The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK

Animal Welfare 2005, 14: 361-365
ISSN 0962-7286

An illustration that statistical design mitigates environmental variation and ensures unambiguous study conclusions

KH Gore and PJ Stanley*

Pfizer Global Research and Development, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK
* Contact for correspondence and requests for reprints: phil.stanley@pfizer.com

Abstract

This paper highlights the essential need for appropriate statistical design and randomisation in laboratory animal studies. Using an example of a 21 day weight gain study in mice, we show that without the use of an appropriate statistical design and randomisation, incorrect conclusions may have been drawn. We used an experimental design that allowed comparisons to be made between five treatments that were free from systematic error. Two alternative designs that are practically attractive, yet had no statistical basis, are also described in this paper and the potentially incorrect conclusions highlighted. The use of appropriate statistical design is ethical because it results in clear, unambiguous conclusions. Conclusions that may be biased or ambiguous will require verification by further research and this, in the long term, is contrary to the reduction element of the Three Rs.

Keywords: *animal welfare, environmental variation, Latin Square, reduction, statistical design, Three Rs*

Introduction

Since Russell and Burch (1959, reprinted 1992) publicised the need for replacement, reduction and refinement, the requirement for statistical input into animal studies has become apparent (Festing *et al* 2002). Statistical input is often used only to reassure the researcher that the number of animals used is sufficient; however, statistics has a much more significant role to play in the Three Rs than just sample size estimation and reducing the number of animals used in a single study. Good statistical design and randomisation ensures that comparisons between treatments are free from systematic error (Cox 1958), and assures the ethical use of animals; choosing the wrong design, or a complete lack of design, can lead to incorrect conclusions.

To illustrate the importance of appropriate statistical design, we used an example of a 21 day weight gain study in female CD–1 mice (*Mus musculus*), which investigated the effect of a control vehicle and a test compound administered at four different doses. In this study, mice were housed singly in cages across three racks using a Latin Square design in each rack, thereby ensuring that all five treatment groups were represented in each row and column of every rack. Differences between the racks, and between the rows within the racks, were discovered. If a housing design had been used which did not allow these differences to be identified, it would not have been possible to say with any certainty whether the results were solely due to the doses of the compound or to the position of the cage. Consequently, the conclusions obtained from the study may have been misleading. Experimental designs in which treatments are systematically arranged appear practically attractive to minimise dosing errors, yet they can cause treatment effects to be indistinguishable from positional effects or other environmental factors. We show that correct statistical design contributes to the reduction element of the Three Rs as much as the use of sample sizing techniques.

Materials and methods

All animal experiments were conducted in compliance with national legislation and the relevant Codes of Practice in the UK, and were approved by the local ethical review process. All the animals used were of high health status, consistent with the Federation of European Laboratory Animal Science Associations (FELASA) recommendations, and our ongoing health maintenance programme did not suggest the introduction of any contaminants.

Seventy-five adult female outbred Swiss albino mice (Crl:CD–1 [ICR]: Charles River Laboratories UK Ltd, Margate, UK), weighing 20–40 g, were randomly divided into five equally sized groups. One of five treatments was randomly assigned to each group. The five treatment groups consisted of a control vehicle group and four dosed groups administered with a test compound at 1, 3, 10 or 30 mg kg^{-1}. The mice were group housed until the start of the study and then singly housed, in Techniplast mouse experimental cages (Cage 1264C: Techniplast UK Ltd, Kettering, UK), with solid-bottom floors and a floor area of 410 cm^2. During the study the mice were singly housed to ensure that food and water consumption could be accurately determined for each individual mouse, as these were experimental parameters in the study. All mice had access to environmental enrichment aids consisting of Enviro-dri paper bedding and

https://doi.org/10.1017/S0962728600029699 Published online by Cambridge University Press

Table 1 Mean rate of water consumption (ml day^{-1}) and standard deviation (SD) for each treatment group.

