Guidelines for the Design and Statistical Analysis of Experiments in Papers Submitted to *ATLA*

Michael F.W. Festing

MRC Toxicology Unit, University of Leicester, P.O. Box 138, Lancaster Road, Leicester LE1 9HN, UK

Summary — *In vitro* experiments need to be well designed and correctly analysed if they are to achieve their full potential to replace the use of animals in research. An "experiment" is a procedure for collecting scientific data in order to answer a hypothesis, or to provide material for generating new hypotheses, and differs from a survey because the scientist has control over the treatments that can be applied. Most experiments can be classified into one of a few formal designs, the most common being completely randomised, and randomised block designs. These are quite common with *in vitro* experiments, which are often replicated in time. Some experiments involve a single independent (treatment) variable, whereas other "factorial" designs simultaneously vary two or more independent variables, such as drug treatment and cell line. Factorial designs often provide additional information at little extra cost. Experiments need to be carefully planned to avoid bias, be powerful yet simple, provide for a valid statistical analysis and, in some cases, have a wide range of applicability. Virtually all experiments need some sort of statistical analysis in order to take account of biological variation among the experimental subjects. Parametric methods using the t test or analysis of variance are usually more powerful than non-parametric methods, provided the underlying assumptions of normality of the residuals and equal variances are approximately valid. The statistical analyses of data from a completely randomised design, and from a randomised-block design are demonstrated in Appendices 1 and 2, and methods of determining sample size are discussed in Appendix 3. Appendix 4 gives a checklist for authors submitting papers to *ATLA*.

Key words: Experimental design, statistical methods, animal experiments, in vitro, sample size.

Introduction

In vitro methods have the potential to replace the use of animals in many scientific applications. However, it is vital that experiments using these methods are well designed and correctly analysed if they are to achieve their full potential. The aim of these guidelines is to help to ensure that papers published in *ATLA* conform to the very highest scientific standards with respect to experimental design and statistical methods.

Definition of an Experiment

An experiment is a procedure for collecting scientific data in a systematic way in order to maximise the chance of answering an hypothesis correctly (confirmatory research) or to provide material for the generation of new hypotheses (exploratory research). Sometimes, an experiment is replicated in different laboratories or at different times, but provided that all replications involve the same scientific objective, and the data are suitably combined in the statistical analysis, it is considered a single experiment. Confirmatory research will normally involve formal significance testing, whereas exploratory research will normally involve looking for patterns in the data, and may not involve formal significance testing. However, there may be some overlap between these two types of experiment.

An Investigation May Involve Several Experiments

Where two or more experiments (not replications of the same experiment) are presented in a paper, this should be clearly indicated. Preferably, the experiments should be labelled by numbers or letters.

Experiments and Surveys

A "controlled" experiment is one where some treatment or other manipulation is under the control of the experimenter, and the aim is to discover whether the treatment is causing a response in the experimental subjects.

In contrast, a survey is used to find associations between the effects of some variable, which is not usually under the control of the scientist, and some characteristic of the subjects being investigated. These guidelines are concerned with controlled experiments.

Experimental Design

A well-designed experiment will avoid bias, and will be sufficiently powerful to be able to detect effects likely to be of biological importance. Where it is necessary to determine which variables, such as time, culture medium and cell line, are most important in influencing the results, and whether they interact, factorial designs can be used. In order to ensure that results are repeatable in time or in different laboratories, experiments are sometimes replicated with randomised block designs, but the resulting data must be correctly analysed (see Appendix 1).

Experiments should normally be designed to test a clearly stated hypothesis or other scientific objective, and should not be so complicated that mistakes are made in their execution. Written protocols should always be used. Experiments should be designed so that they can be subjected to a statistical analysis, with the method of analysis being planned when the experiment is designed (though modifications may be needed, if the results do not come out exactly as expected).

The "Experimental Unit"

Each experiment will involve a number of *experimental units*, such as a cage of animals, an animal, or a flask, dish or well of cells, which can be assigned at random to a treatment. In principle, any two experimental units must be available to be assigned to different treatments. In a multiwell plate, it may be impractical to assign each well at random to a different treatment, so a column of wells may all be assigned to the same treatment. In this case, the experimental unit is the column of wells, and the statistical analysis should normally be performed on the data from the whole column rather than on data from individual wells.

Randomisation

Randomisation of experimental units to treatments is essential, because there are often unknown sources of variation which could bias the results. For example, there may be edge effects in multiwell plates, and incubators do not always have the same environment in all locations.

Blinding

Where possible, experiments should be conducted "blind" with respect to the treatments, with samples coded so that their treatment group is unknown until the data are analysed. This is of vital importance in any comparison between laboratories.

The Use of Formal Experimental Designs

A range of formal experimental designs are described in the literature, and most experiments should conform to one of these. The most common are: completely randomised, randomised block, Latin square, crossover, repeated measures, split-plot, incomplete block and sequential designs.

These formal experimental designs have been developed to take account of special features and constraints of the experimental material and the nature of the investigation. Within each type of design, there is considerable flexibility in terms of choice of treatments and experimental conditions, but standardised methods of statistical analysis are usually available. For example, where experiments produce numerical data, they can often be analysed by using some form of the analysis of variance (ANOVA), if necessary following a scale transformation. Investigators are encouraged to state which of these designs they have used, as this helps to clarify exactly what has been done.

The completely randomised design is the simplest of all designs, but it has limitations which make it unsuitable for some *in vitro* studies, though it is widely used for whole-animal experiments. In contrast, the randomised block design, with replication in time, is widely used for *in vitro* experiments. These two designs are discussed in more detail, with examples, in Appendix 2.

Factorial Designs

Factorial experiments are ones in which a number of independent variables, such as culture medium and cell line, are altered within a single experiment. Strictly, a factorial "design" is really an arrangement of treatments which can be used independently of the formal experimental design. Thus, the example given in Appendix 2 is of a randomised block experimental design, but with a factorial arrangement of the treatments. The factors are the presence or absence of 12-*O*-tetradecanoylphorbol 13-acetate and the presence or absence of genistine, resulting in a 2 × 2 factorial layout. The purpose of such simple factorial designs is usually to see whether the factors interact or potentiate each other, but it also provides a way of studying the effects of both treatments in a single experiment.

In some situations, there are a large number of factors which might influence the results of an experiment, such as cell line, medium, supplements, culture conditions, time, and the presence or absence of other treatments. Special "fractional factorial" designs could be used to explore which of these have a large effect on the results, without having to use excessive numbers of experimental units, though a professional statistician may need to be consulted, to ensure that the designs will be used effectively.

