Use of Factorial Designs to Optimize Animal Experiments and Reduce Animal Use

Robert Shaw, Michael F. W. Festing, Ian Peers, and Larry Furlong

Abstract

Optimization of experiments, such as those used in drug discovery, can lead to useful savings of scientific resources. Factors such as sex, strain, and age of the animals and protocol-specific factors such as timing and methods of administering treatments can have an important influence on the response of animals to experimental treatments. Factorial experimental designs can be used to explore which factors and what levels of these factors will maximize the difference between a vehicle control and a known positive control treatment. This information can then be used to design more efficient experiments, either by reducing the numbers of animals used or by increasing the sensitivity so that smaller biological effects can be detected. A factorial experimental design approach is more effective and efficient than the older approach of varying one factor at a time. Two examples of real factorial experiments reveal how using this approach can potentially lead to a reduction in animal use and savings in financial and scientific resources without loss of scientific validity.

Key Words: animal testing alternatives; animal use alternatives; case report; factorial analysis; models, animal; research design; statistical methods

Introduction

There is considerable scope for reducing the number of animals and scientific resources used in research by designing better experiments (Festing 1994, 1995a,b; Festing and Lovell 1996). Some experiments are performed repeatedly with only minor variations, and even very small improvements in the design can lead to substantial savings of animals over a period of time. Animal experiments form a necessary part of the late stages of the drug discovery process—An animal model may be used to screen large numbers of compounds with only the identity of the compounds changing between experiments. A typical experiment, which may involve three or four groups of approximately eight animals treated with different candidate compounds and a larger control group, may have the aim of finding the compounds that have a potentially useful effect. Batches of vaccines and other biologicals are often tested in animals using a standard protocol, with the aim of measuring the biological activity or toxicity of the batch. Even in basic research, some procedures (e.g., the preparation of cDNA) use complex methods that may be used repeatedly even though individual experiments may vary. If all of these experiments and associated techniques were optimized to use the smallest number of animals consistent with detecting a given response, there would be a substantial reduction in animal use and important savings in scientific resources.

One method of optimizing such experiments is to use factorial experimental designs (FEDs[1]) to discover which factors influence the outcome of the experiment and what levels of these factors lead to an experiment with the greatest sensitivity. The aim is usually to maximize the signal/noise ratio so that the numbers of experimental subjects required to detect a given treatment response (or "signal") is minimized by using power and sample size calculations. The procedure involves using a vehicle control and a known positive control treatment and attempting to maximize the mean difference, herein designated *treatment effect*. In this article, variables investigated for their influence on the treatment effect are termed *factors.* It is also useful to know which factors are relatively unimportant in influencing response so that less attention is given to controlling them.

The factors to be studied can be any variables the investigator can control, including direct animal-related characteristics (e.g., sex, strain, age, and dietary and health status) and aspects of the environment (e.g., cage and group size, bedding material, and environmental complexity). There are also many protocol-specific factors (e.g., methods of preparing the animal model; dose level; timing, route, and method of administration of test compounds; and methods and timing of observations). When complex protocols are involved in making the final observations (e.g., in the preparation and hybridization of cDNA in microarray experiments), then many factors that affect the variability of

The following authors are in Global Enabling Science and Discovery, Research and Development, AstraZeneca Ltd., in Macclesfield, UK: Robert Shaw, M.Sc. CStat., is Statistics Team Leader; Ian Peers, Ph.D., is Statistics Team Leader; and Larry Furlong, Ph.D., CStat., is Director of Statistical Science. Michael F. W. Festing, M.Sc., Ph.D., D.Sc., CStat., BIBiol., is a Senior Research Scientist at the MRC Toxicology Unit, University of Leicester, UK.

[1] Abbreviations used in this article: ANOVA, analysis of variance; AUC, area under the curve; FED, factorial experimental design; OVAT, one variable at a time.

the measurements may also need to be investigated. Sometimes animals are used as sources of tissue or cells in in vitro experiments, and the factors that influence the outcome of these experiments can affect the numbers of animals that are needed. Often this type of experiment will initially involve a 2^k factorial in which k factors are studied, each set at two levels.

Factorial designs using many factors (often of the 2^k series) have been widely used in the manufacturing industry as a means of maximizing output for a given input of resources (Cox 1958; Montgomery 1997). The strategy is to use the factorial design to identify the most important factors and levels of the factors that determine output and then to use these factors in normal production. A similar approach has been used in optimizing output from biological systems. For example, six factorial designs were used to study the effects of medium composition, incubation conditions, and associated microflora on the production of type G *Clostridium botulinum* toxin (Calleri et al. 1992) in vitro. A fractional factorial design was used to optimize enzyme-linked immunosorbent assay tests (Reiken et al. 1994), and a 2^4 factorial was used to optimize the conditions for freezing rat liver slices (Maas et al. 2000). Similar methods have been used to optimize the signal in DNA microarray experiments (Wildsmith et al. 2001). In vivo applications have been more rare but have been used, for example, in studying the effect of genotype, diet, and exercise in the accumulation of body fat in rats (Metzger et al. 2000); the effects of strain and dose levels of chloramphenicol on mouse haematology (Festing et al. 2001); and the effects of carcinogenic mixtures on the development of lung tumors in mice (Nesnow et al. 1998). However, in some of these cases, FEDs have been used to gain an understanding of the factors influencing the observed response, rather than to optimize future experiments.

