numbers of lesions on a leaf). In general, an estimation is easier to make by comparison with a reference or an untreated control than it is in absolute values. It should be stressed that the same kind of variable is obtained as with measurements. The results should thus be expressed in the same units. The values obtained or the transformed values, if necessary, can be considered as real estimations of a continuous variable and can be analysed with normal statistical procedures.

The observer should be trained to make the estimations and his observations should be calibrated against a standard. If these conditions are met, estimates may be very accurate. Accuracy may, however, vary with the level of the variable under estimation. For example, if % leaf area affected is being estimated, high and low values are estimated more accurately than intermediate values. This fact has been taken into account in the design of various aids to estimation and scales (see below). It may also necessitate statistical transformation of estimated data.

2.2.3 Ranking

Ranking situates each individual in a position relative to other individuals. The result of ranking is a qualitative ordinal variable. Provided the number of replicates to be compared is relatively small, they can be ranked for an

appropriate variable in the field. In case there is no alternative to ranked observations, non-parametric methods like the analysis of variance of the ranks may be useful statistical procedures. However, the power of such a non-parametric method is usually smaller when compared to parametric procedures. Hence ranking is not an ideal approach and should be avoided particularly if numbers of replicates are small.

2.2.4 Scoring

To score is to assign an object to an unambiguously defined class. The set of such classes is commonly called a scale, especially when, as is usual, the underlying variable is ordinal. Scoring is also used for binary and nominal variables. Scoring is by definition subjective. It may be applied to a wide range of objects: ordinated or not, continuous or discrete. It may be useful for qualitative variables as well as for quantitative variables that can only be measured accurately at great expense. Its main advantage is that it is usually quick and non-destructive, and can be used to characterize a whole plot by a single value. The number of steps on a scale is its sensitivity. This should not be too low, otherwise the results provide little useful information, or too high, otherwise the scale becomes impractical to use.

Scales are adapted to specific purposes and there is not, in general, a single universal scale for a certain type of variable. The EPPO standards give numerous examples of scales whose use is recommended (Appendix 1) for the assessment of particular crop/pest combinations. In general, certain simple rules have been followed for these scales, especially in defining the extremities. The lowest point of an ordinal scale (no effect) should be given as 1 (not 0 – reserved in many recording systems for missing observations) and the highest value on the scale should correspond to the greatest value of the effect, with intermediate steps in order.

Table 2 sums up the different modes of observation and the various types of variables obtained.

2.3 Use of scales in visual estimation and in scoring

Often there is some confusion between visual estimation and scoring. The operations are superficially similar, but their result is different: visual estimation gives rise to a series of estimated values of a discrete or continuous quantitative variable; scoring gives numbers in classes. The classes of a scoring scale are often represented by sequential numbers (e.g. 1–9), but that does not mean that the intervals between the scale values are the same. If intervals differ, it is not advisable to attempt statistical analysis without specialist advice, or to estimate statistical parameters without thought. Any statistics derived by such computation should be interpreted with great caution. Scale values could as well be represented as letters of the alphabet, which emphasizes at once their nature as an ordinal variable and the dangers of too simplistic an approach.

Scales can, however, also be used as aids to visual estimation ('ordinal variable with intervals'). If the scale values are actual values of a quantitative variable (as in a visual key of % leaf area affected), then the observer assigns the values on the scale, or interpolates intermediate values according to his judgement. The values obtained, suitably transformed if necessary, are estimates of a continuous variable and may be analysed accordingly, using the usual statistical procedures. It is important to stress that, if the observer has the resources (time, manpower, experience) to make even more precise estimates, or even measurements, the resulting data can be analysed with even greater accuracy and power. There is, however, no purpose in making a relatively accurate estimate (for example, of % leaf area affected), and then substituting a much less accurate scale value. The advantages of scoring (speed and simplicity) only exist if the observer scores directly into the appropriate score class (which they have been trained to do) without making any attempt to estimate more accurately.

