Teaching Experimental Design

Derek J. Fry

Abstract

Awareness of poor design and published concerns over study quality stimulated the development of courses on experimental design intended to improve matters. This article describes some of the thinking behind these courses and how the topics can be presented in a variety of formats. The premises are that education in experimental design should be undertaken with an awareness of educational principles, of how adults learn, and of the particular topics in the subject that need emphasis. For those using laboratory animals, it should include ethical considerations, particularly severity issues, and accommodate learners not confident with mathematics. Basic principles, explanation of fully randomized, randomized block, and factorial designs, and discussion of how to size an experiment form the minimum set of topics. A problem-solving approach can help develop the skills of deciding what are correct experimental units and suitable controls in different experimental scenarios, identifying when an experiment has not been properly randomized or blinded, and selecting the most efficient design for particular experimental situations. Content, pace, and presentation should suit the audience and time available, and variety both within a presentation and in ways of interacting with those being taught is likely to be effective. Details are given of a three-day course based on these ideas, which has been rated informative, educational, and enjoyable, and can form a postgraduate module. It has oral presentations reinforced by group exercises and discussions based on realistic problems, and computer exercises which include some analysis. Other case studies consider a half-day format and a module for animal technicians.

Key Words: animal experiments; statistics; education and training; ethics of animal use; experimental design; course content; delivery methods; assessment

Introduction

There is concern over the quality of biomedical studies, and it appears that good design and understanding of the fundamental principles of experimental design cannot be assumed.

As early as 1957, it was pointed out that the efficient designs developed by Fisher and other statisticians in the 1930s were not being widely used by animal researchers (Hume 1957), and this is still a problem. A survey by Kilkenney and colleagues (2009) of 271 papers in a variety of journals noted that only 62% of studies that could have used factorial designs did so and commented “it seems 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.” Even when factorial designs are used, they may not be properly analyzed (Nieuwenhuis et al. 2011).

Many reviews have raised concerns over the quality of studies in a variety of fields. The review by Hurlbert (1984) indicated nearly one-half of the ecological papers surveyed were faulty in that experimental entities were taken as independent when they were not (pseudoreplication) and inappropriate statistical analysis was used; more recently, Lazic (2010) has voiced concerns about pseudoreplication in neuroscience. Kilkenney and colleagues (2009) pointed out many faults in the reporting and design of experiments, bringing to wider attention concerns shown in earlier surveys. Many authors consider there may be problems of inadequate randomization or blinding with a bias in published results the extent of which is unknown. This could explain a preponderance of positive outcomes (Bebarta et al. 2003), poor correspondence between animal experiments and clinical findings (Perel et al. 2006), the variation in reported effects (Scott et al. 2008), and difficulty in repeating published studies in-house (Begley and Ellis 2012; Prinz et al. 2011).

It seems that poor design may be widespread, and teaching in experimental design needs to address this. This was the stimulus to develop courses in experimental design that might improve understanding and help researchers recognize and avoid the problems. The success of these led to a number of derivative courses that have now been run and well received in various countries in Europe, Asia, and America. This article considers how others might develop similar courses. It is principally about teaching experimental design to those involved in work with laboratory or agricultural

Derek J. Fry, MA DPhil MB BCh, is an Honorary Senior Lecturer in the Faculty of Life Sciences at the University of Manchester, Manchester, United Kingdom.

Address correspondence and reprint requests to Dr. D. Fry, Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, United Kingdom or email at derek.fry@manchester.ac.uk.
animals or tissues from them and is aimed at teachers wishing to equip junior researchers to take a critical approach to the current practices they encounter, at those wanting to provide courses that enable participants who have experienced generalized instruction in experimental design or statistics to apply this to animal work, and at staff recruited to provide education in experimental design who are considering what could usefully be included in their courses and how it might be delivered.

The view taken is that teaching experimental design is more education for understanding so that informed choices can be made than training to become competent at a task. In the educational field, it is customary to consider learning outcomes and educational objectives within the different categories of knowledge, understanding/concepts, skills, and attitudes (Cohen and Manion 1983). For experimental design, where it is easy to identify the knowledge to be conveyed, this classification encourages thought as to what skills are being acquired and what attitudes are being engendered.