In this article, we describe the use of FEDs to determine which factors influence the outcome of an experiment and the optimum levels of those factors so that future experiments can be designed to have the greatest possible sensitivity. This use of factorial design is an integral part of the development of a good animal model.

Factorial Experimental Designs

Table 1 The simplest possible factorial design: factor A is a control versus a treatment, and factor B is the sex of the animals

	Factor B	
Factor A	Male	Female
Vehicle control	a	c
Treated	b	d

[a] It is assumed that there are *n* animals in each group.

The theory and practical applications of factorial designs have been described in detail in a number of textbooks (Clarke and Kempson 1997; Cox 1958; Mead 1988; Montgomery 1997), so they will be reviewed here only briefly, in a nonmathematical way. In factorial experiments, more than one type of independent variable is varied at a time, but in a structured way. The simplest factorial is a 2*2 design (Table 1). Factor A could be a treatment such as a vehicle control versus a test substance, and factor B could be males versus females (or strain 1 vs. strain 2 or any other factor thought to be relevant). In such a case, there would be four groups: control-male, test-male, control-female, and test-female. With a number of animals per group, there would be four means—one for each group (e.g., a, b, c, and d), respectively. The mean for all animals given vehicle control is (a+c)/2; the mean for the treated animals is (b+d)/2; so the estimate for the effect of drug treatment is ((a+c)-(b+d))/2. It is important to note that *all* of the animals contribute information on this difference, which is known as the *main effect* of the drug. Similarly, all of the animals will have contributed to the main effect of sex, estimated as ((a+b)-(c+d))/2. This multiple use of data is a key benefit of using factorial designs to maximize information on limited numbers of animals.

If the response of the two sexes to the drug treatment is different, then there is said to be an interaction between factors A and B, estimated as ((a-c)-(b-d))/2. These interaction effects (commonly seen in biomedical sciences as synergism, enhancement, or potentiation) between the treatments (factor A in this case) and the other factors tested in the experiment are of particular interest because the aim is to maximize the treatment effect.

This principle generalizes to any number of factors and levels of each factor. For example, an experiment using rats may involve four dose levels of a drug X (one of which may be a zero dose level vehicle control) and both males and females. It would be a 4 × 2 factorial experiment, and the main aim might be to determine whether the dose-response relation is the same in both sexes. When the aim is to screen the possible impact of many (k) different factors on response or output, it is common to use a 2^k factorial design with each factor at two levels (e.g., male/female or treatment/control). Such designs can be used as a rapid screen to identify factors that are of most importance in determining response, and they can be designed to be economical with the use of animals.

In this discussion, we assume that the response is a quantitative rather than a categorical variable so that the experiment can be analyzed using analysis of variance (ANOVA[1]). (A quantitative variable is one that can be measured on a continuous scale [e.g., body weight], whereas a categorical variable is one that takes a limited number of discrete values [e.g., response/no response or large/medium/small].) The two assumptions of normality of the residuals (i.e., deviation of each observation from its group mean) and homogeneity of variances are normally examined as part of the statistical analysis.

All dedicated statistical packages should support the multiway ANOVA used to analyze factorial designs, although elementary packages may not allow unequal numbers in each group. Some packages also provide ready-randomized plans and analyses for a range of different factorial designs, with up to 15 factors, as well as Plackett-Burman designs (not considered here), with up to 47 factors. Other, more specialized experimental design software is available. Note that these designs are widely used in the manufacturing industry, where the factors are often settings on a machine. Observations that would normally be associated with the treatments of an individual animal here are frequently referred to as a "run" in industrial research.

Comparison with the One-Variable-at-a-Time (OVAT[1]) Approach

In contrast to FEDs, the most common historical approach is to vary each factor of interest in turn, keeping all other factors, which may influence the outcome, at a fixed level. This approach, commonly known as OVAT, has certain disadvantages, which include the following:

1. Each group of animals will contribute to understanding the effect of only a single factor, in contrast to FED, in which each animal will contribute to understanding the effect of all of the factors under exploration.
2. The independent investigation of each factor inherent in the OVAT approach overlooks the possible ways in which the effect of one factor can depend on the level of another (i.e., two or more factors may interact with each other).
3. Implementation of FED designs contributes to more efficient use of resources. For example, detailed consideration of all potential factors of interest at the study outset avoids incremental changes to multiple studies over time.

Thus, according to Cox (1958), "... factorial experiments have, compared with the one factor at a time approach, the advantages of giving greater precision for estimating overall factor effects, of enabling interactions between different factors to be explored, and of allowing the range of validity of the conclusions to be extended by the insertion of additional factors." These points imply that in comparison to OVAT, a FED approach will lead to better information with the use of fewer animals.

