

# Refinement and Reduction Through the Control of Variation

**Michael F.W. Festing**

*c/o MRC Toxicology Unit, University of Leicester, Leicester LE1 9HN, UK*

**E-mail:** *mfwf1@le.ac.uk*

**Summary** — The key to doing animal experiments efficiently, while using the minimum number of animals without loss of scientific information, lies in good control of random variation, and recognition and control of “fixed effect” variation, such as the sex or strain of the animals. However, many scientists erroneously assume that the use of outbred, genetically heterogeneous animals is justified, because in some way, they more closely model humans. Unfortunately, all this does is to increase the phenotypic variation, which results in less-powerful experiments. If the aim is to model variation in human responses, this can be done by using a small number of animals from several isogenic strains, without increasing the total number of animals. Reducing inter-individual variation, whether caused by genetic or non-genetic causes, will nearly always result in improved experiments. Fixed-effect variation, such as the sex of the animals, can be taken into account, either by restricting the conclusions to the sex actually used, or by assuming that the other sex would respond in the same way, or by including both sexes in the study, by using a factorial design, without increasing the total number of animals.

**Key words:** *animal experiments, experimental design, reduction, refinement, statistics.*

## Introduction

Good experimental design can lead to a reduction in animal use. It requires an understanding of variation and how it can be controlled either simply by choosing uniform material, or by controlling it in the design of the experiment. There are two main types of variation that need to be taken into account. First, variation can be caused by so called “fixed effects”, such as the sex, strain and age of the animals, which can be controlled by the scientist at a level thought to be appropriate. Thus, the scientist can choose whether to use males or females or both, and a choice will be made, taking into account the implications of the resulting choice. Second, the variation can be caused by so called “random effects”, such as inter-individual differences and measurement error, which cannot be fixed by the researcher. It is important to understand these two sources of variation and the ways in which experiments can be designed to take account of them.

## Genetic Variation and the Use of Outbred Stocks in Research: a Student Exercise

A scientist using laboratory mice or rats is faced with four main classes of stock from which to choose. First, “outbred stocks” are usually closed colonies of genetically heterogeneous animals, which go under names such as “Wistar” and “Sprague-Dawley” rats and “Swiss” or “CD-1” mice. Second, there are inbred strains, which have

been produced by many generations of brother × sister mating and are like immortal clones of genetically identical individuals. Third, there is a range of stocks carrying mutations and genetic polymorphisms. These can be transferred to, and maintained on, either an inbred or on an outbred genetic background by a process of backcrossing. Fourth, there are now many genetically modified strains, most of which behave like mutants and polymorphisms. The choice of either inbred or outbred animals in general research illustrates the way in which variation can be controlled by the scientist in order to improve the quality of his or her research.

The students (ranging in experience from PhD students to experienced research workers) taking the UK “Module 5” course to qualify them for a project licence to use animals, were told that, “You have developed a compound which you think will help to prevent rejection of transplanted hearts, and you want to test this experimentally. The experiment will involve heart grafts between donor and recipient rats (whose own hearts are not removed). There will be a control group and one treated with the test compound. The following rat strains are available: outbred Wistar and Sprague-Dawley; inbred ACI, F344 and LEW”. They were asked, “Which strains will you use as donor and recipient, and why?”. They were also told that, “You know it is not acutely toxic but need to do a long-term toxicity study with control and treated rats. You discuss this with a toxicologist who points out that you wish to model humans who are genetically heterogeneous. He suggests that you use outbred genetically heterogeneous Sprague-Dawley

rats, the strategy used by virtually all toxicologists in drug development. Do you decide to accept or reject his advice? Give your reasons”.

The students were given 5 minutes to discuss the problem with their nearest neighbour, and then the results were discussed by the class as a whole. This exercise has produced a wide range of results. Probably well over half the groups (exact figures have not been compiled) suggest that an outbred stock should be used as either the donor or recipient, or both, in the heart transplant part of the problem, and well over 90% say that they would accept the advice of the toxicologist and use the Sprague-Dawley rats in the toxicity test. In nearly all cases, the reasons given for choosing the outbred stock was to try to replicate human genetic variation in the group of test animals. In fact, the first part of the exercise arose from a real problem. I received an e-mail from an investigator who had chosen to use inbred F344 rats as the heart donors and outbred Sprague-Dawley as the recipients, as follows: “We transplanted hearts of young . . . Fishers into . . . recipient Sprague-Dawleys. An outbred strain was selected since such animals are usually heartier and easier to handle . . . We are puzzled by our results . . . palpable heart beats were evident in the saline group long after acute rejections . . . were expected . . . Results in the experimental groups varied considerably . . .”.