It is helpful for a tutor to be aware of learning theory, or rather the range of theories of learning (see http://cmapspublic3.ihmc.us/rid=1LGVGJY66-CCD5CZ-12G3/ Learning%20Theory.cmap [accessed on March 23, 2014]), and a number of universities provide background in this for new staff which is publicly available (e.g. Newcastle: http://www.ncl.ac.uk/quilt/resources/teaching/theory.htm [accessed on March 23, 2014]; Stanford: http://www.stanford.edu/class/ed269/hplintrochapter.pdf [accessed on March 23, 2014]). Individuals have different learning styles, and education needs to accommodate both the active learner who will seek out information and the passive one who looks for it being provided; both the rigid learners who tend to stick to a learning strategy that they have found suits them and the flexible learners who can use different approaches; both those who readily proceed from abstract conceptualization and those who need to start from concrete experience; both those who are more and those who are less reflective; and both those who see a problem overall and those who focus on a particular issue. A summary of these different learning styles is provided in Jarvis (2010). For educating adults, it is also important to appreciate the importance of using experience and culture to assist adult learning (Jarvis et al. 2011). Knowles (1973) promoted the view that adult learners are more self-directed than dependent, bring their experience into their learning, are motivated by a need to acquire knowledge and skills, and tend to be problem orientated. These thoughts form a useful basis for considering education provision at the later undergraduate and postgraduate levels and for continuous professional development. Knowles and colleagues (2011) and Jarvis (2010) place these thoughts in a modern context along with discussion of learning theories and the social context of learning.

General Considerations

Basic Teaching Principles

As with any educational endeavor, those arranging an experimental design course should consider the following basic questions: Who is it for? What will be the scope? What does it hope to achieve? How will it be delivered and how will it engage the audience? How will it be assessed and for what purpose? The answers to these questions and the time constraints should determine the particular mix of the interlinked elements of content, pace, and presentation that are appropriate for a particular audience.

There is also the personality and style of the teacher to consider, so there is some art in providing effective experimental design tuition for a variety of learners.

Who is it for?

Within the life sciences, experimental design might be part of a university degree course, postgraduate education, continuous professional development, or specific courses designed for researchers in a particular field. Comments from those in biomedical fields who have experienced general statistics courses indicate difficulty in applying the taught material to their experiments (Fry 2013b), and there is scope for additional experimental design teaching that enables them to do so. There is also a case for including some experimental design tuition in continuous professional development for animal technicians (see below).

What Will Be the Scope?

Although the basic principles would be the same, the emphasis and illustrative examples, content, depth of treatment, and complexities of design considered would be expected to be tailored to the audience. This could range from a simple understanding of the basics to complex design and analysis.

To avoid overload, the scope should be trimmed to the time available and the pace a particular audience can tolerate. This is discussed further in the case studies. Online or distance-learning courses may not have the same time constraints, but there are similar considerations for the placing of points for self-assessment and tutor feedback.

In general, it is better to cover a limited scope well than attempt to be more comprehensive than the time allows. This places the emphasis on good selection.

What Does It Hope to Achieve?

What those taught might be expected to acquire from the teaching is clearly related to the scope, but there is also the consideration of whether the full range of educational objectives will be attempted. Will imparting knowledge and development of understanding be supplemented by improvement of skills and attempt to influence attitude? Is it for training in the sense of gaining competence in specific skills in a particular discipline, or is it education with a broad approach aimed at understanding that can be applied to a range of scenarios?
A tuition model in which the expert tutor delivers personal knowledge and insight to learners who absorb this intellectual food may be cost-effective, and with a good lecturer entertaining and memorable, but it leaves the learner without the confidence in applying the information that practice provides. An alternative model is one where a knowledgeable facilitator encourages understanding by interaction between learners discussing carefully chosen information or scenarios and provides tutored practice. The feedback from the courses discussed in the case studies indicates that a mix of these approaches can be well liked and effective.

How Will It Be Assessed and for What Purpose?

Broadly speaking, assessment can be to help the learner, to assist the teacher, or to show attainment to a standard that can be externally verified, or all three. Assessment designed for learning stimulates the learner to think about aspects of a topic and, with feedback or correction, provides him or her with confidence in what has been acquired and gives indications of where there could be improvement. It can be self-assessment, plenary question and answer sessions, or more formal testing. For the teacher, assessment should show which topics are not being well appreciated, the level of knowledge understanding and skill being reached, and where additional tuition might be directed or future provision altered. Formal assessment for qualifications needs to be of a comparable level across the bodies involved in teaching for the qualification, and from year to year. Standard setting and external verification is a subject in itself. The points made by Witchell (2012) for physiologists on this matter are also relevant for experimental design teachers.