Full Factorial and Fractional Factorial Designs

The number of factors that can be studied will, in practice, range from two to approximately 15 or more, although in whole animal experiments it is unlikely that more than about 10 factors will be tested in a single experiment. When a small number of factors (perhaps up to 4) are considered at two levels each, *full factorial* designs can be applied. In these designs, a group of animals for every combination of each factor is included, as given in the simple examples above. With two factors, there are four animal groups to be considered (assuming 2 levels/factor); with four factors, there are 16; and with seven factors, there are 128 groups. For a larger number of factors, the total number of possible combinations becomes very large. In these situations, *fractional factorial* designs can be used.

A fractional factorial design provides a balanced subset of these groups while maximizing information on factors explored in the study. An example is a design with four factors, each at two levels (called a 2^4 design). There are 16 treatment combinations in all; however, it is possible to carry out a half fraction of the full design, with only eight treatment combinations (with appropriate replication), and still be able to estimate the effect of each factor and the interaction between any pair of factors. A degree of replication for each group is typically applied for two reasons: (1) to avoid losing information on an individual treatment combination through death/humane withdrawal of an animal; and (2) in many cases, it may be impractical due to protocol logistics and housing to include too many different groups. It should be noted that with fractional designs, some of the interactions may no longer be cleanly estimated and may be difficult to interpret. We recommend obtaining expert statistical advice when considering such a design.

A brief example of a fractional design layout is provided in Table 2. Eight factors were identified from a brainstorming session to be explored within an experimental design. A full factorial design would have consisted of $2^8 = 256$ groups. We chose a fractional factorial, which comprises 16 groups representing only 1/16 of the full design. This type of layout is obtainable from many statistical software packages. The design is structured to maximize information on the main effects and low-order interactions of the factors and treatment while sacrificing information on the high-order interactions, which are presumed to be negligible. In summary, fractional factorial designs provide a very powerful approach for reducing the total number of animal groups.

Interactions Involving Treatment and Other Factors

As noted above, one criterion often used for optimizing in vivo screens is to maximize the treatment difference. Any interaction between a factor in the experiment and the treatment may imply scope for improving the sensitivity of the experiment (the signal/noise ratio). For example, in Table 3, the coded white blood cell counts of two strains of mice administered vehicle or chloramphenicol (2500 mg/kg) by gavage are shown (Festing et al. 2001). Data were trans-

Table 2 Example of a fractionated design

Compound	Dose (mg/kg)	Sex	Age	Nutritional state	Time after dosing (hr)	Glucose load (g/kg)	Time after glucose load (min)
A	10	M	10	Fed	4	2	5
B	10	M	10	Fed	6	4	10
A	20	M	10	Unfed	4	4	10
B	20	M	10	Unfed	6	2	5
A	10	F	10	Unfed	6	4	5
B	10	F	10	Unfed	4	2	10
A	20	F	10	Fed	6	2	10
B	20	F	10	Fed	4	4	5
A	10	M	20	Unfed	6	2	10
B	10	M	20	Unfed	4	4	5
A	20	M	20	Fed	6	4	5
B	20	M	20	Fed	4	2	10
A	10	F	20	Fed	4	4	10
B	10	F	20	Fed	6	2	5
A	20	F	20	Unfed	4	2	5
B	20	F	20	Unfed	6	4	10

formed to $\log_{10}(X+1)$, where X is the individual observation to normalize the residuals (deviations of each observation from its group mean), and were analyzed by ANOVA. Residual plots were used to ensure that the assumptions of normality of residuals and homogeneity of variances were satisfied (not shown, but see Example 2 and most modern statistics textbooks). A highly significant interaction ($F_{1,28} = 8.46, p = 0.007$) between the chloramphenicol treatment and the mouse strain reveals that the two strains differed in sensitivity. The least squares means (which account for unequal numbers in each group) shown in Table 3 are given in standard deviation units after the division of each mean by the pooled standard deviation. The use of standard deviation units here makes it easier to estimate sample sizes (see below) and to compare different experiments.

The response was substantially greater in strain CBA than in CD-1, so that if the aim had been to screen compounds to see whether they reduce white blood cell counts in the same way as chloramphenicol (an unwanted toxic side effect in this case), then much smaller sample sizes would have been needed to detect a specified effect using sensitive CBA rather than insensitive CD-1 mice. Indeed, the response to chloramphenicol in the CD-1 mice is so small (and not even statistically significant) that it is doubtful whether this strain could even be used in such a screen. Although the use of CBA mice could result in smaller sample sizes, an alternative would be to use the same number of animals but to have a more powerful experiment capable of detecting a smaller biological response.

Table 3 White blood cell response (in standard deviation units) of two strains of mice given chloramphenicol at 2500 mg/kg[a]

Strain	Vehicle	Chloramphenicol	Difference
CD-1	4.67	4.23	0.44
CBA	4.03	1.51	2.52

[a] Data abstracted from the larger study of Festing MFW, Diamanti P, Turton JA. 2001. Strain differences in haematological response to chloramphenicol succinate in mice: Implications for toxicological research. Food Chem Toxicol 39:375-383.