Table 2 Different modes of observation and types of variables

	Measurement	Visual estimation	Ranking	Scoring
Binary				X
Nominal				X
Ordinal			X	X
Discrete	X	X		
Continuous limited	X	X		
Continuous not limited	X	X		

2.4 Quality of a mode of observation

Modes of observation can be distinguished by a number of qualities:

- 'Accuracy' – absence of bias, in the statistical context;
- 'Reliability' – low variability (or variance);
- 'Precision' – the combination of accuracy and reliability;
- 'Sensitivity' reaction of the mode of observation to a small change in the value of the experimental unit;
- 'Repeatability' – the same (or very close) value given by the same observer to identical experimental units;
- 'Reproducibility' – the same (or very close) value given by different observers to the same experimental unit.

These important qualities will determine the choice of modes of observation for particular purposes, especially in trial series.

3. Statistical analysis of trial results

A decision on the need to conduct a statistical analysis of the results of a trial or a trial series will depend on the results demonstrated and on the purpose of the trial. Statistical analysis is not essential in all trials used for registration

purposes. Statistical analysis is particularly valuable, for example, when comparing effects of treatments at different doses, effectiveness of different formulations of the same product, or effect on yield relative to another treatment.

3.1 Principles

What follows is intended to give an outline of good statistical practice in the analysis of data. It is not, and cannot be, a prescription for all analyses, or cover all situations.

Practitioners should never underestimate the need for professional statistical advice. It is important for practitioners to understand the advice they receive and it is often better for them to perform a simple analysis that they can report and defend with confidence than to accept advice that leads to an analysis that they may understand only partially. The bibliography at the end of these standards may be helpful. It gives several good texts that attempt to reveal the principles of good statistical practice, rather than to provide a series of statistical recipes to be followed blindly.

3.2 Statistical analysis of a single trial

3.2.1 Basic structure and sequence of analysis

EPPO Standards for the efficacy evaluation of plant protection products state that 'Statistical analysis should normally be used, by appropriate methods which should be indicated'. The procedure to be followed can be illustrated by a typical trial in which several test products are applied at a single dose and compared with a reference product, in the presence of an untreated control. Product efficacy is assessed by a measured quantitative variable. The purpose of the trial is to compare the test products with the reference product, and in particular to identify which are the most effective. The sequence of analysis, for a trial properly conducted according to the appropriate EPPO Standard, is the following:

(a) Is the trial realistic, i.e. able to give useful data? This will only be so if pest infestation in the untreated control is sufficiently high and not too variable.
(b) Are the results coherent? Does the reference product give the expected result in comparison with the untreated control?
(c) If these 2 conditions are satisfied, it is then valid to compare the test products with the reference product and possibly to make comparisons between the products themselves. The analysis should aim primarily to estimate the magnitude of the differences or ratios between the test product and the reference product and to provide an estimate of the variability of those estimates by means of standard error, confidence intervals or similar statistic.

Similar schemes can be developed for other efficacy evaluation trials, and in particular for the special case of herbicide selectivity, and for exceptional cases in which there are no suitable reference product and the treatments need to be compared with the untreated control (see section 1.6.1).

If 2 (or more) reference products are included (see example under 1.2.2), the mode of analysis should be defined beforehand. A separate comparison of each reference product to the new product without any adjustments or corrections is recommended. If a comparison of the test product against the combined references is deemed appropriate, then a homogeneity test between the references may be carried out first.

3.2.2 Choice of the method of analysis

Broadly, the type of variable determines the method of analysis. If the variable is quantitative (binary, binomial, discrete or continuous), a parametric statistical method should be used, usually based on Generalized Linear Models (GLM), e.g. analysis of variance, linear regression, logistic regression. If the variable is qualitative, non-parametric methods are appropriate. In performing an analysis of variance, 3 assumptions are made: additivity of effects, homogeneity of variance and normality of the error. The use of non-parametric methods is recommended if these assumptions are not satisfied. However, non-additivity and non-normality can very often be improved, and are not a sufficient reason to analyse data by non-parametric methods, which generally lack power.