	Vehicle	1 mg kg^{-1} day^{-1}	3 mg kg^{-1} day^{-1}	10 mg kg^{-1} day^{-1}	30 mg kg^{-1} day^{-1}
Mean rate (ml day^{-1})	0.03	0.42	0.54	0.64	0.77
SD	0.67	0.53	0.60	0.73	0.79

Table 2 Mean rate of water consumption (ml day^{-1}) and standard deviation (SD) for each rack.

	Rack 1	Rack 2	Rack 3
Mean rate (ml day^{-1})	1.03	0.16	0.26
SD	0.75	0.48	0.52

a cardboard 'dome home'. The cages were stored across three racks, with each rack having space for 30 cages arranged in a 5 × 6 grid. The mice had access to bottled water *ad libitum* and were fed on Rat and Mouse No 1 Maintenance diet (RM1 [E]: SQC, Special Diet Services, Witham, Essex, UK).

Environmental conditions were monitored and controlled at 21°C and 55% relative humidity. Lighting was monitored and controlled using a 12:12 h light:dark cycle (lights on at 0500h). All environmental conditions were validated and determined to be within the original, stringent, engineering specifications; for example, the heat distribution was assessed through the gas decay method.

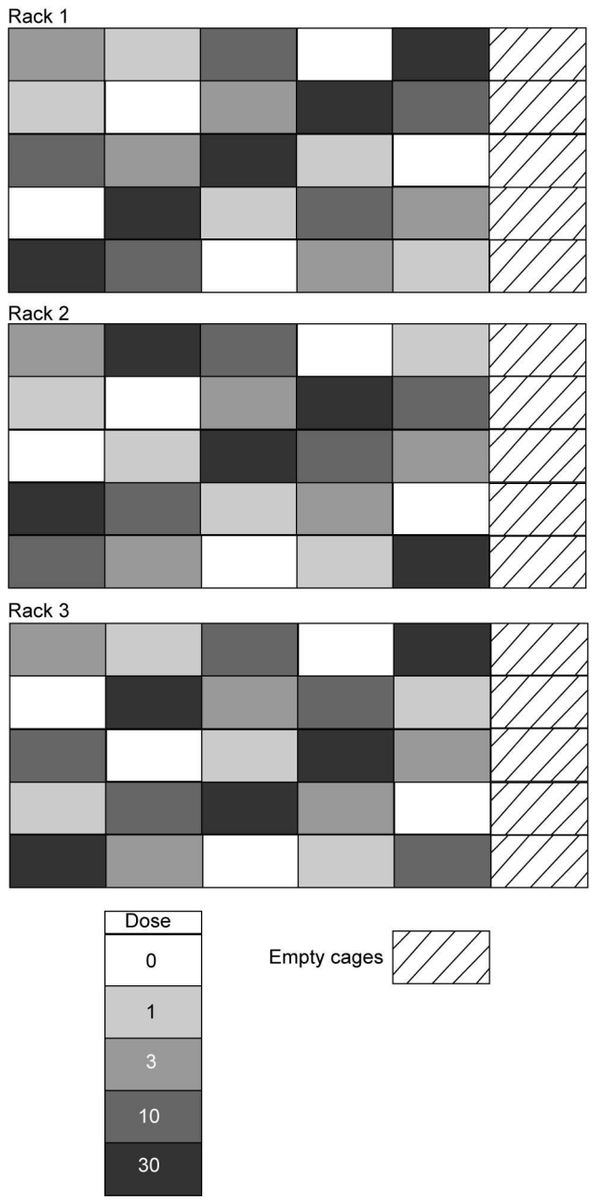
In the week prior to beginning the study, food consumption, water consumption and weight gain were measured on days –6, –3 and –1. During the study, animals were orally dosed twice per day (at 0700h and 1600h) for 21 days, with either the control vehicle or a single dose of the test compound. The animals were observed daily for any changes in their appearance and behaviour. Body weight was also measured daily, whereas food and water consumption were determined every Monday, Wednesday and Friday. Body temperature was measured orally at the same time on each day using a digital thermometer (Comark model 2001: Comark Ltd, Stevenage, UK). On the morning of day 21, animals were euthanased, using exsanguination under anaesthesia, and the perirenal and parametrial fat pads and the gastrocnemius muscle were removed and weighed to give an indication of lean body mass.

