The Need to Improve Experimental Design

Derek J. Fry

Good experimental design is a valuable strategy to address the reduction and refinement principles of the Three Rs.

Among the Three Rs, Reduction seems the one which has least progress to record. This may partly be because some principles of good experimental design were established long ago. Some were codified in 1831, when Marshall Hall published his “Principles of Investigation in Physiology”, which emphasise the importance of having clear objectives, of minimising severity, and of using observation to avoid invasive procedures. Advances in experimental design since then are eclipsed by great advances in anaesthesia, analgesia and in understanding of animal structure, function and behaviour. The scope for the control of severity in experiments, for reliable measurements, for more-precise formulation of experimental questions and better detection of experimental effects has enormously improved. In particular over the 50 years since Russell and Burch developed the Three Rs concept one can trace its spread to worldwide acceptance and see significant improvements in many areas. As well as advances in anaesthesia and analgesic regimes, and much better understanding of animal behaviour, there has been much progress in recognition of pain and distress, a change in attitude among researchers from animals being merely data-tools, and wide acceptance of the need for ethical evaluation. However, the major advances in experimental design developed in the 1920s and 30s have been slow to spread through biomedical studies, and experimental design of animal studies tends to consider only one R — Reduction — and omit consideration of the other two, particularly Refinement.

Over 50 years ago, writing in the UFAW handbook of 1957, Hume commented “techniques for designing experiments on the basis of small-sample theory are available” but “one still sees unjustifiably large samples reported”. He was referring principally to the methods in Fisher’s 1935 work, The Design of Experiments. Sadly, one could say the same today, with a number of recent papers citing the quality of studies as a likely major reason for difficulties in reproducing animal studies, and for the lack of correlation between their findings and results in the clinic. Also, the survey by Kilkenny et al. in 2009 considered that “a large number of the studies assessed did not make the most efficient use of the available resources (including the animals), by using the most appropriate experimental design”.

Although there is some concern that researchers are not clear in the objective of an experiment (this was judged to be the case in 5% of studies in the Kilkenny et al. survey), a much greater problem seems to be that they are not adhering to the fundamental principles which are necessary to ensure that valid comparisons can be made between groups under different experimental conditions. These can be summarised as independent replication, randomisation and control (see Figure 1). To this list may be added ‘blinding’ (concealing the experimental treatment from those allocating experimental material to it, and also from those assessing the outcome). ‘Blinding’ guards against subjective bias, and is particularly important when effects are small. Without adherence to the fundamental principles, the results of an experiment are unreliable and statistical testing is illusory.

Independent replication is essential because of biological variation. It allows the extent of variability to be estimated, and that estimate is the basis for many statistical tests. It means allocating a number of animals or other experimental comparison units to each set of experimental conditions. But the units do have to be truly independent. Placing all test plates in one incubator and control plates in another risks confounding a difference between the incubators as a treatment effect. Similarly, putting ten mice all undergoing one treatment into two cages and another ten all undergoing a different treatment in another two cages superimposes cage effects on any treatment effects. Here, the ”experimental unit” (defined by Festing et al. as “the unit of replication that can be assigned at random to a treatment”) is strictly-speaking the cage and there are only two replicates for each treatment. Commonly, the ten individuals are taken as independent, without taking account of possible cage effects, so the estimates of power and the derived p values may be erroneous.

Randomisation is of crucial importance as it avoids an inherent difference being confounded with the effect of an experimental treatment. A good way of randomising is to give the animals (or other experimental units) numbers, then to use a computer to put the numbers in random order, and then use the experimental units in that order.
Suitable comparisons or controls are essential for the proper interpretation of results. Experimenters should think what the outcomes of an experiment might be, then think what the possible interpretations could be, and then think what controls are needed to avoid misinterpretation. Discussions with experimenters usually indicate that a negative control is routinely included in their experimental design, but other controls are not considered. Importantly, they forget positive controls, which are essential if there are possible causes for lack of effect in an experimental group other than that the treatment is ineffective — a change in animal susceptibility, for example.

The other major failure in experimental design is in the choice of type of design. Often fully randomised group comparison designs are used when there are actually more-efficient arrangements. Use of these more efficient arrangements would result in the gaining of more data and/or the use of fewer animals. Experimental units can be matched into sets according to a characteristic liable to contribute noticeably to variability, such as animal weight-range, age or parentage, cage position in a rack, or the time of starting the experimental procedures. With this ‘blocking’ arrangement, the variability between the sets can then be separately estimated and distinguished from individual variation or that due to treatment. This can provide a more-precise estimate of the effect of treatment, so the ability to detect a real effect is enhanced, and fewer experimental units are needed overall.

Another under-utilised approach is factorial arrangement, in which each experimental group is composed of experimental units with known but different characteristics — both male and female animals, for example, or animals of different strains or age bands. In such designs, biological variation is estimated by using all the individual units and the number in each treatment-characteristic sub-group can be small. Variability due to the characteristic can be separately estimated from variation due to treatment, and overall numbers are much reduced when compared to repeating the treatments for each characteristic (sex, age, strain and so on).