3.2.3 Additivity of effects

It is important to consider whether effects are likely to be additive on the scale on which it is proposed to analyse the response variable. For example, if the response variable is an insect population density, then effects of treatments such as insecticide or fungal attack are likely to be multiplicative, affecting a proportion of the population. Alternatively, if the response variable is the proportion of weeds killed by a herbicide treatment, then effects are likely to be additive not on the natural scale, but on a probit or logit scale. Two common methods are used to amend the natural scale to one that is more realistic: transformations and generalized linear models. Generalized linear models are a form of regression that generalizes the analysis of variance for designed experiments. They are an improvement on transformations in that they address separately and simultaneously the problem of additivity of effects and equality of variances (non-normality). They allow the distribution of the response variable to be specified directly. For example, for insect counts, such a model might specify a logarithmic 'link function' (to address multiplicative effects), and a Poisson distribution for the counts (to address directly the problem of equality of variances and non-normality). Alternatively, a binomial variable may be analysed by using a logit link function (to achieve additivity) and by specifying a binomial distribution (to directly match the data, which may be in the form of r diseased plants out of n treated). There are many parallels between the analysis of deviance that results from the use of a generalized linear model, and

the traditional analysis of variance. In particular, the concepts of sums of squares, degrees of freedom, orthogonal contrasts, chi-squared and *F*-tests, and predicted means with standard errors all have specific parallels in generalized linear models, and may be provided for examination.

3.2.4 Homogeneity of variance

Transformations may address the problem of additivity of effects, but they do not usually ensure homogeneity of variances. This should be checked for independently, although additivity is usually the more important. For counts, a logarithmic transformation will often ensure both additivity and equality of variances. For binary data, binomial data or data in the form of proportions, a logit, probit or complementary log-log transformation will usually ensure additivity, although this may not give equality of variances.

3.2.5 Normality and independence of the error

The distribution of the errors should be normal. Standard tests or graphical displays are available to check this. In practice, the analysis of variance is often robust to departures from normality. It should be confirmed where possible that the errors are independent of treatment factors.

3.3 Analysis of variance

3.3.1 Tables of means

Either a LM (Linear Model), GLM or following transformation, an analysis of variance is recommended. A table of the mean of each of the treatments should then be presented, with an estimate of the variability of the means, usually in the form of a standard error or confidence interval. Such a table places emphasis on the magnitude of effects, and is recommended to overcome the well-known problem that biological relevance cannot be equated with statistical significance, and that effects may be large in magnitude and importance but non-significant due to insufficient power of the analysis or test. Analysis may also employ a generalized linear model, of which the analysis of variance is a special case, or some other appropriate method.

Care should be taken to assign the units to strata properly in the analysis of variance table, with the treatment and blocking structure appropriate to the design adopted. In particular, care should be taken to guard against the well-known problem of pseudo-replication, caused by not allowing for the fact that treatments have not been randomized fully over the sample units, but applied instead to groups of units.

3.3.2 F-tests and orthogonal contrasts

In addition to the presentation of tables of means and standard errors, formal statistical tests, usually *F*-tests, may be performed on the data as a whole. An overall test of all the treatments should not be presented as evidence of efficacy, except in the simplest of cases, since this will in general be contaminated by the information from the untreated control and treatments with poor efficacy. Instead, it may be useful that the treatments sums of squares is divided into components of biological interest by the definition of orthogonal (independent) contrasts.

For example, in the first example, where 8 treatments were compared, there were 5 different test products, 2 reference products and an untreated control. The 8 treatments yield 7 df in the treatments sums of squares. Sensible contrasts to make might be: the untreated control versus the mean of the other 7 treatments (1 df); reference product one versus reference product two (1 df); the mean of the reference products versus the mean of the test products (1 df); differences between the means of test products themselves (4 df). Of these contrasts, the first two are designed to remove nuisance variation of relatively little biological interest, and it is the last two that help to reveal the true objectives of the trial. Each contrast provides a separate *F*-statistic, which may be used to test formally hypotheses of concern. In this example, the important hypotheses might be that, on average, the test products were no better than the reference products, and that there were no differences between the test products themselves. Interpretation of the first of these hypotheses might well be influenced by whether the contrast between the reference products themselves had exposed a large difference. If non-orthogonal contrasts are to be tested, for example the separate 5 contrasts on 1 df between the mean of each test product against a specific reference product, then these should also be done by *F*-test (or *t*-test if appropriate) using the residual mean square from the analysis of variance.