Particular Considerations for Experimental Design

Educating in Experimental Design Has a Different Starting Point from Teaching Statistics

It is notable that textbooks on statistics, even when intended for life scientists, tend to start with the characteristics of populations, sampling, probability distributions, and so on, whereas textbooks on experimental design start with an example of an experimental question or scientific method and an illustrative experimental question. This underlies a difference in approach between statisticians, with a delight in the nature of their subject, and experimenters, for whom statistics is a tool. Those doing animal experiments are likely to be more comfortable with starting with the experiment and proceeding to the limited mathematics needed to understand how to test the result.

The concerns about poor design reviewed at the start of this article point to the topics that should be better understood and therefore given some emphasis in an experimental design course. As the problems are encountered in publications from those already in the field, this is not just a matter for the education of those fresh to animal experimentation, so there is a need for courses aimed at improving the understanding for those already undertaking animal research. So those attending may have a wide range of experience, abilities, and backgrounds. A single course may have participants from a range of specialties who may be dealing with species from fish or invertebrates to nonhuman primates. They may greatly differ in prior knowledge of the topics and research experience, and someone just joining a research group may be alongside an experienced scientist held in high regard.

Technical Terms Used Pose Difficulties

The field of experimental design and analysis has terms with a precise meaning in the context that differs from that in common usage. “Deviation,” “error,” and “residuals” are obvious examples, but there are also more subtle differences. “Replication” as used for experimental units under different treatments is rather different from replication in the sense of repeating a situation under different circumstances or at a different time. There is also inconsistent usage. Standard deviation and standard error may be used interchangeably or with the distinction of the former referring to the population of values and the latter to the population of means from that population. Repeated measures is sometimes a term for cross-over arrangements, and in other cases it refers to serial sampling from the same experimental unit over a period of time on the same treatment. All this gives ample opportunity for confusing the learner or making the subject seem more difficult than it is, so clearly defining terms is particularly necessary in experimental design teaching for an audience unused to statistical parlance. Understanding of terms is also needed for using statistical packages correctly.

Attitude Change May Be an Important Consideration in Teaching about Studies Using Animals

Researchers may bring particular personal or cultural views to the use of animals, so the attitudinal objectives for experimental design teaching for studies using laboratory animals might include promoting recognition of animals as sentient beings, not just data-generators or tools, and appreciation of different levels of sentience so that primates, for example, are not used when rodents could be. Other objectives might be to encourage willingness to take advice and respect for the value of considering statistical analysis early in the design process.
Structuring a Course or Module

Content

It is convenient to consider first a fairly comprehensive list of what might be included in an experimental design course irrespective of target audience or time constraints and then discuss selection from this range of topics to match content and extent of coverage to the expected audience and to suit the time available. Here, suggested possible topics are given in what could be a linear sequence along with the reasons for including them.

General Topics

Scientific method. Undergraduates and researchers coming to experimental design tuition would be expected to know about the scientific method. It may be competently taught at the secondary school level, and if not they could be expected to have covered it in other courses and general knowledge. However, there is value in providing a definition and some scope for discussion. This reminds them that a hypothesis is formulated on the basis of observation or theory and tested by experiment.

It is a convenient stage to mention the comparative nature of biological experiments, that the “null hypothesis” postulates no difference between sets of biological material treated the same except for the matter in question, and that testing involves accepting or rejecting the null hypothesis with a certain level of confidence.

The discussion also provides a hook on which to hang the thought that the outcome of the experiment provides evidence about the hypothesis tested, and the amount of confidence that can be placed in the correctness of the hypothesis is dependent on the strength of that evidence and the quality of the experimentation and not on whether the results meet any arbitrary cut-off level of significance or have been published in a high-impact journal.

Deciding upon an experimental question. Placing this early in the sequence emphasizes that the design of an experiment centers around what it is trying to achieve. As Mead (1988) puts it, “The need for an experiment arises from a question or set of questions to which the research scientist wants to find answers.” This thought leads to discussion or illustration of what might be a good experimental question, namely one worth answering and that gives rise to one or more hypotheses that can be tested by experiments within the time and resources available.

Later on in a course, it may be worth allocating time to development of the skill of identifying a good experimental question and testable hypotheses. There could also be exercises to develop the different skill of stating the experimental question and hypothesis clearly. In their survey of randomly selected published papers, Kilkenny and colleagues (2009) found the objective of the work was unclear in 5%. Groups given the task of determining the objective in a published paper have found they could not be sure in about one-half of the papers studied, and this forms a good exercise for making the point about the importance of clear objectives.