Statistical design

Following standard sample size calculations (Steel & Torrie 1980), a group size of 15 animals per treatment was chosen as the appropriate number of animals to detect a 7% decrease in body weight in comparison with control animals, at the 5% significance level with 80% power.

The 75 mice were arranged across three racks of 25 cages; the far right column of each rack was left empty. Within each rack, the five treatment groups were arranged using a randomised 5 × 5 Latin Square design, which ensured that one cage from each of the five treatment groups was placed in each row and column of every rack. The arrangement used can be seen in Figure 1.

Figure 1



The experimental housing design used for the mice cages, using a randomised 5 × 5 Latin Square design on each rack.

https://doi.org/10.1017/S0962728600029699 Published online by Cambridge University Press

Table 3 Mean body temperature (°C) and standard deviation (SD) for each treatment group.

	Vehicle	1 mg kg^{-1} day^{-1}	3 mg kg^{-1} day^{-1}	10 mg kg^{-1} day^{-1}	30 mg kg^{-1} day^{-1}
Mean body temperature (°C)	37.9	38.1	37.8	38.1	38.1
SD	0.46	0.80	0.85	0.68	0.57

Table 4 Mean body temperature (°C) and standard deviation (SD) for each row.

	Top row	Row 2	Row 3	Row 4	Bottom row
Mean body temperature (°C)	37.2	38.0	38.4	38.3	38.1
SD	0.67	0.64	0.36	0.58	0.49

Statistical analysis

In this study, the main statistical comparisons of interest were the effects exhibited by the four dosed groups compared to the control vehicle group. There were four primary measures of interest: the rates of change in body weight; food intake; feed efficiency (g of body weight gain per g of food consumed); and water consumption. Linear regression was performed on the 21 day data from each animal for each of the four primary measures. The slope of the linear regression of the response over time gave the rate of change for each animal. The calculated rates for all primary measures were analysed using an analysis of variance for a replicated Latin Square, taking into account the rack number, and allowing for different columns and common rows (height) across racks. The secondary measures of interest were: the weight of perirenal fat pads; the weight of parametrial fat pads; the weight of gastrocnemius muscle; and the body temperature on day 20. These were also analysed using an analysis of variance for a replicated Latin Square, taking into account the rack number, and allowing for different columns and common rows (height) across racks. Where the overall F test for the differences among all five treatments in the analysis of variance was statistically significant ($P < 0.05$), the differences between dose and vehicle were tested with t-tests using a pooled estimate of the variance from each analysis of variance. Estimated treatment means (results for the primary and secondary measures of interest averaged across animals within each treatment group), differences between the four dosed groups and the vehicle group, and 95% confidence intervals of the differences were also calculated but are not presented in this paper.

The randomised 5 × 5 Latin Square design used on each rack allowed the differences between the racks, and the positional effects of rows and columns, to be investigated in addition to the comparison between the five treatments. However, the objective of the study was to investigate treatment effects, not environmental effects, and so no formal statistical testing was performed on these environmental effects. Any conclusions drawn were purely observational. All interesting findings are presented with means and standard deviations. All statistical analyses were performed using GenStat® for Windows, version 6.1 (VSN International Ltd, Hemel Hempstead, UK).

Results

The aim of this paper is to highlight the impact that the lack of statistical design and randomisation would have on any study conclusions; therefore, only a brief summary of the key results of the original study is reported below and no interpretation of the results is given.

