REVIEW ARTICLE

Reduction of animal use: experimental design and quality of experiments

Michael F. W. Festing

MRC Toxicology Unit, Hodgkin Building, University of Leicester, Leicester LE1 9HN, UK

Summary

Poorly designed and analysed experiments can lead to a waste of scientific resources, and may even reach the wrong conclusions. Surveys of published papers by a number of authors have shown that many experiments are poorly analysed statistically, and one survey suggested that about a third of experiments may be unnecessarily large. Few toxicologists attempted to control variability using blocking or covariance analysis.

In this study experimental design and statistical methods in 3 papers published in toxicological journals were used as case studies and were examined in detail. The first used dogs to study the effects of ethanol on blood and hepatic parameters following chronic alcohol consumption in a 2×4 factorial experimental design. However, the authors used mongrel dogs of both sexes and different ages with a wide range of body weights without any attempt to control the variation. They had also attempted to analyse a factorial design using Student's *t*-test rather than the analysis of variance. Means of 2 blood parameters presented with one decimal place had apparently been rounded to the nearest 5 units. It is suggested that this experiment could equally well have been done in 3 blocks using 24 instead of 46 dogs. The second case study was an investigation of the response of 2 strains of mice to a toxic agent causing bladder injury. The first experiment involved 40 treatment combinations (2 strains $\times$ 4 doses $\times$ 5 days) with 3–6 mice per combination. There was no explanation of how the experiment involving approximately 180 mice had actually been done, but unequal subclass numbers suggest that the experiment may have been done on an *ad hoc* basis rather than being properly designed. It is suggested that the experiment could have been done as 2 blocks involving 80 instead of about 180 mice.

The third study again involved a factorial design with 4 dose levels of a compound and 2 sexes, with a total of 80 mice. Open field behaviour was examined. The author incorrectly used the *t*-test to analyse the data, and concluded that there was no dose effect, when a correct analysis showed this to be highly significant.

In all case studies the scientists presented means $\pm$ standard deviations or standard errors involving only the animals contributing to that mean, rather than the much better estimates that would be obtained with a pooled estimate of error. This is virtually a universal practice. While it is not in itself a serious error, it may lead scientists to design experiments with group sizes of at least 3 animals, which may result in an unnecessarily large experiment if there are many treatment combinations.

In conclusion, all 3 papers could have been substantially improved, with higher precision and the use of fewer animals if more attention had been paid to better experimental design.

Keywords Experimental design; statistics; laboratory animals; mice; dogs; reduction in animal use

Accepted 16 February 1994

Are animal experiments generally well designed? Such a question is rarely asked, yet it has important implications both ethically, and for the efficient use of scientific resources. A well designed experiment should be capable of answering the problems to which it is addressed with a high degree of precision, it should be an appropriate size, it should be unbiased, it should make efficient use of resources, and it should be correctly analysed so that no information is wasted. A key point in designing good experiments is to control the variability of the experimental material and all processes such as laboratory determinations that result in the final numerical measurements or counts. Success in this respect should lead to a reduction in animal use (see Russell & Burch 1959, chapter 6). Failure to design experiments correctly at best will result in efficient use of resources and a waste of animals. At worst it will lead to incorrect conclusions. Neither outcome is ethically desirable.

Relatively few attempts have been made to assess the quality of design of animal experiments, though there have been several papers which have looked at the statistical methods which have been used once an experiment has been completed. For example, Sterling (1971) found great difficulty in interpreting the results of toxicity experiments on 2,4,5-T because few of the papers used the correct statistical analyses. He found that in many cases Student's *t*-test was used when the more powerful analysis of variance would have been more appropriate. Similarly Benignus and Muller (1982) and Mitchell (1983) commented on the poor quality of statistical analysis of papers published in the neuro-toxicological sciences, again focusing on problems presented by the repeated use of Student's *t*-test and the resulting high level of false positive results. This led Muller *et al.* (1984) to prepare an excellent set of guidelines for appropriate statistical methods in toxicological experiments. Poor statistical methods are not confined to toxicology. Altman (1982) found that 'The general standard of statistics in medical journals is poor. . .' and concluded that the reasons for this are that in the majority of cases no statistician is involved in the study, and the statistical training of research workers is usually inadequate. He also suggested that the training of statisticians was not sufficiently practical and was usually too general to include many of the techniques specific to medical statistics. He suggested that many scientists would be only too pleased to get some expert assistance, but they have nobody to provide it.