Contrasts and hypotheses of interest should preferably be specified in advance, at the design stage, and used sparingly. Tests should not be done merely because a *post-hoc* preliminary analysis showed differences that appeared large and might be significant if tested. Consistency is often a better guide to the presence of a real effect than isolated significance tests, particularly if the power of the test is low. For example, if a test product proved more efficacious than a reference product at each of 11 distant sites, but not significantly so in any of them, common sense would argue that the consistency of the results were important (indeed, a 2-tailed binomial test could be used to argue that the probability of a result as extreme as this, if there were no real difference between the treatments, was <0.001).

3.3.3 Multiple test procedures

For registration purposes not all pairwise comparisons are of relevance and not all orthogonal contrasts can be considered in a registration application. Of all the possible $(k(k-1))/2$ pair wise comparisons only a few are relevant to demonstrate the efficacy of a test product. For example, let us consider a trial where 7 treatments were compared, with 5 different test products, one untreated control and one reference product. According to the rules described under

3.2.1, several relevant tests should be done. First, the relevance of the trial should be demonstrated by testing the level of infection in the untreated control against a predefined infection level. Second, testing the difference between the reference product and the untreated control should be done to demonstrate the coherence of the trial. If this is achieved, then the third procedure is to compare each test product against the reference product to attempt to demonstrate at least an equality of effect in relation to the reference product. To perform this last test there are many different accepted parametric and non-parametric procedures that are available in the literature (Hothorn & Bleiholder, 2006).

In a factorial experiment (e.g. multiple doses test), it is not usually helpful to perform all pair wise comparisons among factorial combinations (Perry, 1986). Instead, it is more appropriate to analyse the data according to the treatment structure. Depending on the outcome of a 2-way analysis of variance, it is usually most appropriate to compare marginal means or simple means for a factor with separate levels of the other factor, and vice versa.

The 'standard' multiple comparison procedures according to Tukey (1953) or the widely used Duncan's test (Duncan, 1955) or Newman–Keuls test (Keuls, 1952) perform all-pairs comparisons, which are inherently 2-sided. Much less conservative procedures with adequate comparisons are possible when formulated as one-sided tests. One-sided tests and confidence intervals are biologically appropriate because, for example, the interest is usually in the reduction of an infection, not in its increase. The widely used Duncan's multiple range test and the Newman–Keuls multiple range test do not control the global size of the test (α level), controlling only the local size (α level). Hence, if the test is based on a predefined α level of 0.05, this will only be true when comparing 2 treatment means; with an increasing number of means compared simultaneously the α level increases exponentially. When using multiple test procedures it is recommended to select only those procedures that are known to control the local and the global α level simultaneously.

Since registration field trials to demonstrate efficacy of a new test product will be performed at the final stage of product development, the expected direction of each difference should be clear from the context. Therefore, one-sided tests and one-sided confidence limits are recommended to be used to guarantee some level of power, for the normal number of replicates typically used in field trials. However, it does not exclude the use of other statistical tests mentioned above.

3.3.4 Random effects models

This standard has focussed on regarding treatments as fixed effects. Some practitioners, in some experiments, particularly uniformity trials, may wish to regard the treatment effects as some random sample from a larger unknown population. This is known as random effects modelling. Trials can also contain fixed and random effects, so called mixed models. For such models, the technique of REML (Residual Estimation by Maximum Likelihood) can be recommended. REML may also be used when it is desired to make comparisons between several laboratories or sites, to estimate variance components, or when a design cannot be analysed by analysis of variance because there are so many missing values that it has become unbalanced. Once again, there are certain similarities between REML concepts and quantities and those of analysis of variance. However, expert statistical advice should usually be sought.

3.3.5 Ordinal data

Modern methods for analysis of ordered categorical data have been described by Agresti (1984) and Brunner & Munzel (2002), although some statistical advice may be required to use them wisely. In addition, it may be necessary in some cases to treat integer variables as ordinal variables, if their range of variation is not great enough for them to be considered as continuous and if the trial is nevertheless considered valid.