Comparison of treatment effects

Analysis of the primary measures revealed that there were no statistically significant differences between the five treatment groups for the rates of change in body weight ($F_{4,50} = 0.89$, $P > 0.1$), food consumption ($F_{4,50} = 1.31$, $P > 0.1$), or feed efficiency ($F_{4,50} = 0.87$, $P > 0.1$). However, all four dosed groups exhibited statistically significantly greater rates of water consumption compared with the vehicle group ($P < 0.05$ for all, based on the individual t-tests).

In the analysis of the secondary measures, there were no statistically significant differences between the five treatment groups in the weight of perirenal fat pads ($F_{4,50} = 0.18$, $P > 0.1$), the weight of parametrial fat pads ($F_{4,50} = 0.22$, $P > 0.1$), the weight of gastrocnemius muscle ($F_{4,50} = 0.89$, $P > 0.1$), or body temperature at day 20 ($F_{4,50} = 1.36$, $P > 0.1$).

Investigation of the environmental variation

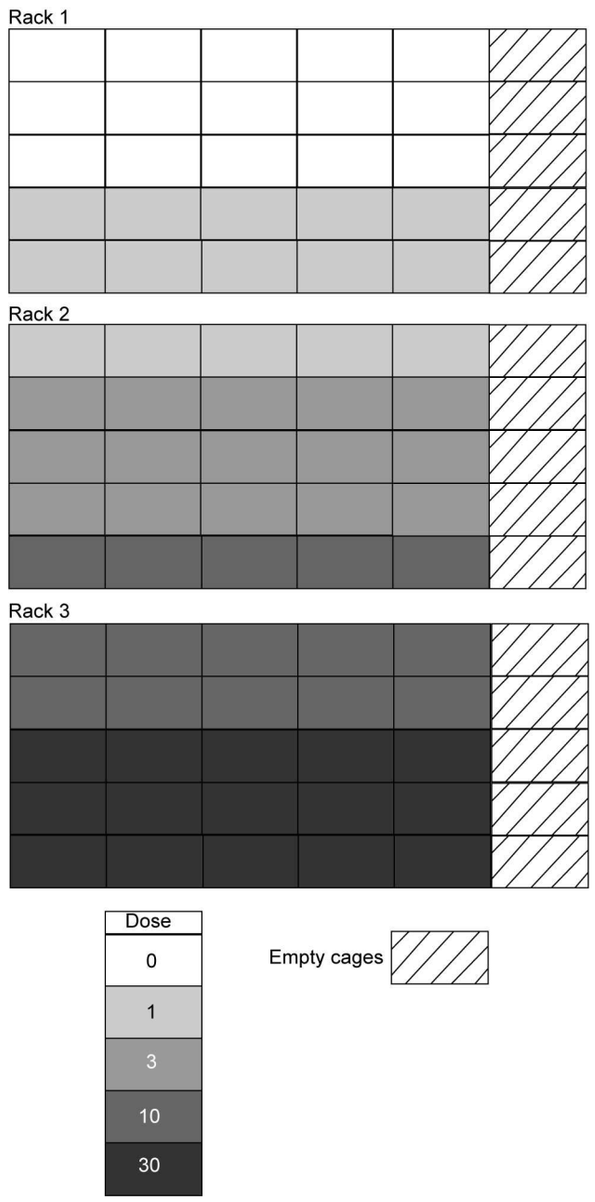
When investigating the environmental effects of rack, row and column position, two further, very interesting findings were observed.

1. In the analysis of the rate of water consumption not only were there statistically significant differences between the five treatments but also there was a clear difference between the racks. Water consumption increased at a higher rate over the 21 days in rack 1 compared with racks 2 and 3. Table 1 shows the estimated mean rate of water consumption and standard deviation for each treatment group and demonstrates how the water consumption rates increased with dose. Table 2 shows the mean rate of water consumption and standard deviation in each rack.