Different kinds of experiments. Some procedures may be regarded as experimental in the sense that they involve investigating a matter that is to some extent unknown but do not have the purpose of testing a hypothesis. Examples are the recording of growth curves and organ weights in a novel strain as well as many observational or correlation studies. In what might be termed “exploratory experiments,” the primary objective is to generate information on which to build a hypothesis or look for patterns, and some definitions of the scientific method include this as a stage in the process. In “pilot experiments,” the intention is to provide preliminary information that can improve the conduct, quality, and efficiency of subsequent hypothesis-testing experiments, typically using small numbers of experimental units. Simply alluding to these various approaches early on corrects any impression that all studies involving laboratory animals are hypothesis testing.

Toward the end of a longer course, there is a place for considering the conduct and sizing of experiments that provide different types of data. Toxicology testing, multi-gene array analysis, neuroanatomical tracing, and neurophysiological investigation of connections by stimulus-response are common enough to be usefully discussed. How often, for example, should the demonstration of an apparent neurological connection be repeated in different animals to provide confidence in the result? Also, designing chemical or radiation mutagenesis to minimize the potential suffering and reducing the amount of wastage in crossing mutant or transgenic animals to produce new lines can be brought into discussions on ethics.

Ethical considerations and the 3Rs: replacement, reduction and refinement. Ethical considerations come into the initiation, conduct, and reporting of experiments. Whether a proposed experiment should be done and the importance of resisting any temptation to falsify or improve results are considerations applicable to a range of experimental situations.

In studies using living animals European Union legislation requires a harm-benefit analysis, which weighs the pain, suffering, distress, and lasting harm the experiment may cause against the benefit expected from the work (European Union 2010). That legislation also requires adherence to the principle of the 3Rs (replacement, reduction and refinement) put forward by Russell and Burch (1959). The same requirement (to consider whether nonanimal procedures could replace the animal use, to reduce the numbers of animals used to the minimum, and to refine the necessary animal use to minimize severity) is in legislations across the world (Bayne et al. 2010; Fry 2012).

The 3Rs need to be considered in relation to meeting the experimental objectives, and all can be integrated into experimental design teaching. Replacement can be discussed in relation to literature searching and setting the experimental aims, questioning not only whether nonanimal alternatives might be used but also whether an alteration of the experimental aim might avoid animal use. Refinement comes into
the decisions on what types of data to gather, the effect of animal discomfort or distress on the reliability and variability of the data gathered, the choice of procedures to be used, and when an experiment should be stopped (see “stopping points” below). Reduction is a major element, because it involves minimizing numbers overall, and avoiding wastage by using an efficient and appropriate design and including in it proper controls, avoidance of bias, and sufficient numbers to detect worthwhile effects.

**Why design?** It provides some level of motivation to explore why the researcher should bother about good design. Participants asked about this typically volunteer three general reasons: to obtain valid results from which safe conclusions can be drawn, to know how widely these may apply, and to use resources efficiently. They may also mention ensuring reproducibility but rarely raise the matter of ethics. However, poor design can be considered unethical if the experiment uses too many animals or fails to get worthwhile data or if it subjects animals to unnecessary severity.

Those unmoved by ethics may be convinced by an approach developed by R. Preziosi (unpublished data) that calculates the cost of either consistent overestimation of numbers needed or of underpowering experiments. Depending on local costing arrangements, just a 10% consistent overestimate could squander up to an amount equivalent to the salary of a researcher worker, and the cost of having to repeat underpowered experiments is even higher.

**Definitions.** As mentioned above, experimental design terms need to be clearly defined, and it is helpful for the participants if this is done before a course tutor starts to use them freely. Although a glossary may be useful, taking all the definitions as one block in an oral presentation is likely to be counterproductive. Providing a set of definitions relevant to the succeeding presentations and repeating some later seems to be effective. In addition to the statistical terms mentioned above, “treatment” merits defining, as a teacher easily slips into talking about comparing control versus treated groups, for example when meaning control versus drug treated, although these should be regarded as two treatment groups from the point of view of design and analysis.

The term “degrees of freedom” needs explanation, as life scientists often have difficulty with it. One can use the approach in Fisher (1960) of equating it to the number of possible comparisons, or take the sum or mean of a set of values and show how many individual values can vary and still get the same sum or mean.