3.3.6 Qualitative data and non-parametric methods

For data that is truly qualitative, for example nominal data, and for certain ranked data, or for data that does not follow a well-known parametric distribution such as the normal, binomial, beta, gamma or Poisson, non-parametric methods could be an useful statistical procedure for data analysis. The power of non-parametric methods when compared to parametric methods are smaller, so they should be used with extreme caution if numbers of replicates are very small. However, the amount of information that such an analysis can impart is for the purpose of this guideline high enough to get useful results when testing the efficacy of a product. Descriptions of the traditional tests can still hardly be improved over those in the text by Siegel (1956) and Brunner & Munzel (2002), which explain clearly which test is appropriate for which set of data. More modern approaches may involve computer-intensive techniques such as randomization tests. Randomization methods can be very useful where parametric approaches should be distrusted, for example if data is very non-normal, or if there is a large proportion of zeroes in the data (if the trial is nevertheless considered valid). Other computer-intensive non-parametric methods are to be recommended to improve estimation, or to calculate better the variability of an estimate. These include 'bootstrapping' and 'jack-knifing', but again may require specialist advice.

3.4 Statistical analysis of trial series

The consistency of treatment effects, e.g. of the comparison of the new product versus the reference product, for different environments (regions, sites) is a relevant and important criterion for registration. Therefore, trial series are preferred to single trials.

3.4.1 Definition

For the purposes of this standard, a trial series can be defined as a set of treatments tested under different environmental conditions in one or several years. The set of treatments belonging to a series should be analysed together using the same statistical model.

3.4.2 Planning

When planning a trial series experimenters should consider defining the trial question and all relevant parameters, i.e. core treatment list, trial design and replicates, number of sites, test methods, etc., that are required to apply the biometrical model planned to be used for the trials series analysis.

3.4.3 Goals

The objectives of analysis are:

- To estimate treatment effects over sites and years;
- To test the interactions between treatments, sites and years. Environmental and other differences between sites and years may confound these factors;
- Possibly also to test the significance of differences between treatments and standards.

3.4.4 Basic structure and sequence of analysis

Before beginning statistical analysis of the results of a trial series, the data of each trial should be validated. This validation applies to 3 items:

- *Methodological validation*: the conduct of all trials should comply with the original protocol;
- *Agronomic and biological validation*: trials should not be disturbed by external or exceptional factors. They should be representative for the region and the year. The reference products in all trials should perform normally. The infection pressure should be suitable (significant level for efficacy trials, weak level for selectivity trials);
- *Statistical validation*: trials should be accurate, showing a typical standard error (or coefficient of variation).

The analysis of trial series is directed to efficacy as well as to the treatment-by-environment or treatment-by-factor interactions. The objective of the analysis of interactions is to demonstrate no meaningful interactions for all or almost all environments and factors, and thus identify any situations where efficacy might be substantially impaired (further information on zonal submissions and evaluations is available in PP 1/278 *Principles of zonal data production and evaluation*). This cannot be demonstrated appropriately merely by the presence of a non-significant global *F*-test of the interaction term. Instead, it is more appropriate to examine all interaction components by means of contrasts, in order to demonstrate the similarity of the treatment effects over all, or at least the majority of environments. Qualitative interactions may be excluded thereby; quantitative interactions should be only tolerated up to a practical acceptable amount. Sites showing no treatment by environment interactions might then be pooled for analysis. Sites showing a level of interaction that is unacceptably large should be analyzed and discussed separately.

3.4.5 Choice of statistical method

As for an individual trial, the statistical methods are determined by the type of variable to be analysed. The methods to be used are the same or similar to those used for an individual trial (e.g. analysis of variance, non-parametric methods). The main purpose of the analysis of a trial series is to measure and test the interactions between the test products and the environment or site, i.e. showing that the differences between products are 'equal' on every site. The trials can be grouped in advance of the analysis, according to appropriate criteria (e.g. soil type, level of infestation), or afterwards, using the analytical methods and results of the interactions to group the trials accordingly.