2. Analysis of the body temperatures revealed that there were no statistically significant differences between the five treatments, yet the average temperatures differed with height in the rack. Mice in the lowest rows of the racks exhibited average body temperatures approximately 1°C

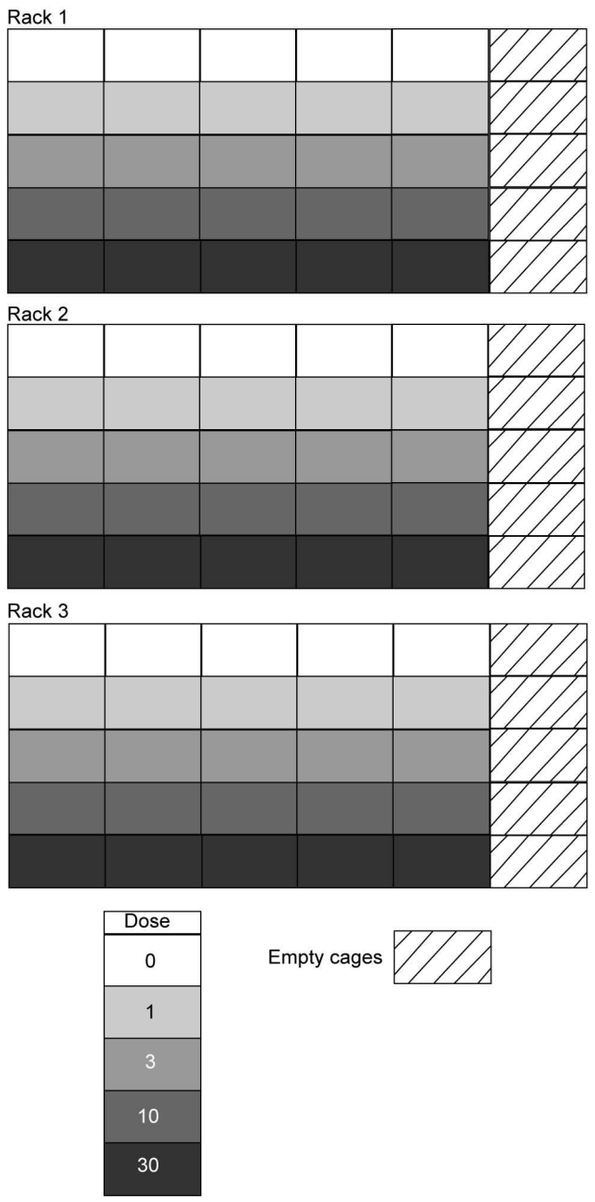
https://doi.org/10.1017/S0962728600029699 Published online by Cambridge University Press

Figure 2



Alternative housing design 1 — doses systematically arranged across the racks.

Figure 3



Alternative housing design 2 — doses systematically arranged in rows on each rack.

higher than mice in the highest rows. Table 3 shows the estimated mean body temperature and standard deviation for each treatment group and demonstrates the consistency of the body temperatures across the treatment groups. Table 4 shows the mean body temperature (°C) and standard deviation for each row across the racks.

It was only possible to detect these environmental effects because of the use of the Latin Square design. For the technical details of how this is achieved using the Latin Square design, the reader is referred to any standard statistical text book, for example Cochran and Cox (1957, Chapter 4). The implications of these findings are discussed in the next section.

Discussion

To emphasise the importance of design and randomisation, this discussion describes two alternative, practically attractive, housing design options and the impact they could have on the treatment comparisons.

Arranging treatments in racks

Practical considerations often influence the layout of animal housing. When animals are dosed daily it may be more practical to systematically arrange the dose groups to minimise dosing errors. Figure 2 illustrates such a design where the racks are filled up systematically. All 15 vehicle group mice are placed on rack 1 and the spaces

https://doi.org/10.1017/S0962728600029699 Published online by Cambridge University Press

are then filled systematically across the racks with each treatment group in turn.