Determining Sample Size

Deciding how large an experiment needs to be is of critical importance with *in vivo* experiments, because of the ethical implications of using animals or humans in research. With *in vitro* experiments, the main constraints are cost, resources and time, and there is usually no serious ethical constraint. Papers involving animals or humans in controlled experiments, which are submitted to *ATLA*, should justify the numbers of animals or people which were used.

Two methods are available for determining sample size. The *power analysis* and *resource equation* methods, described in Appendix 3.

Need for Statistical Analysis

The results of most experiments should be assessed by an appropriate statistical analysis, even though, in some cases, the results may be so clear-cut that it is obvious that any statistical analysis would not alter the interpretation. The materials and methods section should describe the statistical methods used in analysing the results. The aim of the statistical analysis is to extract all the information present in the data, in such a way that it can be interpreted, taking account of biological variability and measurement error. It is particularly useful in preventing unjustified claims about the effect of a treatment, when the results could probably be explained by sampling variation. Note that it is possible for an effect to be statistically significant, but of little or no biological importance. The magnitude of any significant effects should always be quoted, with a confidence interval, standard deviation or standard error to indicate its precision, and exact p-values should normally be given, rather than stating, say, that $p < 0.05$.

Lack of statistical significance should not be used to claim that an effect does not exist, because this may be due to the experiment being too small or the experimental material being too variable. Where an effect is not statistically significant, a power analysis (see Appendix 3) can sometimes be used to show the size of biological effect that the experiment was probably capable of detecting.

Examining the Raw Data

The raw data should be studied for consistency and for any obvious typographical errors. Graphical methods, which are now available in most statistical packages, are helpful, particularly if individual observations can clearly be seen. “Outliers” should not be discarded, unless there is independent evidence that the observation is incorrect, such as a note taken at the time that the observation was recorded, expressing doubt about its credibility. In this case, the reasons for its exclusion should be explicitly stated. It is sometimes useful to do the statistical analysis with and without the suspect data, to see whether this alters the conclusions.

Methods of Statistical Analysis

The method of statistical analysis will depend on the purpose of the study, the design of the experiment, and the nature of the resulting data. Categorical or qualitative data, where counts and proportions are to be compared, will be analysed by using different methods from those used with quantitative or measurement data. In these guidelines, it is only possible to give a brief outline of the main methods which are recommended for papers submitted to *ATLA*.

Statistical Analysis of Quantitative Data and Comparison of Means or Medians

Quantitative data are usually summarised in terms of the mean, “n” (the number of subjects), and the standard deviation as a measure of variation. The median, “n”, and the inter-quartile range may be preferable for data which are clearly skewed. The statistical analysis is usually used to assess whether the means, medians or distributions of the different treatment groups differ. More rarely, and not discussed here, the aim will be to compare the variation within the different groups.

Quantitative data can be analysed by using “parametric” methods, such as the t test or the ANOVA, or by using non-parametric methods, such as the Mann-Whitney test. Parametric tests are usually more versatile and more powerful, so are preferred, but depend on the assumptions that the variances are approximately the same in each group, that the residuals (i.e. deviation of each observation from its group mean) have a normal distribution, and that the observations are independent of each other. The first two of these assumptions should be investigated as part of the analysis, by studying the residuals (see Appendix 2). The last one depends on good experimental design. If these assumptions are not met, it may be possible to transform the data in such a way that they are met.

Transformations

A scale transformation can often be used prior to a parametric analysis, if the assumptions listed above are not met, though most parametric methods are robust against moderate departure from the assumptions. A log transformation is often appropriate when the dependent variable is a concentration. This cannot be less than zero, and may have several moderately high observations, but may have a small number of very high values. Taking logs (one can be added to each observation, if some are zero) often normalises the data. Where data are expressed as proportions or percentages, such as the proportion of stained cells in a sample of cells, and when many of the observations are less than 20% or more than 80%, the data distribution will be skewed. In this case, a suitable transformation is the logit which is $\log_e(p/[1 - p])$, where p is the proportion, or an angular transformation, as discussed in many textbooks. These stretch out the values that are less than 0.2 or more than 0.8, so normalising the data.

Counts such as the numbers of cells in a haemocytometer square, can sometimes produce data which can be analysed by the ANOVA. If the mean count is low, say less than about five, then the data may have a Poisson distribution. This can be transformed by taking the square root of the observations. However, if the mean count is reasonably high, no transformation may be needed.

Parametric Statistical Analysis

Student’s t test should not be used when more than two treatment groups are to be compared. If there are several groups, it lacks power, and multiple testing increases the chance of a false-positive result. Where there are two or more groups, and particularly with randomised block or more-complex designs, the ANOVA should be used.

The ANOVA is usually used initially to test the overall hypothesis that there are no differences among treatment means. If no significant differences are found, further comparisons of means should not normally be done. Where the ANOVA results are significant, say at $p < 0.05$, and there are several groups being compared, either *post hoc* comparisons or orthogonal contrasts can be used to study differences amongst individual means. A range of *post hoc* comparison methods are available, which differ slightly in their properties. These include Dunnett's test for comparing each mean with the control, and Tukey's test, Fisher's protected least-significant difference test, the Newman-Keuls test, and several others for comparing all means. Authors should state which tests have been used. Note that all these tests use the pooled within-group standard deviation obtained from the ANOVA. The ANOVA followed by individual t tests to compare means, not using the pooled standard deviation, is not acceptable, because each test will lack power. Orthogonal contrasts can be used to compare groups of means or, when dose levels are equally spaced on some scale, to assess linearity and deviations from linearity of response to the dose.

Where there are several dose levels, assessing a dose–response relationship by using regression, or orthogonal contrasts (where appropriate), should be considered in preference to comparing each dose level with the control.

The best estimate of the pooled standard deviation is obtained as the square root of the error mean square in the ANOVA. In fact, this is the only estimate of the standard deviation which is available for a randomised block design. Thus, when presenting means either in tables or graphically, this estimate of the standard deviation should be used. It will, of course, be the same for each group.

Several Dependent Variables

Where there are several dependent variables (characters), each can be analysed separately. However, if the variables are correlated, the analyses will not be independent of one another. Thus, if sampling variation resulted in a false-positive or false-negative result for one character, the same thing may happen for another character. A multivariate statistical analysis, such as principal components analysis, should be considered in such cases (1).