**Basic Principles of Design**

**Suitable comparisons and controls.** The need for controls seems well understood at the undergraduate as well as higher levels. In a study on science undergraduate research experiences in the United States, the importance of controls in research scored highest of all the understandings rated (Kardash 2000), and in course pretests over a period of 6 years, > 80% of the researchers participating have correctly identified when a negative control is necessary (though the figure was lower for animal technicians). However, consideration of controls is needed both for completeness and to provide some corrections. A minority of researchers does not see comparisons as providing controls for one another, and, for example, consider a sham operated control is necessary even when the objective is only to compare a new operative procedure with the current one. More prevalent is the lack of consideration of controls other than a negative one. About half of the groups discussing a scenario that calls for a positive control fail to identify that there should be one.

**It may help to extend the thoughts on controls to illustrate that observed changes may have many possible interpretations and that controls or comparisons are essential to limit the possible interpretations.**

Researchers may also put forward as “controls” the naïve animals that may be used for infection surveillance or to indicate whether any adverse effects seen are actually due to the experimental alterations. These are not strictly controls unless they are needed to interpret the experiment, such as when severity comparison is an experimental objective, but are worth discussing as it is valuable to have additional naïve animals for these purposes.

At some stage, it is appropriate to point out that internal controls are comparisons within an individual, taking out variation between individuals, so they have advantages over external controls. It is also worth voicing strong reservations about historic controls. Even experienced experimenters may not reflect that, although historic controls may give a good guide to the extent of variability in a population, they do not control for any difference in circumstances between previous periods and the time of the current experiment, so there should be some concurrent controls.

**Replication.** An important element that needs to be stressed, in view of the prevalence of pseudoreplication (Hurlbert 1984; Lazic 2010), is that replicates are independent and capable of receiving any of the treatments. Definitions like “repeat application of a treatment to another independent experimental unit” or “repetition of measurement or observation in a way that each repeat can be independent of the others” convey this.

It is worth illustrating the relationship between the number of replicates and how good the estimate of variability is. It can be telling to take data from a two-group comparison with a modest difference in the means and show first 2 values for each group, then 4, then 8, then 16, and so on, and ask the audience at which point they are confident the two groups could be reliably compared.

Replication could be returned to, in courses long enough, to discuss analysis of experiments where there is only one replicate per treatment, as may be the case in certain agricultural disease studies that require isolation facilities.

**Randomization and avoidance of bias.** As experimental design texts from Fisher (1935) onwards make clear, randomizing is fundamental to avoiding bias. However, failure to randomize properly is one of the concerns about study quality highlighted by several papers. So a course should convey that
importance by the way the topic is presented and the time spent on considering it. It calls for more than an instruction that the experimenter must randomly allocate treatments to experimental units. Illustration of the biases introduced by nonrandom allocation, discussion of how to randomize experimental units, and “identify-the-error” exercises using scenarios where proper randomization was not carried out all help participants to appreciate the point.

The experimental unit. In his classic work, Fisher (1960) uses the term experimental unit as one that can apply to a range of biological material, and in teaching about animal experiments it is a useful means of avoiding a ready assumption that individual animals are the replicates and that the number of animals is the appropriate number for statistical testing. Festing and colleagues (2002) define it as “The unit of replication that can be assigned at random to a treatment,” which is concise but relies on an understanding that replicates are independent of each other. This needs stressing, because experimental entities are often taken as independent when they are not (Lazic 2010). Such pseudoreplication is basically a failure to specify the experimental unit correctly. Schank and Koehnle (2009) give a good discussion of dealing with situations when individuals are spatially grouped, as fish in ponds are.

Correct identification of the experimental unit is a skill that a course can help develop, although appreciating what the experimental unit is in designs in which individual animals or groups of animals receive different treatments successively can be difficult for participants.

Because this can be a source of confusion, it is advisable to make clear that measurement is often made on a sample from the experimental unit and that with experimental units made up of several individuals, values from those individuals are pooled or averaged to give the value for the experimental unit.

**Design and Analysis**

The importance of knowing how a design can be analyzed. Statisticians commonly complain that they are consulted only after an experiment has been conducted when they may be able to do little more than say why it cannot be analyzed. Although the experimental questions and many of the considerations are biological in nature, and analysis should not be a tail that wags the design dog, any impression that analysis is merely a technical last element needs to be corrected. A mantra that no experiment should be started upon without good understanding of how its results will be analyzed can be reinforced by anecdotes of when someone knowledgeable in statistical analysis has been consulted too late to salvage any meaningful results.

Severity considerations and planning for humane stopping-points. Although well aware that discomfort and distress can affect their own physiology, experimenters may not think about similar affects in their experimental animals. Inclusion of severity considerations can remind them that when laboratory animals are providing the experimental data, the quality, reliability, and reproducibility of their results could depend on how much the well-being of the animals has been disturbed. This is where an animal technician audience may warm to the subject.