Effect sizes as described in the Table 1 text of Festing et al. 2001 as follows:

Main effect of treatment: {(4.67 + 4.03) − (4.23 + 1.51)}/2 = 1.48

Main effect of strain: {(4.67 + 4.23) − (4.03 + 1.51)}/2 = 1.68

Interaction effect: {(4.67 − 4.23) − (4.03 − 1.51)}/2 = (0.44 − 2.52)/2 = −1.04

Note that the interaction effect is the averaged *difference* in the two responses to chloramphenicol across the two strains, [(0.44 − 2.52)/2] = −1.04.

Effect Size and Sample size

Two main strategies for determining sample size, the *resource equation* method and the *power analysis* method, have been described (Festing et al. 2002). The resource equation method depends on the law of diminishing returns and is based on the suggestion of Mead (1988), that there should be 10 to 20 degrees of freedom for the error term in the ANOVA. Although this guide is useful in some circumstances, it does not account for the effect size of scientific interest or the variability of the experimental material as assessed by the standard deviation of the character(s) being measured. The method is also difficult to apply to some factorial designs because of the desirability of having equal or nearly equal numbers in each group so the total number used is some multiple of the number of treatment combinations. Therefore, given that some estimate of the standard deviation is nearly always available, a power analysis should always be used in preference to the resource equation method for the type of FEDs described here.

Power analysis calculations require estimates of the standard deviation among experimental units, which must come from previous experiments or the literature, the effect size of biological interest, the required power, the significance level, and the alternative hypothesis (whether a one- or two-tailed test is appropriate). The effect size is the magnitude of the difference between treatment and control means, which the experiment is to be designed to detect. The larger the effect size, the greater the statistical power and the likelihood that statistical significance will be attained, other things being equal. A standardized effect size index, *d,* which is the effect divided by the pooled standard deviation, can be used to compare studies. This is a pure unitless number, and Cohen (1988) has suggested using it to determine sample size by specifying whether the research worker is interested in a "small," "medium," or "large" effect. In research in psychology, Cohen suggests that in the two-sample case, a small effect would be one where $d = 0.2$, a medium one would have $d = 0.5$, and a large one $d = 0.8$. Effect sizes of interest vary between disciplines. For example, effects classified as small can range from 0.13 in education to 0.55 in sociology. In animal research, effect sizes are likely to be large relative to other types of research because large doses of active compounds are often given to ensure that a response is detectable. Hwever, to date, no one has suggested small, medium, or large values for *d* in animal experiments.

The power of an experiment is the probability (sometimes given as a percentage) of detecting the specified effect size and calling it significant at the specified level. Typically, a power of 80 to 95% is specified, although in screening some vaccines, a power as high as 99% is specified because of the serious consequences that would follow if the experiment failed to detect a toxic batch of vaccine. The significance level, which is the probability that an effect will appear to be caused by the treatment although in reality it is due to chance, is usually set at 5%.

As an example, using Table 3, chloramphenicol given at a dose of 2500 mg/kg caused a 0.44 standard deviation change in the mean of CD-1 mice and a 2.52 standard deviation change in the mean of CBA mice. As explained in the table footnote, the interaction effect is −1.04 standard deviations. Suppose the aim is to set up a program to screen compounds for their effects on white blood cell counts using chloramphenicol at this dose as the positive control. A power analysis can be used to estimate the sample size that would be needed to detect changes of these magnitudes assuming, for example, a 5% significance level and a 90% power. Using nQuery Advisor (Statistical Solutions, Cork, Ireland), such experiments would require 90 mice per group using CD-1 mice, but only four mice per group using CBA mice. These differences are extreme because the CD-1 mice were very insensitive and CBA are very sensitive to the treatment.

A clear distinction must be made between the sample size requirements for the development of an in vivo assay using FEDs (a concern of this article) and the requirements for performing the assay itself. Traditional power calculations can be applied to meet the latter case, based on information gathered from the FED. The number of animals used in developing new screens or models should reflect the importance of this stage; indeed, using suboptimal conditions in model development will increase the number of animals required to achieve sufficient power in the routine screen, which may extend over many months so that the numbers of animals used in the screen can accumulate significantly. It is the total number of animals required to develop new models or assay conditions using a FED that is critical, rather than the number of groups or numbers within a group, because the results of all animals are used to assess all factors.

Sample Size Estimation in Practice

If the aim is to optimize an existing experiment, then the investigator should have a good idea of the likely magnitude of the treatment response as well as an estimate of the standard deviation of the character being measured. Thus, the actual response in standard deviation units observed in previous experiments can be calculated. The FED planned to optimize the experiment should probably be designed to detect an effect somewhat less than what is actually observed from the positive control. Because half of the animals in this type of FED will receive the vehicle and the other half the positive control, a first approximation of sample size can be obtained by assuming that the two groups are to be compared using a two-sample *t*-test. This approach will underestimate the required sample size to some extent because the standard deviation will be based on fewer observations in the FED than in a two-sample design; however, if the experiment is not too small, the underestimation should not be too serious. In Table 4, the total num-

Table 4 Standardized effect size *d* and sample size for a two-sample *t*-test assuming a two-tailed test, a significance level of 0.05, and a power of 95%

Standardized effect size	No./experiment[a]
0.6	148
0.8	84
1.0	54
1.2	40

[a] Note that this is the total number per experiment rather than the number per group.

bers (not numbers per group) required in a two-sample *t*-test, assuming a two-sided test and a 95% power, are shown. In this case, it is possible to obtain a rough estimate of the total numbers of animals to be used in the FED, split among the various treatment groups.

