Randomized Block Experimental Designs Can Increase the Power and Reproducibility of Laboratory Animal Experiments

Michael F. W. Festing

Abstract

Randomized block experimental designs have been widely used in agricultural and industrial research for many decades. Usually they are more powerful, have higher external validity, are less subject to bias, and produce more reproducible results than the completely randomized designs typically used in research involving laboratory animals. Reproducibility can be further increased by using time as a blocking factor. These benefits can be achieved at no extra cost. A small experiment investigating the effect of an antioxidant on the activity of a liver enzyme in four inbred mouse strains, which had two replications (blocks) separated by a period of two months, illustrates this approach. The widespread failure to use these designs more widely in research involving laboratory animals has probably led to a substantial waste of animals, money, and scientific resources and slowed down the development of new treatments for human and animal diseases.

Key Words: experimental design; randomized block; repeatability; animal experiments; reproducibility

Introduction

A fundamental assumption in experimental biology is that if an experiment is well designed, correctly executed, properly analysed, and adequately documented, the results should be reproducible, apart from the occasional type I error (false positive) associated with the chosen significance level. However, several recent publications have found that excessive numbers of animal experiments are unreproducible (i.e., the results could not be repeated by different investigators). For example, Begley and Ellis (2012) attempted to repeat 53 landmark experiments concerned with cancer research but were only able to do so with six of them. In some cases, the original authors were unable to repeat their own experiments. In another paper, investigators (Scott et al. 2008) noted that there were more than 50 reports of drugs that alleviated the symptoms of amyotrophic lateral sclerosis (ALS) in the standard transgenic mouse model of this disease, but only one had any effect in humans. A detailed study showed that there are a number of confounding factors that need to be controlled when using this model. The authors devised a better protocol, which controlled these confounding factors, and then rescreened these 50 drugs plus another 20. They found that none of the drugs was effective in the mouse model. Similarly, Prinz and colleagues (2011) were able to reproduce only 20 to 25% of the results of 67 published studies and claimed that, within the pharmaceutical industry, it is accepted anecdotally that less than half of academic papers give reproducible results.

In many cases, lack of repeatability is due to failure to adhere to some of the most basic requirements of good experimental design. A survey of 271 animal experiments showed that 87% of papers did not report randomization of treatments to subjects (although this does not necessarily mean that it was not done), and 86% did not report blinding in situations where it would be appropriate (Kilkenny et al. 2009). Such failures can lead to biased results and unreproducible (in the same laboratory) or unreproducible (in a different laboratory) experiments. Lack of reproducibility in other laboratories may also be caused by treatment x environment interactions. For example, animal houses may differ in the physical environment, management, or microbiota in such a way as to alter the relative treatment differences. Results may also be unreproducible because the wrong strain of animals was used. There is no effective genetic quality control of outbred stocks. In one study, for example, the investigators obtained 26 weekly samples of 30 Sprague-Dawley rats from a commercial supplier and tested them for response to a synthetic polypeptide, a response controlled by a single gene in the major histocompatibility complex (Simonian et al. 1968). On average, about 80% of the rats were responders and, for the first 12 weeks, the percentage of responders in each sample varied about this mean. However, in weeks 13, 17, 18, 19, and 20, only about 5% of the rats were responders. These rats cannot have come from the same colony and may have responded differently to other experimental treatments, but there was no indication from the breeder that different rats had been supplied. There have been other examples of the wrong animals being supplied by commercial breeders (Festing 1982). And, of course, any single experiment has a 5% chance of getting a false positive result due to statistical
Randomized Block Experimental Designs

The “randomized block” (RB) design is a generic name for a family of experimental designs in which the experimental material is split up into a number of “mini-experiments” that are recombined in the final statistical analysis. Typically, in each block there is a single experimental unit to which each treatment is assigned (although there can be more than one). They include crossover designs, within-subjects designs, matched designs, and Latin square designs. These have a number of useful properties and should be more widely used in research involving laboratory animals. They can be used to:

1. Spread the experiment over a period of time and/or space. So if there are, say, four treatments and the cage of animals is the experimental unit, each block will consist of four cages. Block 1 might be started this week, block 2 next week, and so on. If six blocks are needed, then the experiment will be extended over a 6-week period. Each block may involve a different batch of animals, which could be of a slightly different age or weight with possible differences in the batches of diet, or its age since it was manufactured. Cages may be placed at different levels in a rack. None of these variables is of interest, so they are removed as a block effect in the statistical analysis. If the relationship between the observations alters, then treatment effects will become statistically less significant. If the relative values remain unchanged, then this implies a good level of repeatability with real treatment effects being more likely to be detected.