So how should we address what seems to be a widespread failure to keep to important principles and use advances in design developed in the 1930s? One approach that shows signs of proving successful is education in experimental design by means of workshops of a particular pattern. That pattern involves a mix of information provision with group problem-solving which uses that information, and with discussion with experts in experimental design, statistical testing and control of severity. The comments here come from experience with one- to four-day workshops of this type, run in several countries since 2008. Indications of the need for such education come from the continuing demand for places, from workshop participant comments (e.g. “general statistics courses are too theoretical and too far away from my situation/experiments”), and from survey responses. Workshops can emphasise the important fundamentals and provide opportunities to consider a range of types of design. They can also consider refinement, alerting to the effects the severity of procedures and animal distress may have on the results, and suggest how to design for minimal severity.

Participants’ comments and external assessments of the workshops rate them as informative, educational and enjoyable. Consistently, well over 90% of the participants agree that the workshops exposed them to new knowledge and practices. The routine testing of the knowledge and understanding of the participants has to date always shown a marked improvement between pre-tests and post-tests. The results from a three-and-a-half-day course in an EU country, shown in Figure 2a, are typical of the changes in particular understandings recorded. The distributions of pre- and post-test scores shown in Figure 2b are from a two-day course in an Asian country. They illustrate the marked shift to higher scoring found so far in all of the workshops. These tests are more like quizzes than examinations, and some element of the improved scores could be improved understanding of the quiz format. However, the greater confidence with which the concepts become part of the group discussions and questioning of experts later in the workshops also show appreciation of the concepts and improved understanding of when different types of design might be used. The long-term influence of these workshops has still to be evaluated, but some evidence of long-term effects was picked up in a survey by Howard et al. When ‘change in attitude’ was tested, as it was in one RSPCA-led workshop run in an Asian country, a distinct shift was seen. Figure 3 indicates the shift found with one of the questions asked. It shows the numbers of workshop participants giving different levels of agreement with the statement “One animal

Figure 1: Basic principles in designing experiments

Independent replication — repeating the ‘treatment’, i.e. the particular set of experimental conditions, with a number of independent biological units

Randomisation — arranging that inherent differences in biological units or the measurement process are equally likely to occur with any of the experimental groups

Control — including comparisons which allow valid interpretation of results obtained under different experimental conditions
per cage should be the routine practice for housing rats and mice undergoing experiments”. At the end of the workshop the level of disagreement with routine single-housing had markedly increased.

These workshops have the advantage of providing some limited evidence of effectiveness, but they are only one of a number of approaches to providing education in experimental design. Experimental design has been included in RSPCA courses that also cover refinement, for example. The taught component of certain biomedical, biological and agricultural postgraduate courses in universities will include it, and it is a required element in UK project licensee training and FELASA training for “persons responsible for directing animal experiments”. These would all be expected to have an impact. The concerns raised in publications such as references 5–9, and meeting publication guidelines such as the Gold Standard Publication Checklist or the ARRIVE guidelines, should also focus attention on improving the design of animal experiments.

However, feedback from former workshop attendees and comments from researchers in many countries indicate that there is still much to be done. Workshop participants return to supervisors who are unwilling to change their time-honoured approaches, and submitted papers meet referees or editors who expect at least six animals per group, or are unfamiliar with blocking or factorial designs. The level of

---

**Figure 2: Pre-test and post-test results**

- **Understanding tested**
  - Replication
  - Controls
  - Randomisation
  - Experimental unit
  - Factorial design
  - Large control group design

**a)** Percentage of correct responses to questions testing understanding of basic principles and use of different types of design. The differences between pre- and post-test responses for this EU-based course are typical of those found.

**b)** Pre- and post-test scores. The difference in the histograms for pre- and post-test scores in this course run in an Asian-country is typical of those found.

**Figure 3: Change in attitude of participants**

Workshop participants were asked to indicate their level of agreement with the statement “One animal per cage should be the routine practice for housing rats and mice undergoing experiments”.

---

Workshop participants were asked to indicate their level of agreement with the statement “One animal per cage should be the routine practice for housing rats and mice undergoing experiments”. At the end of the workshop the level of disagreement with routine single-housing had markedly increased.

These workshops have the advantage of providing some limited evidence of effectiveness, but they are only one of a number of approaches to providing education in experimental design. Experimental design has been included in RSPCA courses that also cover refinement, for example. The taught component of certain biomedical, biological and agricultural postgraduate courses in universities will include it, and it is a required element in UK project licensee training and FELASA training for “persons responsible for directing animal experiments”. These would all be expected to have an impact. The concerns raised in publications such as references 5–9, and meeting publication guidelines such as the Gold Standard Publication Checklist or the ARRIVE guidelines, should also focus attention on improving the design of animal experiments.

However, feedback from former workshop attendees and comments from researchers in many countries indicate that there is still much to be done. Workshop participants return to supervisors who are unwilling to change their time-honoured approaches, and submitted papers meet referees or editors who expect at least six animals per group, or are unfamiliar with blocking or factorial designs. The level of
knowledge and confidence reached during a workshop, an undergraduate or postgraduate module, or licensee or FELASA training may well be insufficient of itself to sustain arguments for designs unfamiliar to the home laboratory or referees.

So there is a challenge here for all involved with animal experiments to be open to possibilities for improving design. It would be very sad if, in another 50 years, the comments of Hume, quoted above, still applied.

Dr Derek J. Fry
Faculty of Life Sciences
University of Manchester
Oxford Road
Manchester M13 9PT
UK
E-mail: dj@fry39.fsnet.co.uk

References


12 Hooijmans, C.J., Leenaars, M. & Ritskes-Hoitinga, M. (2010). A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. ATLA 38, 167-182.