Both the design and statistical analysis of experiments in papers published in 2 toxicology journals were examined by Festing (1992). It proved to be quite difficult to assess the quality of experimental design because most papers gave insufficient information on exactly how experiments were conducted, but it was concluded that there was ample scope for improving the design of animal experiments. More recently, Festing (submitted) attempted to assess whether animal experiments in toxicological research were the 'right' size. Seventy-eight papers published in 2 toxicological journals were surveyed. Thirty-three of these used animals, and reported the results of 48 experiments. Although it is difficult to decide how large an experiment should be, Mead (1988) suggested that for most experiments with quantitative end-points there should be about 10–20 degrees of freedom for estimating experimental error. The number of degrees of freedom for error is given by $DF_{error} = (N - 1) - (T - 1) - (B - 1)$, where N is the total number of observations, T is the number of treatments, and B is the number of blocks and/or covariates. For example, an experiment using 30 mice with 3 treatments, done in 2 blocks (i.e. with the experiment split into 2 identical halves in order to increase precision) would have $(30 - 1) - (3 - 1) - (2 - 1) = 26$ degrees of freedom for error, so would probably be

Based on presentation at the symposium 'Developments in laboratory animal science', held on 24 September 1993 to celebrate the tenth anniversary of the Department of Laboratory Animal Science, University of Utrecht, Netherlands

unecessarily large (see Beynen *et al.* 1993 for more details). On this basis, about a third of the experiments appeared to be unnecessarily large. There was also little evidence that research workers were attempting to control variability using blocking or covariance analysis, and although about 30% of experiments used a factorial arrangement of treatments, few of them analysed these correctly. Overall, only 13/48 experiments appeared to have been analysed correctly, with the incorrect use of Student's *t*-test being the main mistake.

The aim of this paper is to focus on 3 case studies where it appears as though lack of understanding of statistical methods and experimental design have reduced the effficacy of animal experiments. As the aim is to give constructive suggestions rather than destructive criticism, the examples given are anonymous and have been disguised, though copies of the papers have been given to the Editor and referees. All papers come from refereed toxicology journals published within the last 3 years. Case study 1 was the first animal experiment found when scanning a recent issue of a toxicological journal to which the author subscribes, and case study 2 was one of the papers in a single issue of a journal which had been passed to the author by a colleague with a note that it showed a strain difference in a toxic response. Case study 3 was from a single issue of the journal taken from the display rack in the library. In no case was an attempt made to pick particularly 'bad' examples. Whilst it is not possible to generalize from just 3 papers, many of the errors in design and statistical analysis are typical of those found in larger surveys (e.g. Festing 1992 and submitted).

Table 1 Number of animals per group in the first case study

Anaesthetic	No ethanol	Ethanol 30 days
Control	6	6
A	5	5
B	5	5
C	5	5

Case study 1

Aim of the experiment

The aim of this experiment was to study possible interactions between 3 non-barbiturate anaesthetics and ethanol 'In view of the fact that in some cases . . . a patient, either temporarily or chronically intoxicated with ethanol, has to undergo surgical treatment. . . .'

Description of the experiment

The experiment involved a total of 42 '. . . mongrel dogs of both sexes and [of] different ages, weighing 5.5 to 11 kg.' Half the dogs were maintained with 12% ethanol instead of water for 30 days before the anaesthetic treatments were carried out. The 'control' groups of 6 dogs per ethanol treatment consisted of 2 animals anaesthetized with anaesthetics A, B, and C, respectively, with the blood and liver biopsy being collected immediately. In the other groups a single anaesthetic was used and the animals were kept in '. . . a deep

Table 2 Part of a table from case study 1 showing the effect of anaesthetics on blood parameters of dogs not treated with ethanol (mean ± standard deviation)

Blood parameter	Control	Anaesthetic A	Anaesthetic B	Anaesthetic C
AST (units	12.0±4.12	10.0±2.82	15.0±6.53	10.0±5.21
ALT (units)	15.0±5.23	20.0±7.55	10.0±3.54	20.0±5.50
ALP (units)	6.7±3.27	9.0±1.55[b]	3.5±0.94	5.5±0.95

[a]$P<0.05$; [b]$P<0.02$; [c]$P<0.01$; [d]$P<0.001$
Note that '*P*' values of 0.05, 0.02, 0.01 and 0.001 are indicated in the footnote though in the original table only a single number (9.0±1.55[b]) has any indication that it is different from the control. This is confusing as it suggests that some superscripts may have been omitted by mistake. See text for further comments

anaesthetic state (clinical signs) for 3 h. . .' before samples were taken. The results were assessed using 5 blood and 5 liver biochemical parameters. The experimental treatments and number of animals were as shown in Table 1.