Non-parametric Tests

Several of these methods, such as the Mann-Whitney test, replace the individual observations by their ranks, resulting in the loss of some information; hence, these methods often lack power in situations where parametric tests are appropriate, but they may be more powerful in situations where the parametric test is not appropriate.

There are several non-parametric tests for equality of population means or medians. The Wilcoxon rank sum test and the Mann-Whitney test are equivalent. These are the non-parametric equivalents of the two-sample t test. The Kruskal-Wallis test is the non-parametric equivalent of the one-way ANOVA, where several groups are to be compared, and a non-parametric equivalent of a *post hoc* comparison can be used, provided that the overall test is significant (2). A version of the Wilcoxon test can also be used as the non-parametric version of the paired t test. The Friedman test is the non-parametric equivalent of the randomised block or repeated measures ANOVA. There are several other non-parametric tests which are appropriate for particular circumstances.

Correlation

The product–moment correlation is the commonest method for assessing the strength of *linear* relationship between two numerical variables, X and Y. Both X and Y are assumed to be subject to sampling variation. It does not assume that variation in X causes variation in Y. Where the data are shown graphically, regression analysis is sometimes used to give the best-fitting straight line. However, there are two lines that can be fitted, namely, the regression of X on Y and the regression of Y on X. Both should normally be plotted, if there is no suggestion of a causal relationship. Note that a change of scale will alter the correlation, and that a non-linear relationship will result in a low correlation, even if the two variables are strongly associated. In such circumstances the correlation of ranks may be more appropriate. There are several other forms of correlation, depending on whether the variables are measurements or ranks, or are dichotomous.

Regression

Regression analysis can be used to quantify the relationship between a variable *X*, which is presumed to cause changes in a variable *Y*. The *X* variable is assumed to be measured without error. Linear regression can be used to fit a straight line of the form $Y = a + bX$, where *a* and *b* are constants which are estimated from the data by using a method called least-squares. Quadratic regression can be used to fit a curve to the data points. Many other types of curve can be fitted, some of which have useful biological interpretations.

Regression analysis and the ANOVA are closely related so that a regression, say of response on dose level, can sometimes be included as part of the ANOVA, by using orthogonal polynomials (3).

The most usual statistical test in regression analysis is of the null hypothesis that there is no linear relationship between *X* and *Y*. A test to determine whether there is a quadratic relationship, would be a test of whether a curve gives a significantly better fit than a straight line.

Categorical Data

Categorical data consist of counts of the number of units with given attributes. These attributes can be described as "nominal" when they have no natural order, such as different cell lines. They are described as ordinal, when they have a natural order, such as low, medium and high dose levels, which may also be defined numerically. Such categorical data are often presented in the form of tables or, possibly, as proportions or percentages.

Proportions or percentages should be accompanied by a confidence interval or standard error, and "n" should be clearly indicated. The usual method of comparing two or more proportions is a contingency table χ^2 analysis, which tests the null hypothesis that rows and columns are independent. The method is only accurate if none of the expected values are less than five. Where some cells have very small numbers, Fisher's exact test should be used. Other acceptable methods of analysis are available, and are described in various texts.

Presentation of the Results

Where individual means are quoted, they should be accompanied by some measure of variation. If the aim is to describe the variation amongst individuals which contribute to the mean, the standard deviation should be given. Avoid using the ± sign. It is better to use a designation such as "9.6 (SD 2.1)", because this avoids any confusion between standard deviation and standard error. Where the aim is to show the precision of the mean, a confidence interval should be used (preferably) or a standard error (for example, 9.6 SE 1.2), but in this case "n" must also be indicated. Where two means are being compared, the difference between them should be quoted, together with a confidence interval.

Non-parametric data should quote medians and the inter-quartile range or some other estimate of variation. Where proportions or percentages are given, a standard error or confidence interval and "n" should also be given. Significance levels should not be quoted without indicating the size of an effect, as statistical significance and biological importance are not synonymous.

Computers sometimes give outputs with excessive numbers of digits. These should be rounded to take account of the precision of the raw data.

Graphical Presentation of Data

Graphs showing individual points rather than error bars are preferred. Where error bars are shown on graphs or bar diagrams, there should be a clear indication of whether these are standard deviations, standard errors or confidence intervals, and the number of observations should be clearly indicated in the text or figure caption.

References and Further Reading

There are numerous textbooks on statistics and experimental design. Most are directed at specific disciplines such as agriculture, psychology or clinical medicine, but the methods are general and applicable to *in vitro* experiments. A few are listed below, some of which are general textbooks, while others are more specialised.

References

1. Everitt, B.S. & Dunn, G, (2001). *Applied Multivariate Data Analysis,* 342pp. London & New York: Arnold.
2. Sprent, P. (1993). *Applied Nonparametric Statistical Methods*, 342pp. London, Glasgow & New York: Chapman and Hall.
3. Altman, D.G. (1991). *Practical Statistics for Medical Research*, 610pp. London, Glasgow & New York: Chapman and Hall.

Further Reading

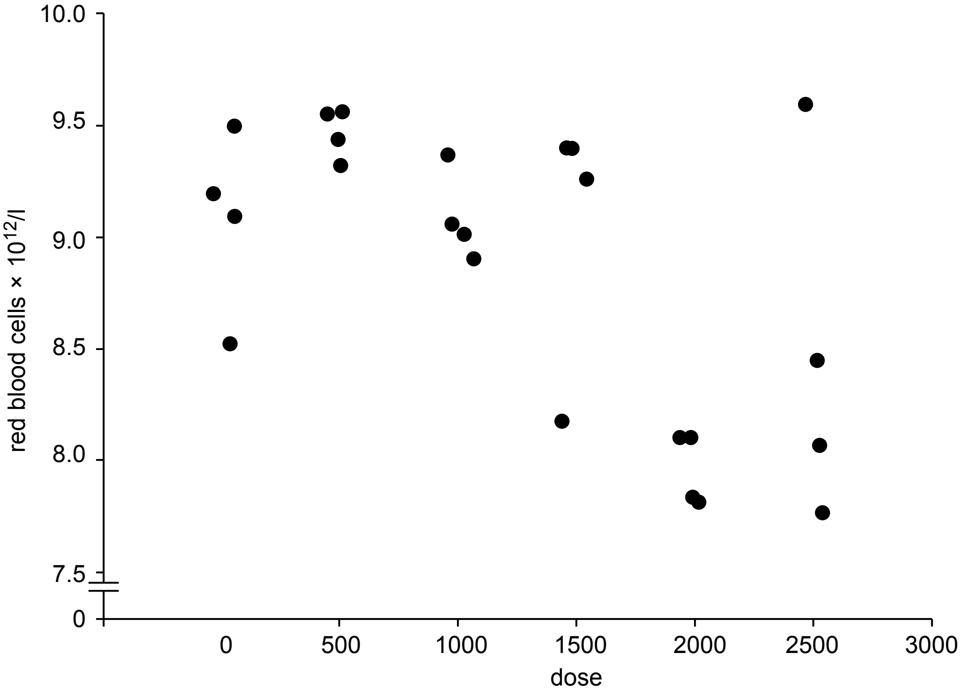
1. Cox, D.R. (1958). *Planning Experiments,* 208pp. New York: John Wiley & Sons
2. Howell, D.C. (1999). *Fundamental Statistics for the Behavioral Sciences*, 494 pp. Pacific Grove, London & New York: Duxberry Press.
3. Mead, R. & Curnow, R.N. (1983). *Statistical Methods in Agriculture and Experimental Biology*, 335pp. London & New York: Chapman and Hall.
4. Maxwell, S.E. & Delaney, H.D. (1989). *Designing Experiments and Analyzing Data*, 902pp. Belmont, CA, USA: Wadsworth Publishing.