Clearly, the genetic heterogeneity was causing severe problems, because if some of the control hearts were not even being rejected, it would have been impossible to detect whether or not a test compound could prolong heart graft survival. A much better strategy would have been to use two inbred strains as donor and recipient. A pair of strains such as ACI and F344 are known to differ at the major histocompatibility complex (MHC), so transplants between this pair would be rejected quite vigorously and at a relatively uniform time. This uniformity means that quite small sample sizes would be needed to detect a difference in mean time to rejection, other things being equal. If the aim had been to model a less acute form of rejection (given that humans may be tissue typed) then a pair of strains such as F344 and LEW might have been chosen. These are known to be similar at the MHC, so graft rejection would have been slower, but still relatively uniform among individuals. Note that “modelling” humans, in this respect, means replicating the sort of range of responses (acute or chronic) that may be encountered in human subjects.

In accepting the advice of the toxicologist and choosing an outbred stock for the toxicity test, most students again gave as their reasons the need to try to replicate the genetic variation found in the human population. However, a toxicity test is also a controlled experiment with a treated and control group, which are compared at the end of the exper-

iment. If there is a lot of experimental “noise”, it will tend to obscure the effects of the treatments. This can be seen if the genetic variation is grossly exaggerated. Suppose a “genetically heterogeneous” group of animals was assembled and divided strictly at random into a treated and control group.

The experiment might consist of the following. In the control group, there might be a mouse, rabbit, cat, frog and horse. In the treated group a cat, dog, rhesus monkey, zebra, and rat. The only difference between a mouse and a rabbit is genetics, so this would be a “genetically heterogeneous” group. But it would clearly be a very poor experiment, because the treated and the control group would differ even before the experiment started for virtually any character, so sample size would need to be enormous to take account of this. A similar effect, though less extreme, would be seen with the use of outbred stocks. Suppose, however that the scientist had available two mice, two rabbits, two cats, two frogs, etc. Then, she could do an experiment in which the two mice, rabbits, cats, etc. were assigned to the treated and control groups, respectively. In this way, the treated and control group would be balanced for the two species. This would provide a much more powerful experiment because the treated and control groups would be identical (with respect to species) at the start of the experiment, but the range of susceptibility would still be sampled as the difference between species.

Unfortunately, it is impossible to do an experiment like this, using outbred stocks, because each individual is genetically different. It is, however, possible to do one using several inbred strains. Basically, turning the random inter-individual variability found in the outbred stock into a fixed effect using at least two genetically identical individuals substantially increases the power of the experiment (1).

A similar argument can be seen with respect to the way that the sex of the animals is treated. Assuming a researcher wants to include both males and females in an experiment, he or she could argue that by taking a random sample of weanling animals and dividing them at random into a treated and control group, it is almost certain that both males and females would be represented in the experiment. Table 1 shows a hypothetical example of just such an experiment. A statistical analysis using a two-sample *t* test shows that there is no significant difference in the scores between the two groups. However, suppose the researcher now identifies the two sexes (Table 2). By a lucky chance, there are equal numbers of males and females in each group. However, it is clear that the females are scoring slightly lower than the males, and when this is now taken into account in a new statistical analysis, exactly the same difference is now statistically significant. Thus, by identifying the sex of the animals rather than leaving it to chance, the

**Table 1: A hypothetical experiment using unsexed animals**

	Control	Treated	Difference
	34	42	
	46	42	
	35	51	
	42	48	
	42	44	
	42	38	
	44	36	
	43	45	
	39	39	
	34	44	
	46	44	
Mean	40.64	43.00	2.36
SD	4.50	4.34	

*The data might, for example, be some behavioural score. When means are compared using the *t* test, the difference in means of 2.36 units is not statistically significant ( $p = 0.224$ ), and the pooled within-group standard deviation is 4.42 units.*

**Table 2: The same hypothetical experiment as Table 1, but in this case, the females have been identified with an "X"**

	Control	Treated	Difference
	34X	42	
	46	42X	
	35X	51	
	42	48	
	42	44	
	42	38X	
	44	36X	
	43	45	
	39X	39X	
	34X	44	
	46	44	
Mean	40.64	43.00	2.36
SD	4.50	4.34	

*Note that by chance, there were the same number of females in each group. However, it is now possible to take the sex of the individuals into account by doing a two-way analysis of variance. As a result, the within-group pooled standard deviation is reduced to 2.48, and the identical difference between the means of 2.36 units has become statistically significant ( $p = 0.032$ ).*

researcher has substantially increased the power of the experiment to detect a treatment difference. Exactly the same effect could be obtained with any genetic locus that influences a character of interest. Genotyping the animals in such cases will increase the power of the experiment. Unfortunately, genotyping is expensive and time-consuming and is only practical for a few loci. If the animals are genotyped at many loci, it will be found that the treated and control groups cannot be balanced.