The authors state that 'Results were expressed as mean values ± s.d. The significance of (any) difference was tested by Student's *t*-test.' The results are presented in 5 tables. Table 2 shows part of a summary table in which some blood parameters were compared among treatments. The authors overall conclusion was that 'The above results indicate that the changes observed are probably due to the action of ethanol, and not to the action of anaesthetics. However, the possibility of synergism cannot be excluded.'

Assuming that this is a worthwhile animal experiment, it can be seriously criticized on a number of grounds:

Critique of case study 1

Highly heterogeneous experimental units The precision of an experiment depends critically on the size of the experiment and the homogeneity of the experimental material. Even quite a small reduction in the within-group standard deviation can lead to a dramatic increase in precision. These dogs were mongrels of both sexes and of very uneven weights and ages, yet they were apparently allocated at random to the experimental treatments. It should have been possible to reduce the size of the experiment or obtain more precise results by choosing a more homogeneous group of animals. If this was not possible, then some of the heterogeneity might have been removed using a randomized block experimental design.

Because of the use of such heterogeneous material, this was almost certainly a low precision experiment. Experiments with low precision will often fail to pick up biologically important treatment effects, as was possibly the case with this experiment in that the authors felt that interactions could not be ruled out.

Choice of treatments The 'control' groups consisted of 6 animals which were anaesthetized (2 by each anaesthetic), and then the blood and liver samples were taken immediately. If the 3 anaesthetics differed in their influence on any of the end-points, then the within-group variation would have been increased, and the precision would be reduced. It would probably have been better to do a pilot experiment to compare the acute effects of the 3 anaesthetics. If they did not differ, then any one of them could have been used.

Incorrect statistical analysis This experiment involved a quite complex 'factorial' experimental design, i.e. there were 2 factors: the 4 anaesthetic treatments (including the controls) and the ethanol treatment. It is incorrect and confusing to attempt to analyse such an experiment using Student's *t*-test, which is only appropriate for comparing 2 experimental treatments rather than the 8 treatment combinations found in this experiment. Use of the *t*-test in such circumstances can lead to increased numbers of false positive results and to false negative results. Moreover, there may be a tendency for research workers not familiar with the analysis of variance to design experiments which can be analysed by the *t*-test. This would imply group sizes of about 5 individuals in order to get sufficient error degrees of freedom. With the analysis of variance, which uses a pooled estimate of the standard deviation, smaller group sizes are possible (see below).

Table 3 Analysis of Variance (approximate[1]) of a blood (ALP) and liver (GSH) parameter to show the correct layout

Source	DF	MS (ALP)	MS (GSH)
Ethanol	1	384.4**	35.15
Anaesthetic	3	54.35**	435.70**
A×E	3	14.64**	221.00**
Error	32	2.03	19.36
Total	39		

[1]The exact analysis cannot be done without access to the raw data due to the extra number of animals in the control group

A more acceptable way to analyse these data would have been to use the analysis of variance (ANOVA). An example of such an ANOVA for 2 of the characters (serum alkaline phosphatase (ALP) and liver glutathione (GSH)) is given in Table 3, reconstructed approximately from data presented by the authors. The analysis of ALP suggests a highly significant effect due to ethanol, and anaesthetics, and a highly significant anaesthetic × ethanol interaction. The latter implies that the effect of the 3 anaesthetics differed depending on the ethanol treatment. With respect to GSH the authors concluded that 'treatment with ethanol caused a very slight increase in GSH content. All 3 anaesthetics caused a decrease in the GSH content in ethanol-treated dogs, but this decrease was not statistically significant.' In fact, the use of a pooled standard deviation (from the ANOVA) and a suitable range test shows that the increase in GSH due to ethanol was not statistically significant, but that all 3 anaesthetics caused a statistically significant ($P<0.05$) depletion of liver GSH.