Appendix 1

The Completely Randomised Design

This is the most common design with animal experiments, but is less common with *in vitro* experiments. The experimental unit is often an animal or a dish or well of cells, which can be assigned at random to a treatment group. There can be any number of experimental units, which should be chosen to be as homogeneous as possible. Typically, the experiment will be performed at one time in one location or, alternatively, time and location will be assumed to have negligible effects on the experimental material. The design can accommodate any number of treatment groups, and unequal numbers in each group usually present no problem. It can usually be analysed by using the analysis of variance (ANOVA), provided that the data are appropriate. Where this is not the case, non-parametric methods can be used. The disadvantage of the design is that it cannot take account of variation among the experimental units over time or in different laboratories. Table I shows the red blood cell (RBC) counts of mice administered a test compound "X" at various dose levels. The mice were assigned to the treatments at random, with the restriction that four mice were assigned to each dose level. Dose levels were

Figure 1: Plot of red blood cell counts against dose level of compound "X" to show scores for each animal



Note that some "jitter" has been added on the X-axis, so that the points do not sit on top of each other.

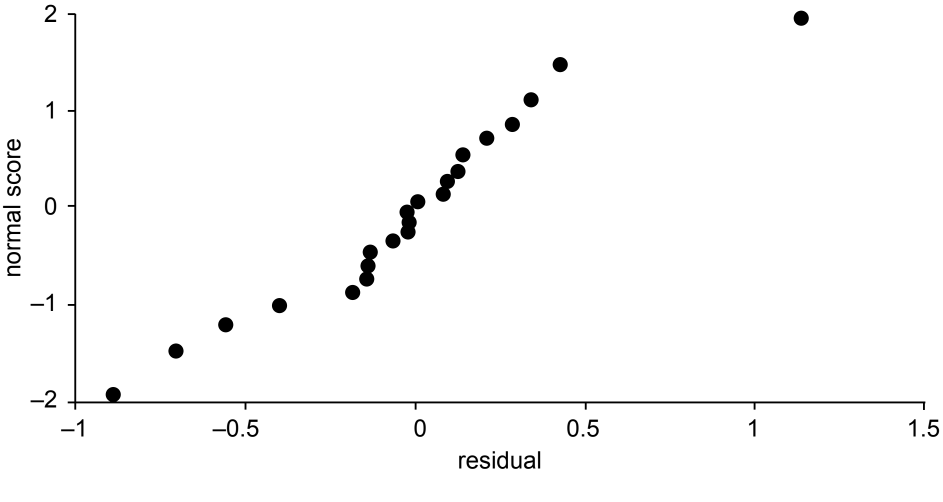
Table I: Red blood cell counts (units) in mice given various doses of a test compound

	Dose level					
	0	500 mg/kg	1000 mg/kg	1500 mg/kg	2000 mg/kg	2500 mg/kg
Mouse 1	9.50	9.44	9.06	9.40	8.10	9.60
Mouse 2	8.52	9.32	8.90	9.40	8.10	8.07
Mouse 3	9.20	9.55	9.37	8.17	7.82	8.45
Mouse 4	9.09	9.56	9.01	9.27	7.83	7.77

equally spaced on an arithmetic scale. The aim of the experiment was to determine whether the test compound affected RBC counts in a dose-dependent manner.

The statistical analysis has been performed by using the MINITAB statistical package (1). Dedicated statistical packages should generally be used in preference to spreadsheets. The first step is to look at the raw data. Figure 1 is a plot of the individual observations against dose level. There is some visual evidence that the RBC counts are lower at the higher dose levels. As the data are quantitative, the ANOVA can be used to test the null hypothesis that there is no difference among dose level groups, pro-

Figure 2: Normal probability plot of residuals for data in Table I



This should be a straight line if the residuals have a normal distribution. The outlier at the highest dose level shows up clearly. Although there is some departure from normality, it does not appear to be sufficiently serious to invalidate a parametric statistical analysis.

vided that the residuals have a normal distribution and the variation is approximately the same in each group. Figure 2 is a normal probability plot of the residuals (i.e. the deviations from the group means), provided as an option when using MINITAB. This will be a straight line if the residuals have a normal distribution. Some judgement is necessary in deciding whether this is a sufficiently good approximation to a straight line to be accept-