Acknowledgements

EPPO acknowledges with thanks the detailed recommendations of H. Bleiholder and L.A. Hothorn for the revision of this Standard in 2006.

References

Agresti A (1984) *Analysis of Ordinal Categorical Data*. Wiley, New York (US).

Bauer P, Röhmel J, Maurer W & Hothorn LA (1998) Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* **17**, 2133–2146.

BBA (1980) *Richtlinie für Planung, Durchführung und Auswertung von Versuchen mit Pflanzenbehandlungmitteln: 1. Versuchsplanung; 2. Versuchsdurchführung*. Biologische Bundesanstalt für Land- und Forstwirtschaft, Braunschweig (DE).

BBA (1982) *Richtlinie für Planung, Durchführung und Auswertung von Versuchen mit Pflanzenbehandlungmitteln: 3. Auswertung des Einzelversuches; 4. Sachregister*. Tabellen, Biologische Bundesanstalt für Land- und Forstwirtschaft, Braunschweig (DE).

Brunner E & Munzel U (2002) *Nichtparametrische Datenanalyse*. Springer, Berlin (DE).

CEB (1981) Rôle et implantation des témoins sans traitement dans les essais de produits phytosanitaires. ANPP-DT4. J Arnoux

CEB (1983) Principes d'appréciation des effets des produits phytosanitaires dans les essais de plein-champ. ANPP-DT5. Y Ribrioux

CEB (1986) Utilisation des tests statistiques dans l'interprétation des essais de produits phytosanitaires. ANPP-DT6. J Arnoux - JP Gouet

CEB (1990a) Les réseaux d'essais. ANPP-DT9. JP Gouet

CEB (1990b) Les unités expérimentales. ANPP-DT10. JP Gouet

Cochran WG & Cox GM (1957) *Experimental Design*, 2nd edn. Wiley, New York (US).

Cox DR (1958) *Planning of Experiments*. Wiley, New York (US).

Crawley MJ (1993) *GLIM for Ecologists*. Blackwell Scientific, Oxford (GB).

Crowder MJ & Hand DJ (1990) *Analysis of Repeated Measures*. Chapman & Hall, London (GB).

Cullis BR & Gleeson AC (1991) Spatial analysis of field experiments – an extension to two dimensions. *Biometrics* **47**, 1449–1460.

Dagnelie P (1969) *Théorie et Méthodes Statistiques*, Vol. 2. Duculot, Gembloux (BE).

Denis JB & Vincourt P (1982) Panorama des méthodes statistiques pour l'étude de l'interaction génotype x milieu. *Agronomie* **2**, 219–230.

Denis JB (1980) Analyse de régression factorielle. *Biométrie-Praximétrie* **19**, 15–34.

Denis JB, Gouet JP & Tranchefort J (1980) Méthodes d'étude de la structure de l'interaction génotype x *milieu et de recherche d'un modèle explicatif à effets fixes: application à l'analyse des résultats d'un réseau d'essais de variété de blé tendre. In: *Biométrie et Génétique* (Eds. J-M Legay, JP Masson, R Tomassone), pp. 98–109. Société Française de Biométrie, Paris (FR).

Dobson AJ (2002) *An Introduction to Generalized Linear Models*, 2nd edn, Chapman & Hall, CRC/Boca Raton (US).

Duncan DB (1955) Multiple range and multiple F tests. *Biometrics* **11**, 1–42.

Dyke GV (1988) *Comparative Experiments with Field Crops*. Griffin, London (GB).

Finney DJ (1971) *Probit Analysis*, 3rd edn. Cambridge University Press, Cambridge (GB).

Finney DJ (1978) *Statistical Method in Biological Assay*, 3rd edn. Griffin, London (GB).

Finney DJ (1980) *Statistics for Biologists*. Chapman & Hall, London (GB).

Gouet JP & Philippeau G (1992) *Comment Interpréter les Résultats d'une Analyse de Variance?* ITCF, Paris (FR).

Gouet JP (1974) *Les Comparaisons de Moyennes et de Variances. Application à l'Agronomie*. ITCF, Paris (FR).