In this paper multiple end-points have been measured, and a strong case could be made for a multivariate analysis of the data using a MANOVA (multivariate analysis of variance) or a method such as principal components analysis (PCA, see for example, Festing *et al.* 1984). These may help to explore the pattern of response to the experimental treatments, and show up any of the variables which behave differently. For example, in this study, the main response to both alcohol and the anaesthetics was for the serum and liver enzymes to increase. However, this was not true of GSH, which decreased. This pattern is obvious when PCA analysis is used. However, a multivariate analysis of this sort is rarely done, and would certainly need to be carried out in collaboration with a statistician, as such techniques can be misleading if used by inexperienced people.

Poor presentation of the results Presentation of results is not usually regarded as an important part of experimental design, yet it can have an impact on the way people design experiments. There are 2 objections to the results presented in this paper, one relatively trivial (means to be compared should be in columns), the other with substantial implications for experimental design (a pooled estimate of the standard deviation should be used).

Means to be compared should be in columns Results really should be presented in such a way that they can be easily understood. Table 2 shows a small portion of one of the 5 tables used to present the results in this paper. A reader would be interested in comparing the different groups for a particular blood or hepatic parameter. The human eye finds it easiest to compare figures which are in a column. In this paper the data for each parameter are presented in rows, and in order to compare the ethanol and no-ethanol groups it is necessary to compare different tables. Fortunately, the authors provide a table in which all the tabular data is given again in a single table, though still with the treatments as column headings. Obviously, the editor was not short of space in allowing exactly the same numerical data to be presented twice.

Table 4 Re-drawn table showing effects of anaesthetics on blood parameters of untreated and ethanol-treated dogs (again, only part of the data is shown)

Treatment		Blood parameter, μcat dm^{-3} serum		
Anaesthetic	Ethanol	AST	ALT	ALP
A	N	10	20	9.0
B	N	15	10	3.5
C	N	10	20	5.5
A	Y	25	40	14.3
B	Y	30	37	8.4
C	Y	25	40	14.8
Pooled SD (24 DF)		6.02	8.52	1.43

Note that the means for AST and ALT in the original paper were presented with one decimal place, but this was always zero, and it appears as though all except one have been rounded to the nearest 5 units

Pooled standard deviations should be used Another feature of the tables, which is virtually universal in the toxicology literature, is that each mean is presented ± the standard deviation (SD) or standard error (SE) using the individual values that contribute to the mean. This is distracting, but there is a more serious objection. Each SD is based on only n-1 degrees of freedom, where n is the number per group. In order to get good estimates, research workers will tend to use relatively large group sizes, which will tend to inflate the size of the whole experiment. Yet, in using a t-test or an ANOVA it is assumed that the within-group standard deviations for each treatment mean are the same. It follows that a better estimate of the SD for each group would be a pooled SD from all groups. The main exception would be if standard deviations really do differ between groups, which is not unusual. It is not uncommon for mean and SD to be correlated. In those circumstances it is usually possible to use a transformation of the data to eliminate such 'heterogeneity' of the variance. If a pooled SD is used, group sizes can be smaller and means can be presented more clearly. Table 4 shows some of the data displayed so as to show treatment means more clearly. In this case a pooled SD based on 24 degrees of freedom has been given for each column. This is a much better estimate of the true SD than the individual values based on only 4 degrees of freedom. Mead (1988) noted the '. . . very large proportion of experiments, perhaps as high as 85%, which take the form of randomized complete block designs.' With such designs only a pooled estimate of error is available. If these designs were more widely used in toxicology (they are rare), such mindless presentation of means ± SDs would be avoided.

In this case, there also appear to be some arithmetical errors in the table. Re-calculation of the comparison of control and anaesthetic A for ALP gives a t-value of 1.53 with 9 degrees of freedom, which does not exceed the critical value for $P = 0.05$, so this difference is not in fact statitistically significant. A point that was not clear in the original is that although means for AST and ALT were presented with one decimal place, this was always zero, and apart from a single value all appear to have been rounded to the nearest 5 units. Table 4 shows that ethanol about doubled all the serum enzyme levels. Anaesthetic B also had the lowest enzyme levels without ethanol, but had average levels of AST and ALT with ethanol. Whether these differences were statistically significant would have been clear if an ANOVA had been conducted for each blood and liver parameter.