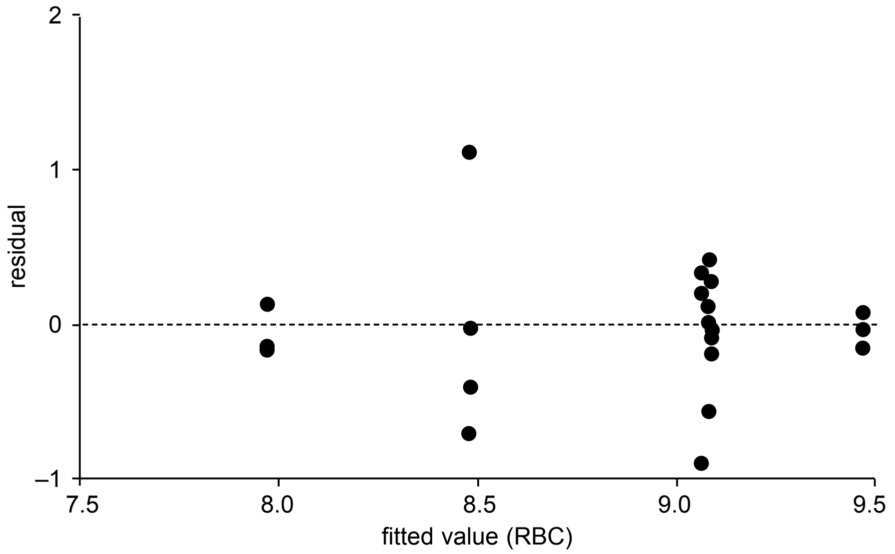
Table II: One-way analysis of variance (ANOVA) for the data in Table I, with post hoc comparisons with Dunnett's test, obtained by using the MINITAB statistical package

```
One-way ANOVA: red blood cell (RBC) versus dose

ANOVA for RBC
Source     DF       SS        MS       F       P
dose        5     5.850     1.170    5.64    0.003
Error      18     3.734     0.207
Total      23     9.584

                                   Individual 95% CIs for mean
                                   based on pooled SD
Level      n      Mean      SD     ----+---------+---------+---------+-
0          4     9.0775   0.4101               (--------*--------)
500        4     9.4675   0.1124                      (--------*--------)
1000       4     9.0850   0.2014               (--------*--------)
1500       4     9.0600   0.5965               (--------*--------)
2000       4     7.9625   0.1588    (--------*--------)
2500       4     8.4725   0.8015          (--------*--------)
                                   ----+---------+---------+---------+-
Pooled SD =      0.4555               7.70      8.40      9.10      9.80

Dunnett's comparisons with a control

Family error rate = 0.0500
Individual error rate = 0.0129

Critical value = 2.76

Control = level (0) of dose

Intervals for treatment mean minus control mean

Level    Lower     Centre     Upper   ----+---------+---------+---------+-
500     -0.4994    0.3900    1.2794                 (-----------*-----------)
1000    -0.8819    0.0075    0.8969           (-----------*-----------)
1500    -0.9069   -0.0175    0.8719           (-----------*-----------)
2000    -2.0044   -1.1150   -0.2256    (-----------*-----------)
2500    -1.4944   -0.6050    0.2844        (-----------*-----------)
                                      ----+---------+---------+---------+-
                                        -2.0      -1.0       0.0       1.0
```

DF = degrees of freedom; SS = sum of squares; MS = mean square; F = variance ratio; P = probability; CI = confidence interval; SD = standard deviation able. In this case, the deviation is not serious, though the outlier at the highest dose level stands out. This observation should be checked to ensure that it is correct. Figure 3 is a plot of fits (i.e. group means) versus residuals (deviations from group means), to see whether the variation is approximately the same in each group. Again, such a graph should be an available option with all good statistical packages. In this case, the groups with most variation seem to be those with middle dose levels. The absence of any clear-cut pattern suggests that most of the variation is random, and does not increase as the means increase. The two plots of Figures 2 and 3 suggest that the assumptions for a parametric analysis are reasonably well met. The results of the ANOVA are shown in Table II, with Dunnett's test used for the *post hoc* comparisons. The ANOVA table has a p-value of 0.003 for the null hypothesis that there are no differences between dose level groups. The dose levels, n and individual means and standard deviations are shown with individual 95% confidence intervals (CI) for each mean shown graphically. The pooled standard deviation is 0.4555 units. The meaning of abbreviations such as DF, SS and MS will be given in the reference manual for each software package, or in statistical texts.

The Dunnett's test comparisons of each mean with the control level have a family error rate of 0.05. That means that 5% of similar experiments would be expected to give one or more false-positive results, when in fact there is no real difference between the groups. The individual error rate is 0.0129. This is less than 0.05, because of the need to take account of the fact that five tests have been conducted. Had each comparison been judged "significant" with $p \leq 0.05$, the chance of a false-positive conclusion for the whole experiment would have been much higher than 0.05. The lower part of the table shows the difference between the mean of each group and the control mean ("Center") and the upper and lower 95% confidence intervals for this difference. If the confidence inter-

Figure 3: Plot of fits (group means) versus residuals for RBC data



There is no good evidence from this plot that the variation differs between groups. In particular, the variation does not seem to be greater in groups with the highest means.

val does not cover a difference of zero, it will be concluded that the difference is significant at $p < 0.05$. In this case, only the 2000 dose level differs significantly from the control level by –1.11 units, with a 95% confidence interval of between –2.00 and –0.22 units (rounding to three significant digits).

However, this analysis does not really answer the question of whether there is a dose-related change in the RBC counts. All it shows is that the means differ and that one of the dose levels differs from the control. A regression analysis would really be more appropriate. Table III shows the MINITAB output from a regression analysis of RBC counts on dose levels. This produces an ANOVA table which tests the null hypothesis that there is no relationship between dose level and RBC counts. This hypothesis will be rejected at $p = 0.003$, so it would be legitimate to conclude that there is a significant dose-related effect. Note that in the ANOVAs of both Tables II and Table III, the total sum of squares (SS) is the same (9.584). In fact, the ANOVA and regression are closely related. In Table II, there are five degrees of freedom (DF) associated with dose levels. One of these can be used to test whether there is a linear trend. The SS for this will be 3.2702, as shown in the ANOVA table of Table III. This is equivalent to using orthogonal polynomials to test whether there is a linear trend. It is possible to use another DF to test whether a quadratic curve would give a better fit than a straight line. In this case, it does not (details not shown). At the top of Table III, the regression equation is shown as RBC = 9.39 – (0.000432 × dose). This says that the intercept (RBC count when dose is zero) is 9.39 units, and that it declines by 0.000432 units for every unit increase in the dose level. The R-Sq (adj) value indicates that

Table III: Regression analysis of the red blood cell (RBC) data of Table I

Regression analysis: RBC versus dose

The regression equation is RBC = 9.39 – (0.000432 × dose)

Predictor	**Coef**	**SE coef**	**T**	**P**
Constant	9.3945	0.1939	48.46	0.000
dose	–0.0004323	0.0001281	–3.38	0.003

S = 0.5357 R-Sq = 34.1% R-Sq(adj) = 31.1%

Analysis of variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	1	3.2702	3.2702	11.39	0.003
Residual error	22	6.3139	0.2870		
Total	23	9.5842			