In this housing design the treatment comparisons are inseparable from the potential differences between the racks. When estimating the difference between the vehicle group and the highest dose group we are also including the difference between rack 1 and rack 3. The impact of this housing design is clear when you consider the first of our two environmental findings: *water consumption increased at a higher rate over the 21 days in rack 1 compared with the other two racks*. If we had run the study using the housing design outlined in Figure 2, and in the analysis of water consumption performed the comparison of vehicle and the highest dose group, we could have incorrectly concluded that the highest dose group exhibited a *significantly smaller increase* in water consumption than the vehicle group. This conclusion is purely due to the water consumption being different in the different racks — it is not a difference caused by the test compound. In our actual study, having allowed for environmental effects, we found that the highest dose group exhibited a much *greater increase* in water consumption than the vehicle group, as shown in Table 1.

Arranging treatments in rows

Figure 3 illustrates another practically simple housing design where the five vehicle group mice are housed on the top row of every rack, five from the lowest dose group on the row immediately beneath the vehicle group and so on with the highest dose group being placed on the lowest rows. Five mice from all five treatment groups are still housed on each rack, but the arrangement within a rack is very systematic.

In this housing design the treatment comparisons are inseparable from the potential height differences. When comparing the difference between the vehicle group and the highest dose group we are also comparing the top and bottom rows of the racks. The impact of this is clear when you consider the second of our two environmental findings: *mice in the lowest rows of the racks exhibited average body temperatures approximately 1°C higher than mice in the highest rows*. If we had run the study using the housing design outlined in Figure 3, and in the analysis of body temperatures performed the comparison of vehicle and the highest dose, we could have incorrectly concluded that the highest dose *significantly increased* body temperature. This conclusion is purely due to height in the rack — it is not a change caused by the test compound. In our actual study, having allowed for environmental effects, we found that the average body temperatures were *very similar* in all five treatment groups, as shown in Table 3.

Conclusions

We have shown that different study conclusions would have been made regarding the effects of the test compound depending upon the housing arrangement of the treatment groups in the study. The two alternative arrangements discussed would have led to incorrect conclusions with respect to either changes in water consumption or body temperature.

It is important that statisticians and scientists fully understand the animal housing and environment if an optimised design is to be generated. The existence and origin of environmental variation is not obvious but can greatly influence the study outcome if ignored. In our study, uniformity of the housing and the environment was not assumed, even though the caging and racking were standardised and the room fully validated to stringent engineering specifications. There are many possible sources of environmental variation, including measurement, technical, equipment and human error. However, we have shown that environmental variation, irrespective of the source, can be mitigated by using a simple statistically optimal design or layout.

Animal welfare implications

The reduction element of the Three Rs, and involvement of statisticians and statistical thinking, is sometimes seen as purely getting the sample size for a single study to be as small as possible. However, reduction is more about choosing an appropriate sample size so that you run the right sized study and the correct programme of work. Furthermore, it is about running a study in which your treatment comparisons are free from any potential systematic error so that you can have well placed confidence in the results.

Running poorly designed studies, such as the alternative theoretical designs presented in this paper, is against the principles of the Three Rs. In our examples, both alternative layouts would have led to more mice being used, either through recognising the incorrect conclusions and running another weight gain study, or through not recognising the error and performing further studies investigating effects that had nothing to do with the compound of interest.

Acknowledgements

We are grateful to Simon Lewis for the collaboration in the development of this study and for allowing us to use the data. We also wish to acknowledge the Animal Resources staff for their excellent technical assistance and care of the animals.

References

Cochran WG and Cox GM 1957 *Experimental Designs, 2nd Edition*. John Wiley & Son: New York, USA

Cox DR 1958 *Planning of Experiments*. John Wiley & Son: New York, USA

Festing MFW, Overend P, Gaines Das R, Cortina Borja M and Berdoy M 2002 *The Design of Animal Experiments. Laboratory Animal Handbooks Number 14*. RSM Press: London, UK

Russell WMS and Burch RL 1959 (reprinted 1992) *The Principles of Humane Experimental Technique*. Universities Federation for Animal Welfare: Wheathampstead, UK

Steel RGD and Torrie JH 1980 *Principles and Procedures of Statistics, 2nd Edition*. McGraw-Hill: New York, USA