Next, the total number of treatment combinations should be worked out according to the number of factors, the treatment to be explored, and the number of levels of each factor (see Successful Implementation of a FED below). Example 1 below is a 3*2*2 design with 12 treatment and factor combinations, and Example 2 is a 2*2*2*2*2 factorial with 32 combinations.

The next step is to propose an appropriate level of replication (n) within each group. It is generally most appropriate to balance the designs by having equal numbers in each group. Thus, the total number of animals used will usually be some multiple of the number of treatment combinations. For Example 1, three animals per group would give a total of 36 animals (as actually used) and, according to Table 4, would probably be capable of detecting only large effect sizes of about d = 1.2 standard deviations. For Example 2, two animals per group (as actually used) would be able to detect only an effect of just over d = 0.8. In fact, the average detected effect was d = 0.88.

When the FED has been carried out, the combination of factors that provides the best response in the positive control will be known. At that time, there should be a good estimate of the standard deviation, and the information can then be used to design an optimum screening experiment close to the correct size using the power analysis method.

Successful Implementation of a FED

Setting up and performing the FED should involve a collaboration between scientist and statistician. We propose following the sequence of steps described below as preparatory work, carrying out the experiments, and follow up.

Preparatory Work

1. Define the objectives of the study and resource constraints.
2. Chart the experimental procedure in a flow chart.
3. Brainstorm all possible factors that could cause variation in the response(s) of interest.
4. Prioritize the factors so that at least the most important ones are included in the design of the first FED. Other factors might be considered in later experiments.
5. Consider the number of factors and appropriate dose levels to derive the number of treatment combinations. Multiply by the number of animals per group to obtain the total number of animals to be used, as outlined above, and verify that the number of animals appears reasonable according to Table 4. Order the animals in a timely manner. Animals may need a couple of weeks to adjust to a new environment.
6. Consider whether to use a completely randomized or a randomized block design. Blocking breaks the experiment up into a number of "mini-experiments" within which the material (animals) can be as homogeneous as possible, thereby increasing precision (see Cox 1958, Montgomery 1997, or any standard statistical text for details).
7. Plan how the animals are to be housed and what randomization procedure is to be used.
8. Consider the practicality of the design. Can any steps be taken to aid logistics and to minimize the risk of potential errors?

Note that in some cases, a few follow-up experiments may be worthwhile in establishing full confidence in the conditions derived, or to explore in more detail outside the ranges initially addressed.

Performing the Experiments

1. Ensure that any observations and deviations from the protocol of the design are recorded with the results because they may influence the statistical analysis.
2. Record all relevant data from experiments, even if it is considered unlikely that they will influence the conclusions.

Followup

1. Analyze the data. Visualize (plot) the data for individual animals before moving on to analysis of mean responses. Graphical study of residuals plots (Example 2) can also detect outliers and ensure that the assumptions underlying the ANOVA are met.
2. Pay attention to the robustness of the experiment (i.e., the extent to which the results are influenced by minor variation in conditions that cannot be adequately controlled). This aspect is particularly important in cell-based and enzyme assays. A randomized block design with replication over time, and/or in different laboratories or by different staff, will often give a good indication of the robustness of the experiment by the extent to which the results are repeatable under these conditions.

Examples of Factorial Designs

Example 1: Full Factorial Design

The objective of this study was to identify conditions with a new animal model to maximize the sensitivity for testing compounds in a screen. The following factors were included: time of fasting (0/2/4 hr), age of rat (young / old), and treatment (control/treated). All combinations of these factor settings were included in 12 groups of three rats, according to the design layout given in Table 5.

Table 5 Example layout of a 3*2*2 factorial design used to optimize a protocol for screening in drug development (for details see text)

Group	Event time (hr)	Age	Treatment
1	0 (–)	Old (+)	Treat (+)
2	0 (–)	Old (+)	Control (–)
3	0 (–)	Young (–)	Treat (+)
4	0 (–)	Young (–)	Control (–)
5	2 (0)	Old (+)	Treat (+)
6	2 (0)	Old (+)	Control (–)
7	2 (0)	Young (–)	Treat (+)
8	2 (0)	Young (–)	Control (–)
9	4 (+)	Old (+)	Treat (+)
10	4 (+)	Old (+)	Control (–)
11	4 (+)	Young (–)	Treat (+)
12	4 (+)	Young (–)	Control (–)

Response measurements were taken on each animal over several time periods, and these data were summarized by evaluating area under the curve (AUC[1]), a useful summary measure of response. The data were analyzed by ANOVA, which led to the conclusion that the age of animal had no effect on the AUC results. However, there was a significant treatment by fasting time interaction. In Table 6, the average AUCs for the different treatment and fasting time combinations are shown. Note that all 36 animals provide information on these relations, and the 4-hr fasted time point provides a better response than either the nonfasted or 2-hr fasted animals.