Hollander M & Wolfe DA (1973) *Non-parametric Statistical Methods*. Wiley, London (GB).

Horn M & Vollandt R (1995) *Multiple Tests und Auswahlverfahren*. Gustav Fischer Verlag, Stuttgart (DE).

Hothorn LA & Bleiholder H (2006) Statistical aspects of efficacy evaluation of plant protection products in field trials – a comment to the EPPO PP1/152(2) guideline. *Bulletin OEPP/EPPO Bulletin* **36**, 31–45.

Hughes G & Madden LV (1992) Aggregation and incidence of disease. *Plant Pathology* **41**, 657–660.

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**, 187–211.

Keuls M (1952) The use of studentized range in connection with an analysis of variance. *Euphytica* **1**, 112–122.

Little TM (1978) If Galileo published in HortScience. *HortScience* **13**, 504–506.

McCullagh P & Nelder JA (1983) *Generalized Linear Models*. Chapman & Hall, London (GB).

Mead R & Curnow RN (1983) *Statistical Methods in Agriculture and Experimental Biology*. Chapman & Hall, London (GB).

Mead R (1988) *The Design of Experiments: Statistical Principles for Practical Applications*. Cambridge University Press, Cambridge (GB).

Nelder JA (1971) Contribution to the discussion of the paper by O'Neill and Wetherill. *Journal of the Royal Statistical Society Series B* **36**, 218–250.

Parker SR, Whelan MJ & Royle DJ (1995) Reliable measurement of disease severity. *Aspects of Applied Biology* **43**, *Field experiment techniques*, 205–214.

Patterson HD & Williams ER (1976) A new class of resolvable incomplete block designs. *Biometrika* **63**, 83–92.

Pearce SC, Clarke GM, Dyke GV & Kempson RE (1988) *Manual of Crop Experimentation*. Griffin, London (GB).

Perry JN (1986) Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology* **79**, 1149–1155.

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

Perry JN (1994) Sampling and applied statistics for pests and diseases. *Aspects of Applied Biology* **37**, 1–14.

Perry JN (1997) Statistical aspects of field experiments. In: *Methods in Ecological and Agricultural Entomology* (Ed. Dent DR & Walton MP), pp. 171–201. CAB International, Wallingford (GB).

Plackett RL (1981) *The Analysis of Categorical Data*, 2nd edn. Griffin, London (GB).

Preece DA (1982) The design and analysis of experiments: what has gone wrong? *Utilitas Mathematica* **21A**, 201–244.

Rasch D, Herrendörfer G, Bock J, Victor N & Guiard V (1996) *Verfahrensbibliothek, Versuchsplanung und -auswertung*. Band I. R. Oldenbourg Verlag, München (DE).

Rasch D, Herrendörfer G, Bock J, Victor N & Guiard V (1998) *Verfahrensbibliothek, Versuchsplanung und -auswertung*. Band II. R. Oldenbourg Verlag, München (DE).

Siegel S (1956) *Non-parametric Statistics for the Behavioral Sciences*. McGraw Hill, New York (US).

Tukey JW (1953) *The Problem of Multiple Comparisons*. Mimeographed monograph. Princeton University, Princeton NY (US).

Appendix 1

Examples of scales used in EPPO Standards for the efficacy evaluation of plant protection products

1. Nominal scale

Leaf discolouration in potato (EPPO Standard PP 1/135 *Phytotoxicity assessment*):

(1) chlorosis;
(2) yellow veins;
(3) yellow spots;
(4) whitening.

2. Ordinal scales without quantitatively defined intervals.

Assessment of cabbage roots for *Plasmodiophora brassicae* (PLADBR) (EPPO Standard PP 1/39 *Efficacy evaluation of fungicides against* Plasmodiophora brassicae):

(1) no swelling visible;
(2) very slight swelling, usually confined to lateral roots;
(3) moderate swelling on lateral and/or tap roots;
(4) severe swelling on lateral and/or tap roots.