The best way around this is to use the same total numbers, but to use small numbers of animals from several inbred strains in a factorial experimental design. For example, a study of the effect of butylated hydroxyanisole, an antioxidant commonly used to preserve food, on a liver enzyme activity was done using two treated and two control mice of each of four inbred strains; a total of 16 mice (Table 3). There was quite a large response to the antioxidant ( $p < 0.001$ ), with some evidence that the strains differed in response ( $p = 0.03$ ). Had an outbred stock been used, then the experiment would have needed about the same number or even more mice, but it would have been impossible to tell whether the response was under genetic control.

## Control of Variation by Experimental Design

The lesson to be learned from this example is that uncontrolled variation results in "noise", which may obscure a true treatment effect, and that, where it is possible to control for this variation, the power of an experiment can be increased. Some "time and space" variables may also need to be taken into account at the design stage of an experiment. It may not be possible to house all the animals in the same room or on the same shelf. It may be necessary to make complex measurements on the animals, and this may take time, so that not all the animals can be measured on the same morning, day or week, and there is no assurance that things such as biological rhythms, micro-environments and even personnel will remain the same for the whole experiment. A completely randomised design will ignore these sources of variation, which may grossly inflate the estimates of within-group variation, leading to a requirement for large sample sizes or experiments that lack power. A randomised block, Latin square or crossover design should be considered in these circumstances (2). The randomised block design, for example, splits up the experiment into a number of replicates or blocks, typically with one animal (or other experimental unit) in each treatment group. Thus, if the experiment has five treatment groups, a block will consist of five animals, one on each treatment. Blocks can be replicated over a period of time, or in different rooms or even on different shelves in the animal

**Table 3: Liver enzyme (ethoxy resorufin O-decarboxylase, EROD) activity in control mice and in those treated with butylated hydroxyanisole (BHA) in the diet**

Strain	Liver EROD activity (arbitrary units)	
	Control	Treated
A/J	7.05	17.70
129/O1a	7.55	16.15
NIH	8.95	15.60
Balb/c	7.85	23.05

*The experiment was done as a randomised block experimental design. Thus, the first block consisted of two mice of each strain assigned at random to either the treated or control group. This was repeated about two months later, giving a total of 16 mice used in the experiment. Note that about this number of mice would have been needed had a single outbred stock been used. Data were analysed by a two-way analysis of variance. Effect of BHA treatment was significant at  $p < 0.001$ , and the treatment  $\times$  strain interaction was significant at  $p = 0.03$  largely because the response in Balb/c mice was slightly greater than in the other strains. The biological significance of this is debatable.*

house. The use of such designs can substantially increase the power of the experiment at no extra cost, but with a slight increase in the complexity of the statistical analysis. Mead (3) noted that, in agricultural research, about 80% of experiments use a randomised block experimental design, and he suggested that more-advanced designs might often be appropriate. However, very few papers using research animals use such designs, suggesting that there is scope for improvement.

### Random Variation and Sample Size Determination

It is important to use an appropriate sample size. If it is too large, then animals will be wasted. Too small and scientifically important effects may be missed. Sample size determinations can also be used to explore the effect of better control of random variation. Even a small reduction in the within-group standard deviation can result in a substantial reduction in the required sample size, as estimated using power and sample size calculations.