The main outcomes of applying FED in this example were as follows:

- The design contributed to the development of a better animal model for screening compounds in that a switch was made from using a nonfasted animal to using a 2-hr fasted animal.
- The improved model opened up a window (treatment control difference) for testing compounds nearly twice as large as previously seen.
- The wider window enabled follow-on dose response studies to be completed at much lower doses than previously possible, allowing the screen to pick out more candidate drugs than before and possibly contributing to the welfare of the animals.
- The wider window means that the sample size could be reduced without compromising the sensitivity of the screen and/or that smaller responses could be detected.

Table 6 Summary of treatment*time interaction in Example 1 (data expressed in standard deviation units)

Combination	Event time (hr)	Treatment	Average area under the curve	Difference
1	0	Treat	5.09	
2	0	Control	7.64	2.55
3	2	Treat	5.21	
4	2	Control	9.91	4.70
5	4	Treat	5.85	
6	4	Control	10.79	4.94

Example 2: Full 2^5 Factorial Design Used to Help Develop a Model for Testing Agents That May Reduce Cancer

Multiple lung tumors can be induced in some strains of mice exposed to a carcinogen such as urethane. These cases could be used as a model to test compounds that might prevent or reduce the incidence of cancer. For example, diallyl sulphide, one of the active ingredients of garlic, may help to protect against cancer (Hong et al. 1992). The proposed protocol would be to administer the test compound to mice for a period of time, then expose them to a carcinogen, and after an appropriate period (usually about 5 mo), sacrifice the mice and count the number of tumors on the surface of the lungs. This protocol is quantitative, with the number of small (2- to 3-mm diameter) tumors ranging up to about 30 and exceptionally as high as 100 or more. Because the tumors are very small and the animals are sacrificed well before they become sick, the endpoint is humane.

To develop the model, a FED was used to test the following: (1) strain**,** using two strains of mice known to be susceptible (A/J and NIH); (2) sex (males and females); (3) diet (RM1 expanded diet or RM3 pelleted diet); (4) carcinogen (urethane or 3-methylcholanthrene); and (5) treatment (diallyl sulphide, compared with vehicle, administered by gavage over a period of 3 days before administration and 2 days after administration of the carcinogen, which was given by intraperitoneal injection). Thus, this was a 2^5 factorial design with 32 combinations of factors and treatments and two mice in each group, with the aim of maximizing the observed treatment effect. Full details of dose levels and so forth are not discussed in this article because the aim is to demonstrate the principle of the use of such designs.

The data were transformed by taking the square root of the (tumor count plus one) to normalize the residuals and equalize the variation in each group (Festing et al. 1994). A normal probability plot of the residuals (deviations from group means) is shown in Figure 1. Although the bulk of the observations form a good straight line, implying a normal distribution, there are six points that could be considered to be outliers. These points represent three pairs of observations in which the two mice in the group differed more than

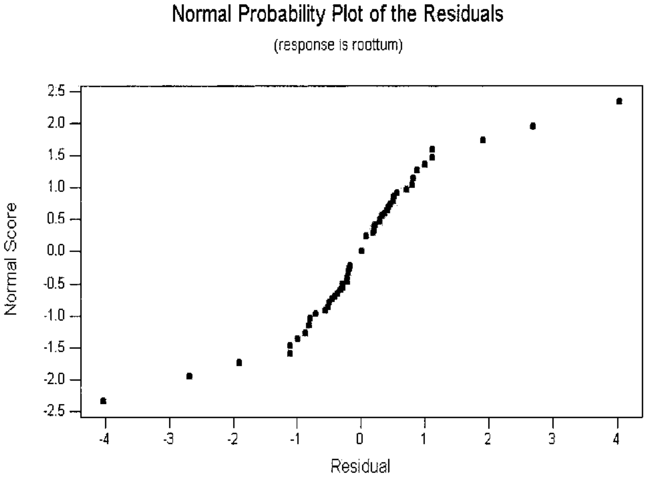

Figure 1 Normal probability plot of the residuals (deviations from model group means) versus the "normal score" for Example 2. This plot is provided by a standard statistical package as a check of whether one of the assumptions presumed in analysis of variance is met. If these deviations are normally distributed, this line should be straight, which is the case for most of the middle points although there are six points that are outliers due to large differences between three sets of animals receiving the same treatments. See text for more details.