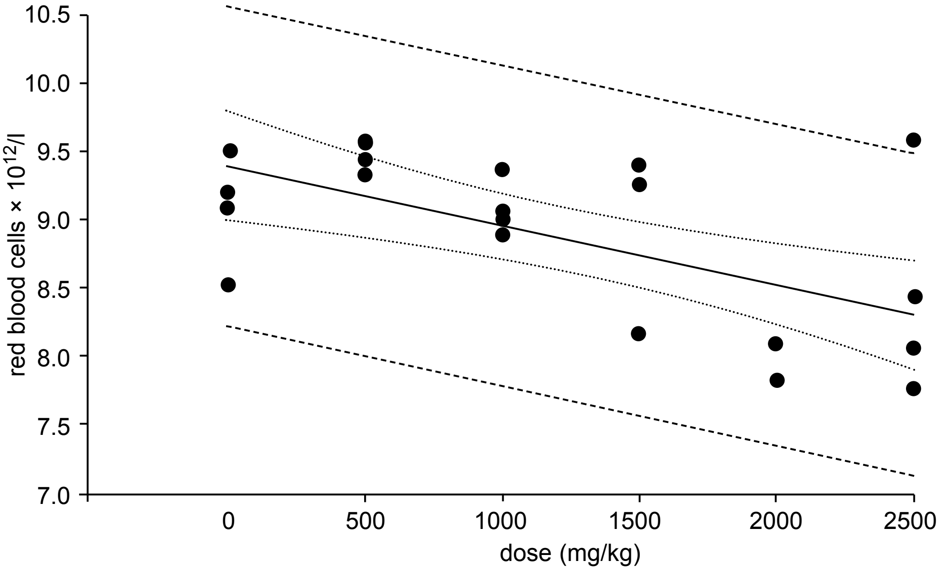
Unusual observations

Obs	**dose**	**RBC**	**Fit**	**SE fit**	**Residual**	**St resid**
21	2500	9.600	8.314	0.194	1.286	2.58R

R denotes an observation with a large standardised residual.

DF = degrees of freedom; Coef = coefficient; SE coef = standard error of the coefficient ; T = Student's T; P = probability; SS = sum of squares; MS = mean square; F = variance ratio; SE fit = standard error of fit; St resid = standard residual.

Figure 4: Regression plot of the red blood cell (RBC) count scores versus dose level, showing the best-fitting straight line



——— *Regression,* -------- *= 95% confidence interval,* ·············· *= 95% prediction interval for individual points.*

RBC = 9.39452 – 0.0004323 dose, S = 0.535721, R-Sq = 34.1%, R-Sq(adj) = 31.1%.

31.1% of the variation in RBC counts is associated with a linear variation in the dose level.

At the very bottom of the table there is a note about unusual observations, in this case, number 21, which is the very high observation in the top dose group shown in Figure 1.

Finally, it may be appropriate to present a graph of the results with the best fitting straight line, as shown in Figure 4. This also shows the 95% confidence intervals for the mean at each dose level (inner dotted lines) and the 95% prediction level for individual points (i.e. 95% of observations should fit within the outer dotted lines).

Appendix 2

The Randomised Block Design

This is a common design for *in vitro* studies with the experiment being replicated in time and/or in different laboratories. A "block" is like a "mini-experiment", with all treatments being represented.

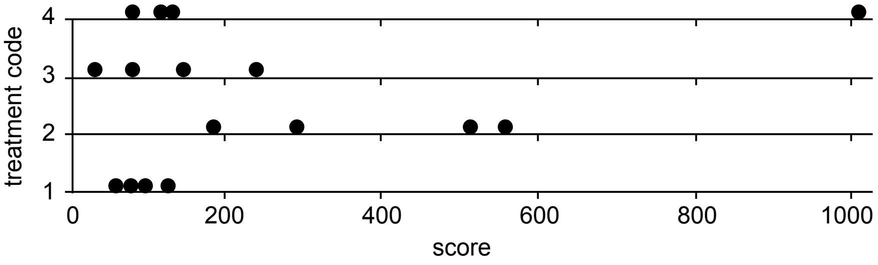
Table IV shows the results of an experiment in which four flasks of a transgenic cell line were assigned either to a control medium, medium with 12-*O*-tetradecanoylphorbol 13-acetate (TPA; T), medium with genistine (G), or medium with both T and G. This is a factorial arrangement of the treatments. Cultures were scored for the number of cells showing a particular phenotype (real data, but disguised). The "mini-experiment", usually called a block or a replicate, was repeated four times on different days, providing 16 observations. The aim was to find out whether T and

Table IV: Data used to illustrate the method of statistical analysis of a randomised block factorial experimental design

Treatment[a]	Block 1	Block 2	Block 3	Block 4
0	100	81	62	128
T	514	187	294	558
G	35	82	148	241
T+G	120	84	134	1011

[a]0 = medium alone, T = medium + 12-O-tetradecanoylphorbol 13-acetate; G = medium + genistine; T+G = medium + TPA and genistine.

Figure 5: Dotplot of raw data for the randomised block design



Blocking has been ignored for this plot, and treatments have been coded 1 (control) to 4. Note that the variation in group 2 seems to be greater than that in groups 1 and 3, and there seems to be an outlier in group 4.

G affected the results, and whether they interacted or potentiated each other. There was considerable variation between the blocks, with the values in block 3, for example, being about half of those in block 4. Such differences are common with *in vitro* studies, and they must be removed in the statistical analysis, if they are not to completely obscure the treatment differences. Note that the minimum block size is the same as the number of treatments, but it can be larger. For example, in the above experiments, there could have been eight flasks in each block, with two being assigned to each treatment.

The first step in an analysis should be to study the raw data. Figure 5 shows a dotplot of each treatment group, ignoring the blocking factor and coding the treatments simply as 1–4. This suggests that there may be an outlier in group 4. This should be checked, to ensure that it is not a typographical error, though in this case it is a valid observation. The dotplot also suggests that groups with low mean values appear to be less variable than groups with higher mean values. This could present a problem, as the ANOVA assumes approximately equal variation in each group.