A better design

This experiment involved a total of 42 dogs, which is quite a large experiment for this species. It is possible (though impossible to confirm) that a randomized block design in which each block consisted of homogeneous animals could have achieved the same precision with fewer animals. For example, 3 blocks each consisting of 8 animals all as similar as possible could have been used. Thus, block 1 might have consisted of 8 large males, block 2 of 8 small males, and block 3 of 8 average females. A single block and the analysis of variance table would then have looked like the ones shown in Tables 5 and 6. This experiment would involve only 24 dogs. Other designs might have been possible. If these were not terminal experiments, then it might have been possible to use the same animals to study the effects of more than one of the treatments, with suitable recovery times between them. This would have been a 'crossover' design which might have been

Table 5 Suggested layout of a single block. This would be repeated on 3 occasions

	Ethanol	
Anaesthetic	No	Yes
Control	1	1
A	1	1
B	1	1
C	1	1

Block 1: 8 dogs as uniform as possible

Table 6 Skeletal layout of the ANOVA for each character if the experiment had been designed as a complete block design with 3 blocks

Source	DF	SS	MS
Blocks	2		
Ethanol (E)	1		
Anaesthetic (A)	3		
E×A	3		
Error	14		S^2
Total	23		

much more precise than the randomized block design, and would have used far fewer animals. Ideally, a statistician should have been part of the research team which designed the experiment.

Case study 2

Aim of the experiments

The aim of this paper was to study a strain difference in the response of mice to a compound causing bladder toxicity. Two mouse strains were used, and the strain names (coded here as strains X and Y) and their source was given. The only other details of the animals were that 'The animals were maintained on a 12/12-light dark cycle and were provided with food and water *ad libitum*.'

Description of the experiments

The paper is typical of many toxicity studies in that it presents data from a series of experiments, but it is often extremely difficult to discover exactly how many experiments were conducted and how many animals were used. Very often results are presented, but it is not clear whether they represent data from a new experiment. However, it appears as though there was a total of 6 different experiments involving about 340 mice in total. Only the first experiment is discussed here.

In this experiment the compound was administered to animals as a single i.p. injection at 4 dose levels (including the vehicle dosed controls), and damage was allowed to progress for 1–5 days. Bladder damage was assessed by measuring the blood content of the bladders. Apparently the compound was given at all 3 doses to both strains of mice, but the results were only shown graphically and data are only shown for the sensitive strain at all 3 doses and the resistant strain at the highest dose level. The only indication of the number of mice used is a footnote to the results figure stating $n = 3–6$. With mice apparently killed on 5 consecutive days, 2 strains and 4 doses including controls and assuming an average of 4.5 mice per group, this is a $5 \times 4 \times 2$ factorial design involving a total of 180 mice.

The statistical analysis section stated that 'The data are expressed as the mean ± SE and were analysed using the unpaired Student's *t*-test, where appropriate, or a one-way analysis of variance. Post hoc analyses were carried out using the Student-Newman-Keuls test. A P value of <0.05 was considered significant.'

Critique of case study 2

Experiments inadequately described The immediate problem with this and many similar papers is to discover exactly what the research workers did, and why. Papers should clearly state how many experiments were done, and the objectives in each case, and ideally these should be labelled Experiment 1, Experiment 2 etc.

Why were there unequal subclass numbers? The only indication of the number of animals used is a footnote, and the number of animals per group varied. Why should the number per group vary when there were no unplanned deaths? This is not clear, but one possible explanation is that the studies were not carried out as planned experiments with animals first obtained in the required numbers, then acclimatized, and finally assigned to treatment groups at random. Clearly, an experiment involving about 180 mice whch have to be killed on 5 different days, the bladders excised, and various determinations made cannot be done all at once, but the authors do not state how the

experiment was broken down. The easiest way would be to start with one strain and carry out the experiment as and when animals became available. If this was done then any experimental errors associated with the killing of the mice and the subsequent laboratory determinations could become confounded with the treatments.