2. Increase the power of an experiment by matching the experimental units in each block, say, on age, weight, or location in the animal house. This means that powerful experiments can often be done even though the experimental units are somewhat heterogeneous, providing that matching is possible. This is particularly important with large experiments in which it is often difficult to obtain a sufficiently homogeneous group of animals.

3. Take account of material which has a natural structure. Within-litter experiments are an obvious example.

4. Split the experiment up into smaller bits in order to make it more manageable. This would be useful with large experiments and should help to minimize measurement errors because the work can be done under less time pressure.

5. Increase the external validity of an experiment because each block samples a different environment and/or time period.

Sample Size Determination

Either a power analysis or, for smaller more fundamental studies, the resource equation method can be used to set sample size. A power analysis requires an estimate of the standard deviation (for measurement variables). As the magnitude of the likely block effect and the standard deviation are usually unknown, the estimate will probably have to come from an unblocked (completely randomized) experiment. The payoff will be taken in increased power (experience has shown that RB designs are nearly always more powerful than completely randomized designs of the same size, with the possible exception of very small experiments). However, when similar blocked experiments are done frequently, the estimate of the standard deviation from these could be used in future power analyses to estimate the sample size.

The resource equation method (Mead 1988) is

\[
E = \frac{t (t - 1)(b - 1)}{2b}
\]

where \(E\) denotes the expected cost, \(t\) is the number of treatments, and \(b\) is the number of blocks. Randomization is done separately within each block. An individual observation will be made up of a grand mean \(\mu\), a deviation “\(i\)” due to the treatment it receives, a deviation “\(b\)” due to the block, and an individual deviation “\(e\)”:

\[
Y_{ij} = \mu + t_i + b_j + e_{ij}
\]

where \(Y_{ij}\) is the individual observation, \(\mu\) represents the overall mean, \(t_i\) represents a deviation due to the \(i^{th}\) treatment, \(b_j\) represents a deviation due to the \(j^{th}\) block, and \(e_{ij}\) represents a random deviation associated with the individual experimental unit. In this case, \(i = 1 \ldots t\) where \(t\) is the number of treatments and \(j = 1 \ldots b\), where \(b\) is the number of blocks.

Assuming a single treatment factor and a single block factor, the experiment is analyzed by a two-way analysis of variance.
without interaction. However, a factorial treatment structure can also be used (as in the example below), so there can be two or more treatment factors. A Latin square design has two blocking factors, often designated “rows” and “columns.”

Note that with these designs no two observations come from the same block and treatment so the estimate of the standard deviation is obtained as the square root of the error mean square in the analysis of variance.

Randomized block designs can sometimes have within-block replication (e.g., two or more experimental units per block and treatment combination), but this is not discussed here.

An Example

The example given below comes from a series of studies that were aimed at exploring the effect of antioxidants on susceptibility to cancer. In this example, diallyl sulphide (DS), a substance found in garlic, was administered by gavage in three daily doses of 0.2 mg/g to 8-week-old female mice of four inbred strains, and the activity of a number of liver enzymes was compared in treated and vehicle-treated controls. This work was carried out in 1993 as part of MAFF Project FS1710 entitled “Mechanisms of modulation of carcinogens by antioxidants: genetic control of the anticarcinogenic response in mice.” The work was done under UK legislation and all animals were humanely euthanized as directed under the Animals (scientific procedures) Act 1986. Were this an original research paper rather than an example, then full details would need to be given according to the Animal Research: Reporting of in vivo Experiments (ARRIVE) guidelines (Kilkenny and Altman 2010).