Table V: Analysis of variance (ANOVA): LogScore versus replicate, T and G

Factor	Type	Levels	Values			
Replicate	random	4	1	2	3	4
T	fixed	2	1	2		
G	fixed	2	1	2		

ANOVA for LogScore

Source	DF	SS	MS	F	P
Replicate	3	0.74028	0.24676	3.91	0.049
T	1	0.77214	0.77214	12.24	0.007
G	1	0.04627	0.04627	0.73	0.414
T × G	1	0.09994	0.09994	1.58	0.240
Error	9	0.56775	0.06308		
Total	15	2.22637			

Means

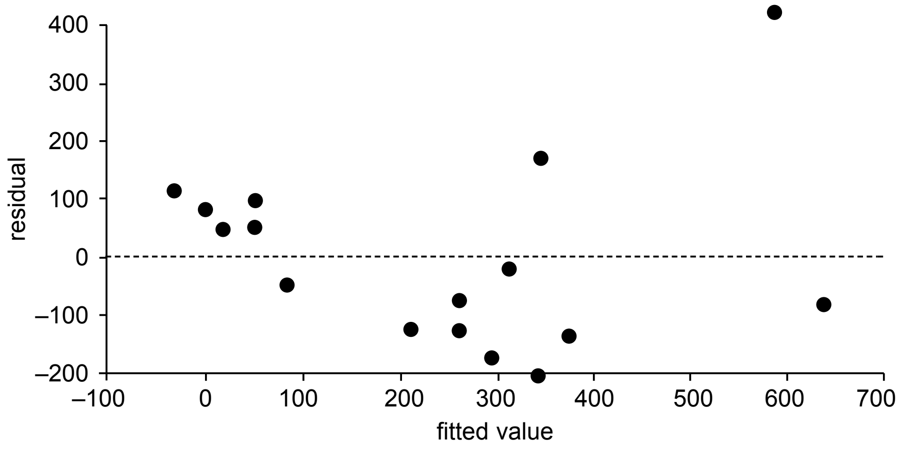
T		N	LogScore
1		8	1.9773
2		8	2.4166

G		N	LogScore
1		8	2.2507
2		8	2.1432

T	G	N	LogScore
1	1	4	1.9520
1	2	4	2.0025
2	1	4	2.5494
2	2	4	2.2838

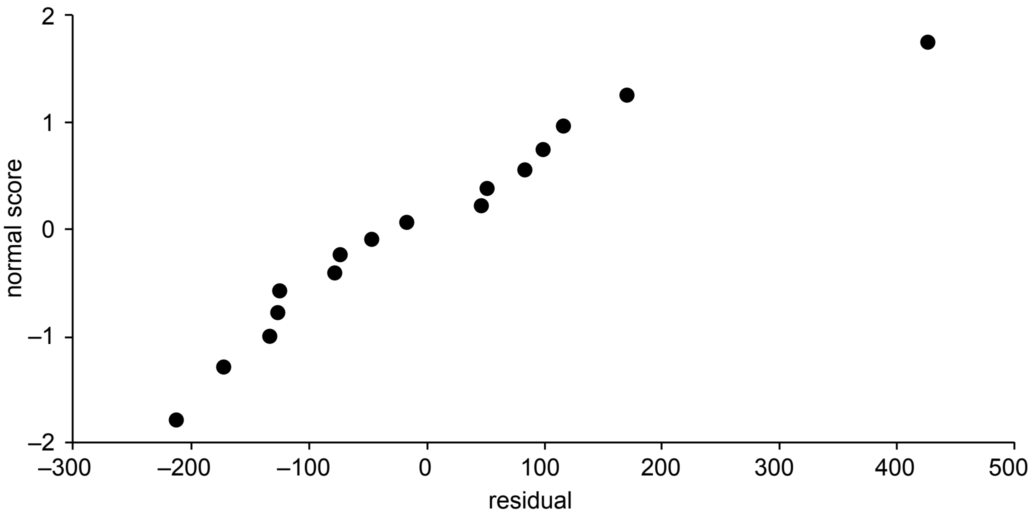
T = medium + TPA; G = medium + genistine; T + G = medium + TPA and genistine.

Figure 6: Residuals versus fits plot for the data in Table IV



The next step is usually to study the residuals plots, as was done with the previous example. Figures 6 and 7 show the two plots. Figure 6 should be a straight line, if the residuals have a normal distribution, but it is slightly curved, and the outlier stands out. Figure 7 suggests that low fitted values are associated with low variation in the residuals, though the effect is not very marked. Some judgement is necessary in deciding whether these departures are sufficiently serious to reject the ANOVA based on these

Figure 7: Normal probability plot of the residuals for the data in Table IV



Note that the variation seems to increase with higher fitted values.

raw data. Generally, departures from the assumptions decrease the power of the experiment, resulting in fewer effects being declared significant. In this case, data were transformed to the $\log_{10}$ of the scores. The residuals plots were again studied (not shown), and it was concluded that the assumptions for the ANOVA were met on this scale. Even the outlier in group 4 was no longer obviously an outlier on the logarithmic scale.

The results of the ANOVA, as produced by MINITAB, are shown in Table V. Note first that the variation between the blocks or replicates has been removed in the analysis. There is no significant interaction between T and G. In other words, the response to G does not depend on whether the cells have first been treated with T, and only the response to T is statistically significant at p = 0.007 (it was 0.034 on the untransformed scale). On the logarithmic scale, treating the cells with T increased the score by 0.44 units. The pooled standard deviation is obtained as the square root of the error mean square, i.e. the square root of 0.06308 = 0.25 units. A confidence interval for difference between cells receiving and not receiving T can be constructed, as described in most text books as $\pm t_{0.05} \times$ SD/root(n). In this case, t is based on 9 degrees of freedom (see ANOVA table), and at the 0.05 level it is 2.262. Thus, the 95% confidence interval is $0.44 \pm 2.262 \times 0.25/3 = 0.44 \pm 0.18$. So the 95% confidence interval for the true effect of adding T is an increase of between 0.26 and 0.62 units on the logarithmic scale. Note that the means on the log scale can be back-transformed to the original scale to provide the geometric means, but the standard deviation and differences between means cannot be meaningfully back-transformed. If necessary, the analysis can be conducted on the logarithmic scale, and the data can be presented on the original scale, provided that this is clearly indicated.

Appendix 3

Methods for Determining Sample Size

There are two main methods for determining sample size, as noted in the guidelines. The *power analysis* method is most appropriate for relatively simple but expensive experiments, such as clinical trials and some animal and *in vitro* experiments. It is particularly useful for experiments likely to be repeated several times, such as drug safety screening. However, it is not always possible to use the method, as it requires a knowledge of the standard deviation of the character of interest, and some estimate of the effect size deemed to be of biological importance. This information may be difficult to obtain for one-off experiments, for complex experimental designs, or for ones involving many dependent variables, such as those using microarrays. In such circumstances, the *resource equation* method, which depends on the law of diminishing returns, may be useful.