How the severity of the experimental procedures may be lessened or how alternative, less invasive procedures may be used is one aspect, and the other is setting humane end points in the sense of cut-off points at which animals will be taken off the experiment to avoid unnecessary suffering. These can be thought of as three types: a severity limit at which the suffering an animal is experiencing is greater than the experiment merits, an objective-achieved end point after which any suffering is unnecessary because the objective has already been achieved, and an objective-unachievable end point at which it is recognized that the objective cannot be achieved so that further exposure of the animal to potential suffering is unnecessary (Fry 1999). It is worth calling attention to the potential loss of data points that may result from the operation of severity limit end points and that some designs would be more sensitive to this than others.

**Signal/noise ratio.** This is essentially a discussion of maximizing effect (choosing the best measures, picking sensitive subjects, increasing the stimulus, and the like) and minimizing variation. It suitably follows severity considerations, because maximizing effect by giving a greater stimulus or disturbance may increase the severity and calls for judgment as to whether that is justified. It also nicely leads into discussion of variability and how to reduce or control this.

**Sources of variability and reducing variability.** As well as reminding the audience of the many sources of “noise” that come from the animals themselves, the people handling them, and the environment they are in (housing, husbandry, social hierarchy, and so on), some time might be spent on participants discussing the particular sources of variability in their work and how they might be reduced. This can be revealing for those who direct experiments but rarely visit the animal house. It is a suitable point to talk about the value of acclimatization, habituation, training, good handling, freedom from infection, and standardization of procedures.

**Blocking.** It is also a good point to introduce the idea of blocking. Fisher (1960) considers that this increases precision when “there is less variation within certain aggregates of [individuals] than there is among different individuals belonging to different aggregates.” Replace “individuals” by “plots of land” and “aggregates” by “blocks” in this quotation to put it in the agricultural context in which Fisher developed the ideas, and the concept of “blocking” can be more readily understood. This can then proceed to different types of experimental units and different groupings or “blocks” with illustrations of blocking by litter, age, position in cage rack, day operated upon, and so on.

**Expressing variability quantitatively.** At some point in the consideration of variability, one needs to ask how a figure can be put on the extent of variation. This is where it is appropriate to talk about populations, probabilities, the “normal distribution” and how that can be described mathematically,
the “standard deviation” as a measure of the spread of a normal distribution, and the distinction between it as referring to the population of values and the “standard error” as referring to the distribution of means of samples taken from that population.

Risk of false positives and false negatives: Type 1 and Type 2 errors. This is a suitable point to discuss significance testing in relation to the normal distribution, the risk at each arbitrary significance level of an apparent effect occurring by chance, and the other risk of underpowering an experiment, that is missing an effect when there is one by having insufficient replicates. These may be termed false positives and false negatives, but it should be made clear that that is with regard to the alternative hypothesis (that there is an effect).

Determining the numbers needed. The six variables that link in power analysis and how five of them allow for an estimate of experimental units needed are readily demonstrated and appreciated. In a computer session, the way in which the size of effect and the variability affect the group size needed can be brought home, in particular how four times the number is needed when the standard deviation doubles. Depending on the statistics package used locally, it may be appropriate to point out that packages may do power analysis on the basis of a two-group comparison and can give conservative estimates for the numbers needed in other designs. The alternative, which may be better for complex designs, is using the resource equation (Mead 1988), which gives an estimate of the number of experimental units for the whole experiment. The diminishing returns approach that underlies this is easily understood, but explaining the various treatment, block and error terms in the equation \((T + B + E = N−1)\) sensibly follows after blocking has been dealt with.

This is a suitable point to consider “sequential experiments” whose size is not determined beforehand but which are continued up to preset cut-off points, and Waterton and colleagues (2000) provide a good indication of what is involved in these.

The sizing of experiments with the objective of seeing whether there is a change in occurrence, and for which the estimate of numbers needed is based on probability, could also be discussed. Examples are studies of low-incidence characteristics in crosses of genetically altered animals and alterations of the incidence of prion disease with different treatments.

T-tests and analysis of variance. Before passing on to particular designs, it is helpful to consider analysis. The t-test can be presented as a standardized signal/noise ratio and audiences reminded of testing not just of a comparison of two groups but also the differences in an internal comparison. It is logical to pass on from the t-test to a one-way analysis of variance (ANOVA) and present the concept of ANOVA and the steps involved in assigning the variability to different sources, going on to two-way ANOVA. To aid understanding and give some practice, an exercise using ANOVA to analyze a simple block arrangement could be used. There should also be mention of some of the post-hoc tests. However, this needs to be no more than pointing to the different nature of the common ones, like the Bonferroni correction and Tukey and Dunnett tests, and why you might choose them. The detail of how to perform these is best taken from texts.