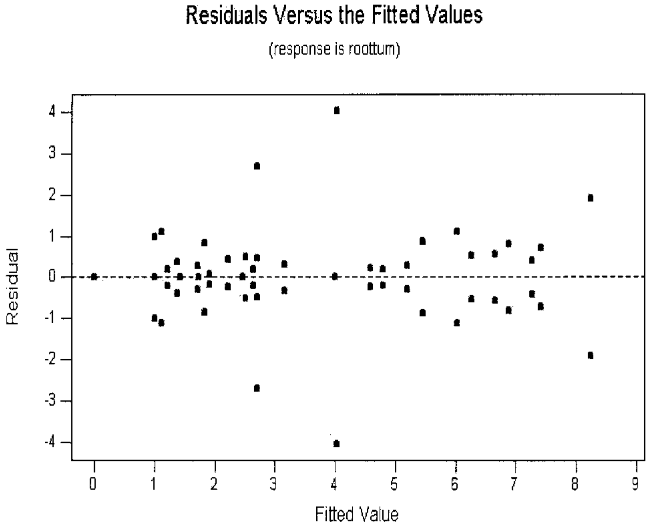

Figure 2 Plot produced by a standard statistical package of residuals (deviation from group means) versus model group means (see test Example 2). Absence of any trend or pattern implies that the variation is approximately the same in each group. Note that in this case, there is a symmetrical pattern about the zero line resulting from two animals in each group.

would be expected by chance. For example, two NIH mice had 0 and 29 tumors, two A/J mice had 0 and 65 tumors, and two A/J mice had 40 and 103 tumors, respectively. The reason for these differences is unknown. There was no evidence of a mistake in recording the numbers. It is possible that the two mice without tumors were injected incorrectly, but this possibility is pure speculation, and no adjustment has been made in the subsequent analysis. The plots of model group means versus residuals (Figure 2) provide no evidence of a pattern that would suggest heterogeneity of variance. Thus, the assumptions necessary for the analysis of variance are adequately met.

The ANOVA revealed that there were no significant three-, four-, or five-way interactions ($p > 0.05$ in all cases). Statistically significant strain*carcinogen ($F_{1,32} = 26.5$, $p < 0.0005$) and treatment*carcinogen ($F_{1,32} = 5.0$, $p = 0.032$) interactions were found. The main effects of treatment, strain, and carcinogen were also statistically significant ($p < 0.05$).

The treatment*carcinogen interaction is of interest in optimizing the model. The means, in standard deviation units, are given in Table 7. Diallyl sulphide, the positive control, was only marginally effective in reducing the number of tumors induced by urethane, resulting in a mean reduction of only 0.19 standard deviations, whereas the result was a reduction of 1.56 standard deviations against tumors induced by 3-methlycholanthrene. In effect, the screen would be effective in detecting compounds that give protection from tumors caused by compounds that act only in a similar way to 3-methlycholanthrene. The highly significant strain*carcinogen interaction was due to strain A/J mice developing more tumors with 3-methylcholanthrene than with urethane, whereas the opposite was true with strain NIH (data not shown). This difference might be worth taking into account when choosing a strain for future screening even though there is no overall evidence that one strain would be better than another.

These results clearly raise several questions about this model, which are not discussed in this article. However, if the aim is to detect compounds that protect against cancer induced by a range of carcinogens, then the protocols would require further development. Thought should be given to the dose levels and timing of administration of the treatments and the carcinogens. However, less attention to the diet or sex of mouse is needed because there was no evidence that these factors played any part in determining the response to diallyl sulphide. Thus, a factorial study of this sort can be

Table 7 Lung tumor response to diallyl sulphide after treatment with two carcinogens (in standard deviation units)

	Treatment		
Carcinogen	Vehicle	Diallyl sulphide	Difference
Urethane	3.18	3.37	0.19
3-MC	1.50	3.06	1.56

used to identify factors that can maximize response in future studies and those that are likely to have little or no effect. In this case, the aim of screening compounds that might reduce tumor incidence with a mode of action similar to diallyl sulphide can probably be done only by using 3-methlycholanthrene as the carcinogen challenge.

Discussion

The objective of this article is to describe the use of FEDs in the development and optimization of animal experiments by exploring the effects of, for example, age, strain, sex, and protocol-specific factors on the sensitivity of the experiment, using the treatment response as an index of sensitivity. Clearly the positive control should have a mode of action that is similar to the presumed mode of action of the chemicals being screened. After finding which factors influence the treatment difference and the combination of these factors and levels of factors that leads to the largest treatment effect, subsequent experiments can be designed with a high level of sensitivity. In other words, sample size can be reduced for equivalent quality of information. With experiments that are repeated frequently with only minor changes (e.g., screening experiments in drug development), the result will be a significant cumulative savings in animals and scientific resources over the life of the screen. Furthermore, as demonstrated above, the optimization process is more efficient compared with traditional approaches such as the OVAT approach.

Two additional points that must be taken into account in designing future experiments are not easily addressed using FEDs alone. The first point is that the experimental protocol must be robust so that the responses are not seriously altered by minor changes in uncontrollable conditions. FEDs can be used to show which controllable factors are important, but they cannot be used easily to investigate the many uncontrollable factors that can affect the results. These factors can be explored, however, by designing the FEDs as randomized block designs when possible. Blocking splits the experiment up into a number of mini-experiments that are repeated over a period of time, in different laboratories, or with different personnel. If the individual blocks give essentially the same results, then the experiment should be robust.