The purpose of the experiment was to assess whether the activity of the liver enzymes was altered by the DS treatment and to see if there were any important strain differences in response. Altogether there were eight treatment combinations: 4 inbred strains × 2 treatments in a factorial arrangement, in two blocks. Both treatment and strain were regarded as fixed effects, block being a random effect. Note that strain is a classification that cannot be randomized. So in this experiment the only randomization was the decision of which of two mice of each strain within a block would receive the treatment and which would be the control. However, the order in which the animals were sacrificed in each block was randomized.

The work was started when the MRC Toxicology Unit was being relocated from South London to Leicester. The new animal house was not yet ready, and the first block of the experiment was done with the mice housed in a plastic film isolator. The second block was done approximately two months later with the animals housed in one of the new animal rooms, so the two blocks had different environmental conditions although these were not quantified. The determinations of enzyme activity were done separately for each block using freshly made up solutions.

The raw data showing the activity of one of the liver enzymes, glutathione-S-transferase (Gst), assayed using the chlorodinitrobenzene (CDNB) method, is shown in Table 1.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Treatment</th>
<th>Block 1</th>
<th>Block 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH</td>
<td>C</td>
<td>444</td>
<td>764</td>
</tr>
<tr>
<td>NIH</td>
<td>T</td>
<td>614</td>
<td>831</td>
</tr>
<tr>
<td>BALB/c</td>
<td>C</td>
<td>423</td>
<td>586</td>
</tr>
<tr>
<td>BALB/c</td>
<td>T</td>
<td>625</td>
<td>782</td>
</tr>
<tr>
<td>A/J</td>
<td>C</td>
<td>408</td>
<td>609</td>
</tr>
<tr>
<td>A/J</td>
<td>T</td>
<td>856</td>
<td>1002</td>
</tr>
<tr>
<td>129/Ola</td>
<td>C</td>
<td>447</td>
<td>606</td>
</tr>
<tr>
<td>129/Ola</td>
<td>T</td>
<td>719</td>
<td>766</td>
</tr>
<tr>
<td>Block means</td>
<td></td>
<td>567</td>
<td>743</td>
</tr>
</tbody>
</table>

* C = vehicle control; T = Treated with DS.

The units are nmol conjugate formed per minute per mg of protein.

There was a large block effect, with all block A values being lower than the corresponding block B values. Why was this? The protocols were identical. It could have been due to slight differences in calibration of instruments or minor differences in reagents and solutions used to assess the enzyme activity. Possibly the animals supplied were of a slightly different age, had a different microflora, were on a different batch of diet, or perhaps the different environment of isolator versus animal room altered their response. There are many variables that can influence such results, and it is impossible to identify or control them all. What is important is the relative magnitude of each observation. This was maintained as shown by the strong correlation of 0.88 between the two blocks, shown in Figure 1. Large block effects are common in this type of design. They highlight the importance of having concurrent controls and randomization, as well as the danger of using historical data where differences of the sort seen here between blocks might be mistaken for treatment effects.

The analysis of variance (ANOVA; Table 2) shows a large treatment effect, no significant difference between strains \(p = 0.091\) but some evidence of a strain by treatment interaction \(p = 0.028\). When a significant two-way interaction is observed, the individual means need to be looked at separately. These are shown graphically in Figure 2 with strain NIH being less responsive to treatment with DS than the other strains. The residual mean square (2957) provides an estimate of the pooled within-group variance, so the standard deviation is 54.5 units. The experiment is a bit small according to the resource equation, with \(E = 7\). But power has probably been increased by using inbred strains and the randomized block design.