Types of experimental design and choice of design. There is a useful sequence of explaining a particular design, providing examples of when it should be used and showing how the results would be analyzed. Fully randomized, randomized block, and factorial designs should be covered, as it seems the first is often used when one of the other two would be better (Kilkenny et al. 2009). For those involved in pharmaceutical screening using animals, the advantage of a large control group can be discussed. Correct identification of the type of design that would best fit different experimental circumstances is a skill that can be developed and can be complemented by practice in analysis.

Discussion of Latin square and cross-over designs, and when they can and cannot be used, would also be appropriate. Whether to include split-plot, incomplete block, and sequential designs in any detail is a judgment that depends on the time available and who is in the audience. There is similarly a judgment as to whether to cover experiments including covarates or those needing multivariate methods in the analysis.

Nested ANOVA. The extent to which this is covered might also depend on the audience. Those working with agricultural animals or fish in experimental ponds are likely to come across situations where the opportunity to separate individuals so they can be independently assigned to treatment is limited by the number of pens or ponds available. There is risk of different effects being super-imposed and analyzing the variability at different levels is advantageous. For those using laboratory rodents, nested ANOVA could help identify the extent of variability at the experimental unit, sampling, and measurement levels and thus at what level to increase numbers in future experiments.

Internal and external validity. Because factorial arrangements provide information on a number of variables, they form a good link to consider the extent to which the results of an experiment can be generalized.

Types of data, nonparametric tests and contingency tables. These are grouped because there is an advantage in taking them together. For each type of data, it is helpful to include examples of that type and how an experiment producing data of that type can be analyzed. However, detail of the numerous nonparametric tests is probably best left to private study.

Assumptions behind parametric statistical tests. This would cover not only the assumptions themselves but how robust standard parametric tests are to these assumptions being violated. The importance of independence and random assignment can be reemphasised and consideration given to unequal sample sizes, unequal variances, and cases where the values in treatment groups do not show a normal distribution around the group mean. Discussion of examples where these “residuals” are not normally distributed leads to discussion of transformations and use of nonparametric tests as alternatives.

Correlation and regression. These are not prominent in laboratory animal studies but merit inclusion to raise
awareness of how to determine whether two variables are correlated and that linear regression analysis may be applicable to some investigations, such as a time sequence. It would also be worth warning that correlation does not mean causation and making the distinction between correlation and linear regression analysis.

**Additional Topics**

**Planning a sequence of experiments.** Discussion of minimizing numbers and severity overall in the way a sequence is organized is unlikely to be a feature in texts on experimental design. So it may be worth including a presentation which covers the use of feasibility studies and pilot experiments to avoid wastage in larger definitive studies and to alert to potential problems and which also shows the advantage of a severity sequence in which experiments involving less invasive and distressing procedures precede greater severity ones (Fry et al. 2010; Gaines Das et al. 2009).

**Statistics programs/packages.** Depending on time available and on whether there is a linked course on statistics, demonstration of how to use one or two commonly available statistics packages or of programming in R or R commander could be included. This should impart understanding of the terms used and improve confidence in using the software for analyses.

**Presentation of data.** Because those attending an experimental design course will be involved in assessing published results and presenting their own, a brief session on presentation could be useful. It might, for example, show how some presentations can be misleading (lines in graphs leading the eye, suppression of the zero in bar charts exaggerating differences, and so on). It could also consider the use of confidence intervals rather than P values, and discuss the merits of box-and-whisker plots and other graphical presentations.

**Comparison of Bayesian and frequentist approaches.** In a more advanced course, it would be worth alerting participants to the difference between these approaches, and Bland and Altman (1998) is a suitable brief summary. However, to do a Bayesian approach justice would take more time than would normally be justifiable for the likely ability to make use of it.

**Further reading.** It is helpful to have recommendations for suitable texts both to accompany the teaching and for reference later, particularly on matters not covered in detail. Most topics in the list above are included in Festing (2010) and/or Fry (2013a). Festing and colleagues (2002) and Ruxton and Colegrave (2011) have been recommended in the courses discussed in the case studies, as they presume little mathematical or statistical understanding and cover many aspects of experimental design at a level suitable for new researchers. A useful supplementary text on statistics is McKillup (2012), which gives an introduction to statistical analysis at a similar level. Morris (1999) deals with experimental design for those working with agricultural animals. Morton (1998) explores refinement aspects of experimental design, and severity issues are also covered in the latest ILAR Guide (National Research Council 2011). Fisher (1935, 1960) is the classic seminal work, and Quinn and Keough (2002), Mead (1988), and Cochran and Cox (1957) can be given as reference works with a comprehensive treatment of experimental design and statistical analysis.