The second point is that the within-group variability can have a large impact on sample size estimates, but this effect cannot be studied easily using FEDs because group sizes are normally small, which is why they are so efficient. Outbred stocks of mice and rats are usually phenotypically more variable than inbred strains due to the genetic variation within the colony, and this variability may mean that more outbred than inbred stocks may be needed (Festing 1976; Festing et al. 2001; Ghirardi et al. 1995). Similarly, subclinical infection can also increase interindividual variability (Gartner 1990). However, large sample sizes are needed to detect statistically significant differences in variability between groups of animals. It is probably best to use only high-quality specific pathogen-free animals of an isogenic strain when they are available, although an outbred stock that is equally sensitive or more sensitive than an inbred strain could be justified.

In conclusion, FEDs can, and should, be used to optimize animal experiments to derive robust conditions efficiently. Particular benefit is gained when these conditions are applied repeatedly with only minor changes in treatments, such as in drug screening and development.

Acknowledgments

The authors thank the following colleagues at AstraZeneca for providing data and information for this article: John Myatt, Debbie Martin, Brendan Leighton, and Annie Atkinson. The data in Example 2 were generated with the support of the UK Ministry of Agriculture, Fisheries and Food (contract FS1710, awarded to M.F.W.F).

References

Calleri de MMC, Mayorga LS, Puig de CON. 1992. Optimization of culture conditions for toxin production of type G *Clostridium botulinum.* Zentralbl Bakteriol 277:161-169.

Clarke GM, Kempson RE. 1997. Introduction to the Design and Analysis of Experiments. London: Arnold.

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale NJ: Lawrence Erlbaum Associates.

Cox DR. 1958. Planning Experiments. New York: John Wiley and Sons.

Festing MFW. 1976. Phenotypic variability of inbred and outbred mice. Nature 263:230-232.

Festing MFW. 1994. Reduction of animal use: Experimental design and quality of experiments. Lab Anim 28:212-221.

Festing MFW. 1995a. Reduction in animal use 35 years after Russell and Burch's Principles of Humane Experimental Technique. ATLA 23:51-60.

Festing MFW. 1995b. Reduction of animal use and experimental design. In: Goldberg AM, van Zutphen LFM, eds. World Congress on Alternatives and Animal use in the Life Sciences: Education, Research, Testing. Vol 11. Larchmont NY: Mary Ann Liebert, Inc. p 43-49.

Festing MFW, Lovell DP. 1996. Reducing the use of laboratory-animals in toxicological research and testing by better experimental-design. J R Stat Soc Series B-Methodol 58:127-140.

Festing MFW, Yang A, Malkinson AM. 1994. At least four genes and sex are associated with susceptibility to urethane-induced pulmonary adenomas in mice. Genet Res 64:99-106.

Festing MFW, Diamanti P, Turton JA. 2001. Strain differences in haematological response to chloramphenicol succinate in mice: Implications for toxicological research. Food Chem Toxicol 39:375-383.

Festing MFW, Overend P, Gaines Das R, Cortina Borja M, Berdoy M. 2002. The Design of Animal Experiments: Reducing the Use of Animals in Research Through Better Experimental Design. London: Royal Society of Medicine Press Limited.

Gartner K. 1990. A third component causing random variability beside environment and genotype. A reason for limited success of a 30 year long effort to standardize laboratory animals. Lab Anim 24:71-77.

Ghirardi O, Cozzolino R, Guaraldi D, Giuliani A. 1995. Within- and between-strain variability in longevity of inbred and outbred rats under the same environmental conditions. Exp Gerontol 30:485-494.

Hong J-Y, Wang ZY, Smith TJ, Zhou S, Shi S, Pan J, Yang CS. 1992.

Inhibitory effects of diallyl sulfide on the metabolism and tumorigenicity of the tobacco-specific carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) in A/J mouse lung. Carcinogenesis 13: 901-904.

Maas WJ, de Graaf IA, Schoen ED, Koster HJ, van de Sandt JJ, Groten JP. 2000. Assessment of some critical factors in the freezing technique for the cryopreservation of precision-cut rat liver slices. Cryobiology 40:250–263.

Mead R. 1988. The Design of Experiments. Cambridge: Cambridge University Press.

Metzger BL, Jarosz PA, Noureddine S. 2000. The effect of a high-fat diet and exercise on the expression of genetic obesity. West J Nurs Res 22:736-748.

Montgomery DC. 1997. Design and Analysis of Experiments. 4th Ed. New York: John Wiley & Sons, Inc.

Nesnow S, Mass MJ, Ross JA, Galati AJ, Lambert GR, Gennings C, Carter WH Jr, Stoner GD. 1998. Lung tumorigenic interactions in strain A/J mice of five environmental polycyclic aromatic hydrocarbons. Envir Health Perspect 106(Suppl 6):1337-1346.

Reiken SR, Van WBJ, Sutisna H, Kurdikar DL, Davis WC. 1994. Efficient optimization of ELISAs. J Immunol Methods 177:199-206.

Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ. 2001. Maximization of signal derived from cDNA microarrays. Biotechniques 30:202-208.

Wu CFJ, Hamada M. 2000. Experiments, Planning, Analysis and Parameter Design Optimization. New York: John Wiley & Sons, Inc.