Assessment of lettuce plants for infection by *Botryotinia fuckeliana* (BOTRCI) (EPPO Standard PP 1/54 *Efficacy evaluation of fungicides against* Botrytis *spp. and* Sclerotinia *spp. on vegetables*):

(1) no attack;
(2) slight attack, infection of basal petioles only;
(3) moderate attack, stem lesion not girdling stems;
(4) heavy attack, stem lesion girdling stems, or upper leaves infected, lettuce unmarketable (including plants completely destroyed by *B. fuckeliana* during the trial).

3. Ordinal scales with defined intervals based on numbers.

Attack by *Venturia inaequalis* (VENTIN) on apple fruits (EPPO Standard PP 1/5 *Efficacy evaluation of fungicides against* Venturia inaequalis *and* V. pirina):
(1) no attack;
(2) 1–3 spots per fruit;
(3) >3 spots per fruit.

Number of lesions on sugarbeet due to e.g. *Scutigerella immaculate* (SCUTIM) (EPPO Standard PP 1/45 *Efficacy evaluation of insecticides against the soil pest complex in beet*):
(1) no lesions;
(2) 1–2 lesions;
(3) 3–5 lesions;
(4) >5 lesions.

Some scales are partly based on number, partly on area, e.g.
(1) healthy leaf;
(2) 1–2 spots per leaf;
(3) more than 2 spots per leaf;
(4) more than 1/3 leaf area infected.

4. Ordinal scales with defined intervals based on continuous variables.

Assessment of infection by *Oculimacula (Tapesia) yallundae* (PSDCHE) and *Oculimacula (Tapesia) acuformis* (PSDCHA) causing eyespot of cereals (EPPO Standard PP 1/28 *Efficacy evaluation of fungicides against eyespot of cereals*):
(1) no symptoms;
(2) <50% of tiller circumference attacked at place where infection is most severe;
(3) more than 50% of tiller circumference attacked at place where infection is most severe, but tissue still firm;
(4) 100% of tiller circumference attacked with tissue rotted (softening).

Usually such a scale is, at least partly, logarithmic.

Tobacco leaf area affected by *Peronospora hyoscyami* (PEROTA) (EPPO Standard PP 1/68 *Efficacy evaluation of fungicides against* Peronospora hyoscyami) causing blue mould of tobacco:
(1) no infection
(2) up to 5% of leaf area affected
(3) 5–25% of leaf area affected
(4) 25–50% of leaf area affected
(5) 50–100% of leaf area affected.

Although these scales are apparently logarithmic, there is practically never a constant logarithmic step from the central value of each class to the next. So, although in theory the linear scores corresponding to a logarithmic scale could be analysed as a continuous variable corresponding to a simple transform of the original variable, this case hardly arises, for the scales are not truly logarithmic. In addition, the assignment of the value 1 to the zero class is heterogeneous with the rest of the scale.

Another point which should be stressed is that the classes are defined by the intervals of a continuous variable. In the case of tobacco leaves affected by *Peronospora hyoscyami* (above), the observer simply estimates the category of infection (e.g. class 4 or in class 5), rather than having to distinguish between e.g. 50% or 51% (which is manifestly impossible).

In a few cases, descriptive categories are mixed with defined intervals e.g. the assessment of apple leaves for *Podosphaera leucotricha* (PODOLE) (EPPO Standard PP 1/69 *Efficacy evaluation of fungicides against* Podosphaera leucotricha):
(1) no powdery mildew;
(2) slight attack (scattered patches of powdery mildew);
(3) moderate to strong attack (up to half the leaf surface mildewed);
(4) very strong attack (over half the leaf surface mildewed; leaf margins beginning to roll in and dry up).

5. Ordinal scales with classes defined by their central values.

These are the scales which are best regarded as aids to estimation. The visual keys are the most frequent (e.g. for *Cercospora beticola* (CERCBE) PP 1/1, *Phytophthora infestans* (PHYTIN) PP 1/2, etc.). The keys usually serve to estimate % surface area affected, and have been accurately calibrated. The steps are usually chosen to conveniently cover the range of attack expected, e.g. 1, 5, 10, 25, 50 and allow appropriate interpolation, rather than in a regular nearly logarithmic sequence (which would be preferable if such a scale were used for scoring).