Power analysis

This method depends on the mathematical relationship between: the effect size of interest; the standard deviation; the chosen significance level; the chosen power; the alternative hypothesis; and the sample size. Any five of these can be fixed, and this will determine the sixth one. The formulae are complex, but software is now available for doing the calculations.

The first step is to decide the type of statistical analysis that will be used to analyse the experiment. Briefly, the analysis will depend on the aims of the experiment, the number of treatments, and type of data which will be produced. If there are two groups with quantitative data, the results could be analysed by using Student's t test, assuming that the data are appropriate. If the aim is to compare the proportions of dead and alive cells in two or more groups, a χ^2 test could be used. If a dose-response relationship is being studied, regression analysis could be used.

The second step is to decide the effect size to be detected. For example, how much of a reduction in neutral red uptake in a cell culture or in RBC levels in a mouse experiment is of biological significance? A 10% reduction may not be of much interest, whereas a 50% reduction may be. Where several different characters are being studied (for example, neutral red uptake, total protein in a well, and a cell count), the calculations should concentrate on the most important of these. With categorical data (dead/alive), the difference in two or more proportions that is likely to be of biological importance must be decided.

The third step is to estimate the standard deviation among experimental units, assuming quantitative data. As the experiment has not yet been conducted, this has to come from previous experiments or from the open literature. Unfortunately, it may be difficult to get an accurate estimation of the standard deviation, and small differences in standard deviation may translate into quite large differences in the estimated sample size. If two or more proportions are to be compared, the standard deviation is a function of the proportions, so does not need to be separately estimated. Note that sometimes the effect size needs to be specified in standard deviation units by dividing the effect by the standard deviation.

The fourth step is to decide on the significance level to be used. Somewhat arbitrarily, this is often set at 0.05, though other levels can be chosen.

The fifth step is to decide what power the experiment should have. This is the chance that the experiment will be able to detect the specified effect and show it to be statistically significant at the specified significance level. One minus the power is the chance of a false-negative result. A power of somewhere between 80% and 90% is often chosen. However, in *in vitro* tests on the safety of some vaccines, where a false-negative result would have serious consequences, a power of 99% may be specified. Thus, power should be chosen according to the consequences of failing to detect the treatment effect.

Next, the alternative hypothesis usually has to be considered. The null hypothesis is usually that there are no differences among treatments, and the alternative is often that there are differences. However, sometimes there are good biological reasons as to why the difference can only occur in one direction. If this is the case, a "one-sided" statistical test would be used.

Finally, these pieces of information need to be put together in order to obtain an estimate of the required sample size. There are a number of dedicated programs which can be used (2), and several recent versions of statistical software provide some power calculations. For a simple comparison of two samples by using the t test or for comparing two proportions, there is free software on the Web (search for "statistical power" by using a search engine such as www.google.com).

The resource equation

There are occasions when it is difficult to use a power analysis, because there is no information on the standard deviation and/or because the effect size of interest is difficult to specify. The "resource equation" method (3) is based on the law of diminishing returns. Once an experiment exceeds a certain size, adding more experimental units gives very little more information. An appropriate size can be roughly determined by the number of DF for the error term in the analysis of variance or t test given by the formula:

$$E = N - T - B$$

where E, N, T and B are the error, total, treatment and block degrees of freedom in the ANOVA. The suggestion is that E should be somewhere between 10 and 20. Thus, for the data given in Table I, E = 23 – 5 = 18, because there are 24 total observations and six dose levels, and for the data given in Table IV, E = 15 – 3 – 3 = 9, because there are sixteen total observations, four treatment combinations and four blocks. Thus, the first of these two experiments is judged to be about an appropriate size, and the second is a bit small, although the blocking will have increased precision quite considerably, so the experiment is probably of an appropriate size. Where there is no ethical constraint and where experimental units are relatively inexpensive, as is the case with many *in vitro* experiments, the upper limit can be substantially increased.

References

1. MINITAB, Inc. 3081 Enterprise Drive, State College, PA 16801-3008, USA.
2. Thomas, L. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America* **78**, 126-139
3. Mead, R. (1988). *The Design of Experiments,* 620pp. Cambridge & New York: Cambridge University Press.

Appendix 4

Checklist for experimental design and statistical analysis of papers submitted to *ATLA*

1) Is/are the aim(s) of the experiment(s) clearly stated?

2) Have you indicated clearly how many separate experiments (not replications of the same experiment) are being reported, and have these been appropriately labelled "Experiment 1", "Experiment 2", etc.?

3) The following points refer to each individual experiment.
   a) The "experimental unit" is the entity (such as a culture dish) which can be assigned to a treatment. It is the unit of randomisation, and for the statistical analysis. Is the experimental unit in your experiment clearly indicated?
   b) Have you described the method of randomisation?
   c) State whether coded samples and "blinding" have been used where possible and appropriate.
   d) Have you indicated the type of experimental design used, such as completely randomised, randomised block, etc., and whether or not you have used a factorial treatment structure?
   e) If using animals or humans, have you indicated how an appropriate sample size was determined?
   f) Have you described and justified, in the *Materials and Methods* section, the statistical methods used? Uncommon statistical methods should be referenced.
   g) If you have used parametric statistical methods (for example, t test or analysis of variance [ANOVA]), have you determined that the assumptions of approximate normality of residuals and equality of variation are acceptable, given that these procedures can tolerate some departure from these assumptions?
   h) When comparing several means with an ANOVA, have you indicated the *post hoc* comparison or other methods you have used?
   i) When presenting means and proportions, have you indicated "n" and either the standard deviation (SD), standard error (SE) or confidence interval (CI), and have you chosen the most appropriate of these?
   j) When using non-parametric methods, have you indicated the medians, or other indication of location, and some measure of variation, such as the inter-quartile range?
   k) Where effects are statistically significant, have you also shown their magnitude and commented on their biological relevance?
   l) Have you quoted actual p-values, rather than using < signs?
   m) When effects are *not* statistically significant, have you assumed, incorrectly, that this means that the treatment(s) have no effect? Consider using a power analysis to indicate the magnitude of treatment effect that the experiment could probably have detected.
   n) When using correlation, have you graphed the data and considered possible effects of scale changes. If the graph is presented, have you shown lines giving both the regression of *X* on *Y* and of *Y* on *X*?
   o) Are all the diagrams necessary and informative? Simple bar diagrams might be better presented as a table giving numerical data. Also, have you considered using scatter diagrams showing individual points, rather than error bars?