Inadequate statistical methods The statistical methods used in this paper are somewhat better than those used in the previous case study, but are still inadequate. A $5 \times 4 \times 2$ factorial experiment must be analysed by a 3-way analysis of variance. The repeated use of a one-way ANOVA is no more acceptable than the repeated use of the *t*-test. Unfortunately, it has been difficult to analyse a 2- or 3-way ANOVA with unequal subclass numbers until recently when suitable statistical packages for PCs have become available. Possibly the authors did not have access to such a package. All means are given ± SE (standard error of the mean), presumably calculated separately for each mean. However, as the '*n*' varies from group to group, these standard errors are completely meaningless. Without knowing '*n*' it is impossible to use them as a measure of the precision of each mean, or to compare the significance of differences between means. Again, assuming no heterogeneity of variance, a pooled estimate of experimental error should be used, and where *n* varies for each mean it should be given.

A better design

This experiment was excessively large. The total of 40 different treatments (2 strains × 4 doses × 5 days) is difficult to manage while still keeping the material homogeneous. The experiment could probably have been done as 2 blocks with one mouse of each treatment group per block. This would involve taking 20 mice of each strain on day 1 and injecting 5 of each of them with each dose level. On day 1 one mouse of each dose level of each strain would be killed (a total of 8 mice), similarly on day 2 etc. The whole experiment would then be repeated. Presumably it would be practical to handle 8 mice per day. This experiment would involve a total of 80 mice altogether. Although this is an improvement on the 180 or so actually used, it is still a large experiment with 39 degrees of freedom for error. It could not easily be reduced further without reducing the number of treatments (say killing only on days 1, 3 and 5) or using an incomplete block or fractional factorial design, which may involve a level of complexity which would require the services of a professional statistician.

In conclusion, there is a good chance that a properly conducted and tightly controlled experiment involving 80 mice would provide more accurate and reliable information than the original experiment which involved about 180 mice.

Case study 3

Aim and description of the experiments

The third case study involved the effects of a compound (PB) which was not strongly toxic on reproductive and behavioural parameters in mice. It was fed to the mice at 3 dose levels plus a control. Only the first experiment will be discussed here. The mice were tested for several behavioural parameters in an open field for a period of 3 min. A total of 10 mice were used per group, with both sexes being involved, giving a grand total of 4 treatments × 2 sexes × 10 mice per group = 80 mice total. Data were presented as means ± SE of each group, and the data were analysed using Student's *t*-test. The means of each group presented with a

Table 7 Groups means for case study 3 re-presented to show the effects of PB on open-field ambulation, using a pooled standard deviation (*n* = 10/group. Pooled SD = 68.3)

PB Dose	Male	Female	Mean
0	162.5	226.3	194.4
0.15	134.3	147.1*	140.7
0.3	114.5	178.3	146.4
0.6	95.8**	145.6	120.7

*$P<0.05$, **$P<0.01$

Table 8 Analysis of variance of PB ambulation data with the dose effect broken down to show linear, quadratic and cubic contrasts

Source	10 mice/gp, 80 total DF	MSq	3 mice/gp, 24 total DF	MSq
Sex	1	45 220**	1	45 220**
Dose	3	19 520**	3	19 520*
Linear	1	46 397**	1	46 397**
Quadratic	1	3 920	1	3 920
Cubic	1	8 244	1	8 244
Sex x dose	3	2 901	3	2 901
Error	72	4 661	17	4 661
Total	79		23	

*$P<0.05$, **$P<0.01$
Note that virtually the same analysis would be obtained with only 3 mice per group if the means were the same

pooled standard deviation are given in Table 7. The authors concluded that 'There were a few significant differences in open field activity in mice administered PB, *but no consistent significant compound or dose related effect*. . . . However, in male mice the decrease in ambulation and rearing tended to be dose related.' (My italics.)

Critique of case study 3

A re-analysis of the authors' data using the analysis of variance is given in Table 8. This shows clearly that there was a statistically significant sex difference, with the females being more active than the males. There was also a highly significant dose effect, which could be attributed to a linear trend towards lower ambulation. There was no sex difference in the response to the toxic compound. If the means were the same, then all these effects could have been detected using only three mice per group (24 mice total) instead of the 80 mice actually used. In other words, the authors used an excessive number of mice and failed through the use of incorrect statistical methods to detect a highly significant treatment effect.