An ANOVA makes three assumptions about the data: (1) The experimental units are independent (i.e., the treatments
have been individually assigned to the experimental units by randomization); (2) The residuals (deviation of each observation from its group mean) have a normal distribution; and (3) The variances are homogeneous. The first assumption is met in this case by randomization. It is good practice to investigate the second and third assumptions using "residual model diagnostic plots," as shown in Figure 3. The top plot is concerned with studying the homogeneity of variances. There should be a scattering of points with no pattern, as is the case here. The lower plot is a normal probability plot of the residuals. If they have a normal distribution, points should lie on a straight line. In this case, there are four points lying off the line, showing that the assumption of a normal distribution of the residuals is slightly marginal. The ANOVA is quite robust to deviations from these assumptions, but a transformation of the scale of measurement can often be used if there is a serious deviation from normality to see if the fit is better. In this case, a log transformation of the raw data provides a slightly better fit (not shown). The analysis of variance of the log-transformed data differs in that there is no statistically significant strain × treatment interaction. It is a matter of judgment whether to use the raw or the log-transformed scales. However, in this case it makes little difference to the conclusions. There is clearly a strong treatment effect with no large strain differences. Even if the strains differed slightly in response, that magnitude of difference is unlikely to be of much biological significance. It is also worth noting that the ANOVA provides a single overall test of whether there is a treatment, strain, and interaction effect. But that is three tests. Applying a Bonferroni correction would imply that a p value of 0.05/3 = 0.017 should be used, in which case the interaction would not be "significant." But it is the overall interpretation that is important. That rarely depends on such minute details of the statistical analysis. Even a well-designed experiment will give slightly different results if it is repeated. If it were important to find strain differences in Gst levels or in the Gst response to chemicals, which was not a purpose in this case,

Table 2 The analysis of variance of the data shown in Table 1. By convention, p values of less than 0.05 are considered "statistically significant" so there is no evidence of strain differences in mean Gst levels, but some evidence of strain differences in response, but also see text

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks</td>
<td>1</td>
<td>1242</td>
<td>1242</td>
<td>42.01</td>
<td>&lt;0.000</td>
</tr>
<tr>
<td>Strain</td>
<td>3</td>
<td>286</td>
<td>95</td>
<td>3.22</td>
<td>0.091</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>2275</td>
<td>2275</td>
<td>76.93</td>
<td>&lt;0.000</td>
</tr>
<tr>
<td>Strain × treatment</td>
<td>3</td>
<td>495</td>
<td>165</td>
<td>5.58</td>
<td>0.028</td>
</tr>
<tr>
<td>Residuals</td>
<td>7</td>
<td>207</td>
<td>29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: Df, degrees of freedom; Sum Sq, sums of squares, Mean Sq, mean square, F is a test statistic, P is the p-value; the probability that the observed differences could have arisen by chance in the absence of a true treatment or strain effect.

...
then further work would be needed with a different experimental design and a wider choice of strains.

Discussion

The purpose of this paper is to bring randomized block experimental designs to the attention of scientists using laboratory animals. Because of their valuable properties, these designs have been widely used in agricultural and industrial research for many decades. Mead (1988), who has experience in medicine, agriculture, and industry complains that about 85% of all experiments are randomized complete block designs, and suggests that investigators should be more flexible in their choice of design. Yet these designs are rarely used in experiments involving laboratory animals. This cannot be because they are unsuitable. There is nothing about research with laboratory animals that sets it apart from all other disciplines. The reason must be that scientists are unfamiliar with these designs. Mead also says that a statistician should be fully involved with a research scientist in designing his/her experiments and that “this is the only efficient approach to designing experiments.” Yet in the last 50 years few statisticians of stature have been closely involved (e.g., published papers or books) in this area of research. The failure over such a long period of time to use the most efficient designs must surely have led to a serious waste of animals, time, and other scientific resources.

The example experiment, to determine the effect of DS on Gst levels, could have been done on a single occasion using 16 outbred CD-1 mice assigned at random to the two treatments. But it would have given neither an indication of whether there was a genetic component to the response nor whether the results were repeatable. The good agreement between the two blocks, separated by two months and a different animal house environment, suggests that the results are likely to be robust. There was no evidence of strain differences in Gst levels, although there was a hint of strain differences in the treatment response, depending on the scale of measurement. Altogether, the randomized block design gave extra information and had higher external validity at virtually no extra cost, with some assurance that the results should be reproducible.

Conclusions

Randomized block experimental designs include within-subject, crossover, and matched designs in which the experimental material is split up into a number of mini-experiments that are combined in the statistical analysis. They are widely used in many other research disciplines and, because of their useful properties, should be more widely used in laboratory animal research. They can be more convenient, more powerful, and can use fewer animals than completely randomized designs. If the blocks are separated in time and there is good agreement between them, then this gives some assurance that the experiment is reproducible. Their more widespread use would save money, animals, and other scientific resources and would speed up the development of new treatments for diseases of humans and animals.

References