Several modern statistical packages and some stand-alone programs are available to help determine sample sizes. There are also a number of free Web sites that offer calculations for the simpler cases. A full description of how sample size is determined is beyond the scope of this paper, but very

briefly, when comparing two samples, the scientist needs to determine the effect size of biological interest that he or she wants the experiment to be able to detect. For quantitative characters, this is the minimum difference between the means of the two groups that it would be worth detecting. For qualitative characters, the proportion affected in the control and treated group that the experiment aims to detect must be specified. For example, if the control is expected to have a 30% incidence of a condition, such as a tumour, the investigator must specify what incidence, such as a 50, 60 or 70%, it would be important to be able to detect in the treated group. A small increase, such as from 30% in the control to 40% in the treated group, will require a large experiment. Next, for quantitative characters, some estimate of the standard deviation is necessary, and this must come either from a previous experiment, a pilot study or the literature. The significance level needs to be specified, though it is usually set at 5%. The power also needs to be specified, and it is usually set somewhere between 80 and 90%, although, if the consequences of failing to detect an effect would be very serious, as in testing the virulence of human vaccines, a higher power may be specified. Finally, the alternative hypothesis needs to be specified. Usually, the null hypothesis is that there is no difference between the groups, with the alternative hypothesis being that there is one. In this case, a two-tailed test is used. However, occasionally there will be some biological reason why the response can only go in one direction, in which case, a one-tailed test can be used. Once all these things have been specified, computer programs available within many statistical packages, and free on the Web, can be used to estimate an appropriate sample size (the actual formulae are not all that easy to use; 4).

### Control of Fixed Effects

There are basically three ways of dealing with fixed effects such as the strain or sex of the animals. The effect can be fixed at one level, such as using only males or strain X. In this case, the author may be very cautious and restrict the conclusions to males and/or strain X, or he or she may simply assume that the results would have been the same if females and strain Y had been used.

At first sight, this extrapolation would appear to be dangerous and unjustified. However, there is an infinite number of these fixed effects, so some extrapolation is always necessary. For example, most scientists would assume that the results would be the same had cages of a different size been used, or if the animals had had a different diet, or if the temperature of the animal house had been a few degrees cooler. However, in some cases, there is real interest in determining the effects of some of these

fixed effects. This can be done by using a factorial experimental design as already described and illustrated in Table 3. Unfortunately, many scientists seem to assume that, if the experiment is to be done, say, in both males and females, this will require twice as many animals. In fact, it can usually be done by using approximately the same total number of animals by using half the number of males that would have been used, and adding an equal number of females. Since the power of the experiment depends largely on the total number of animals, not the within-group sample size, this does not lead to any appreciable loss of power. Such designs provide more information than single factor designs for approximately the same input of scientific resources (2).

Factorial designs can be used to explore the effects of many factors simultaneously and can even be used to optimise conditions for single factor designs that are to be used repeatedly (5). For example, the response to a drug measured as the difference in mean between treated and control animals may depend on the sex, strain, age, prior treatment (e.g. starvation), route of delivery, times of subsequent measurement and method of measurement. Advanced factorial designs can be used to find out which of these factors are important, and which level of the factor gives the best response.

## Conclusions

Good experimental design ensures that experiments give, as far as possible, the right results with sufficient power to detect clinically or biologically important responses but that are not so large that scientific resources and animals are wasted. A good understanding of variation and the way that it can be controlled is essential. Basically, this variation is of two types: variation due to random influences, and that due to fixed factors that can be controlled by the investigator. Random variation is minimised by using animals of a narrow age and weight range,

by using inbred strains if the experiment involves rats or mice, by using careful experimental technique, so as to minimise measurement error, and by using randomised block or other more advanced designs to take account of time and space variation.

Fixed effects are either controlled at a single level, in which case, the investigator has to decide whether to limit the conclusions to the particular factors selected or whether extrapolation to a wider universe is justified. Some extrapolation to the wider universe is essential if science is to progress, because each experiment involves a unique collection of these fixed effects, but many of them are of little importance to the final outcomes. Where a fixed effect, such as the sex or strain of the animal, needs to be considered in more detail, factorial designs can be used. These provide a powerful method of finding out which factors and levels of those factors are important and can be used to optimise experiments with many different factors without the use of excessive numbers of experimental subjects.

## References

1. Festing, M.F.W. (1995). Use of a multi-strain assay could improve the NTP carcinogenesis bioassay program. *Environmental Health Perspectives* **103**, 44–52.
2. Festing, M.F.W., 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*, 112pp. London, UK: Royal Society of Medicine Press Limited.
3. Mead, R. (1988). *The Design of Experiments*, 620pp. Cambridge, New York: Cambridge University Press.
4. Dell, R., Holleran, S. & Ramakrishnan, R. (2002). Sample size determination. *ILAR Journal* **43**, 207–213.
5. Shaw, R., Festing, M.F.W., Peers, I. & Furlong, L. (2002). The use of factorial designs to optimise animal experiments and reduce animal use. *ILAR Journal* **43**, 223–232.