As behaviour is a relatively unstable end point, this experiment might have been better designed as a randomized block in 3 blocks of 8 mice (male and female each at the 4 dose levels). The 8 mice would all be tested on the same day, in random order. As there would be fewer mice, it should be possible to increase the length of the open field test from 3 min to, say 5 min as a way of increasing precision. In fact, the same mice might be tested more than once.

Discussion

Basically, there are 2 approaches to discovering whether or not there is scope for a reduction of animal use by means of better experimental design. Surveys of published papers can be used to assess the size of experiments, the type of treatment structure (i.e. single factor or factorial designs), and the use of methods of reducing variation by experimental design and statistical analysis such as randomized blocks and covariance analysis. Failure to use methods of experimental design, particularly the use of blocking to increase precision and factorial designs to increase the range of applicability, which have been well recognized for more than 50 years implies that there is scope for improvement. Such surveys can also assess whether or not most experiments appear to have been correctly analysed. For example, the use of Student's *t*-test is clearly inappropriate when there are more than two treatments. Surveys of this type were carried out by Festing (1993, submitted), with the suggestion that there is considerable scope for improving experimental design. However, such an approach inevitably deals in generalities.

The second approach, used here, is to study a few published papers in detail. The problem in this case is that only small samples of papers can be studied, and they may not be representative of papers published in the journal or discipline. Therefore, this approach can really only be used in association with surveys. Thus, although it cannot be claimed that these papers are in any way typical of other papers published in the same journals, it is safe to conclude that many of the detailed faults found in the three papers examined here were of the type which is common. For example, failure to mention any way of controlling the variation such as the use of complete block designs or covariance

analysis is in agreement with previous surveys which show that the use of such designs is uncommon. All 3 experiments involved a factorial arrangement of treatments, which is not uncommon according to the surveys, but none of them used the correct methods of statistical analysis. This is also in agreement with previous surveys. The presentation of means ± standard deviations or standard errors estimated from the units contributing to that mean rather than the use of a pooled estimate of error is almost universal, and yet in many cases it does not make much sense. Two of the experiments appeared to involve unnecessarily large numbers of animals, and all three could probably have been conducted using much smaller numbers, and this is also in general agreement with previous surveys (Festing 1992). Thus, in a way, it is legitimate to consider these 3 experiments typical of many such experiments described in the toxicological literature. If so, then it can be concluded that both surveys of several published papers as well as detailed study of individual papers suggest that there is ample scope for reducing the use of animals and improving scientific productivity by better experimental design.

References

Altman DG (1982) Statistics in medical journals. *Statistics in Medicine* **1**, 59–71

Benignus VA, Muller KE (1982) Bad statistics are worse than none (letter). *Neurotoxicology* **3**, 153–4

Beynen AC, Festing MFW, van Montfort MAJ (1993) Design of animal experiments. In: *Principles of Laboratory Animal Science* (van Zutphen LFM, Baumans V, Beynen AC, eds). Amsterdam, London: Elsevier, 209–40

Festing MFW (1992) The scope for improving the design of laboratory animal experiments. *Laboratory Animals* **26**, 256–67

Festing MFW (submitted) Are animal experiments in toxicological research the "right" size? In: *Statistics in Toxicology* (Morgan BJT, ed.). Oxford: Oxford University Press

Festing MFW, Hawkey CM, Hart MG, Turton JA, Gwynne J, Hicks RM (1984) Principal components analysis of haematological data from F344 rats with bladder cancer fed the retinoid N-(ethyl)-all-trans-retinamide. *Food and Chemicals Toxicology* **22**, 559–72

Mead R (1988) *The Design of Experiments*. Cambridge, New York: Cambridge University Press

Mitchell CL (1983) "Bad statistics" revisited (letter). *Neurotoxicology* **4**, 157–9

Muller KE, Barton CN, Benignus VA (1984) Recommendations for appropriate statistical practice in toxicological experiments. *Neurotoxicology* **5**, 113–26

Russell WMS, Burch RL (1959) *The Principles Of Humane Experimental Technique*. London: Methuen and Co. Ltd (reprinted by the Universities Federation for Animal Welfare, 8 Hamilton Close, South Mimms, Potters Bar, Herts EN6 3QD)

Sterling TD (1971) Difficulty of evaluating the toxicity and teratogenicity of 2,4,5-T from existing animal experiments. *Science* **174**, 1358–9