**Delivery**

**Variety**

As indicated in the Introduction, people learn in different ways and the development of understanding is nonlinear. The occurrence of a flash of insight does not have a strict relationship to the amount of information or illustrative material presented. Practice and experience are important in relating knowledge to application, and people may need time to think about information or arguments before appreciating them. Different approaches have different degrees of effectiveness for information provision, improvement of understanding, skill development, and attitude change.

These thoughts suggest there should be variety in both the components of experimental design teaching and in the nature of those components. Oral or text-based presentation including good pictures and diagrams and a logical argumentative progression should help different types of learners and can be supplemented by practical and discussion sessions, with breaks between bodies of taught material.

The case studies illustrate how this might be put into practice and the compromises that can be made to fit to time constraints. The courses examined in detail provide variety by use of oral presentations with strong visual and argument content, group exercises and discussions, short assessments as a learning tool, and in some cases computer sessions. More information on these and how they are conducted is given with the first case study.

Part of variety is lightening the tone. There is opportunity to do this by use of absurd comparisons, from exposure of questionable logic, and by making fun of the poorly replicated, uncontrolled comparisons that appear in newspapers and on television.

**An Educational 3Rs: Repetition, Recapitulation, and Reinforcement**

It helps appreciation for the material to be repeated with rather different delivery or a different context at various stages in a course or module, and for it to be included in an activity that involves the learner using that information. For example, the concept of control may first be introduced in a list of fundamentals, then repeated with a breakdown into different types of controls, then recapitulated when discussing variability and quantitative comparison and reinforced in an exercise or discussion to identify suitable controls or to design an experiment.
Conveying Attitudes

Studies of attitude change in the science classroom point to the importance of the teacher, the message and how it is conveyed (with imagery superior to summarized data), and the active participation of the recipient (Kobella 1989). They also show the lack of a relation between attitude change and gain in factual information. These are relevant considerations for experimental design teaching. The teachers will convey their own attitudes in the topics emphasized, the words chosen, their body language, and so on. They can be persuasive if perceived as expert and credible. In discussion groups, these attitudes may be reinforced by knowledgeable participants, or an appreciation of the arguments that comes from discussion, but could be countered by some with strongly expressed other views.

Online Delivery

Courses such as those considered in the case studies can reasonably take up to only 70 participants, and the demand for education in experimental design may be much greater than such courses can deal with. Online delivery can reach a much wider audience, and the popularity of massive online open courses could indicate this is an approach to consider. The CalTech debate about massive online open courses (Kuzins 2014) brings out the advantages and disadvantages of online delivery. Online delivery does not match the performance experience of the live presentation and the level of interaction with tutors and other participants that occurs in face-to-face discussion. These are aspects of the courses in the case studies that are highly rated, so an online course would be no substitute for these courses. However, the website www.3Rs.reduction.co.uk (accessed on April 30, 2014) provides a good online resource that could be used for online study.

Assessment

There is value in feedback for both learner and teacher. Teacher to learner feedback provides correction and encouragement, and learner to teacher feedback indicates elements or topics that could be better expressed or need more time. Assessment is one means of gaining such feedback and complements end-of-course questionnaires, independent personal debriefings, and the like. Courses or modules that are not to be formally examined should still include some assessment for learning. This is structured to enable individuals or groups to appreciate their deficiencies, to promote individual learning from thinking about questions and problem solving, and to provide confirmation of advance. It can range from simply questioning the audience to a written examination but always with correction and discussion of the responses. With choice questions where there are distractors, it is particularly important to discuss the choices so that incorrect or poorer answers are identified and the reasons why they are not the best response explained.

Assessment of attitude change needs a different approach. An example is given in Figure 1 that summarizes results from an assessment of attitude change by questionnaire. For some statements, there was a notable shift in response after the course, which covered animal welfare and the ethics of animal use as well as experimental design but did not specifically discuss the statements in the questionnaire. Formal evaluation, normally against specified learning outcomes, is needed for taught courses that lead to a recognized qualification. It should include not only assessment of knowledge gained, but also demonstration of understanding by application of the knowledge to different circumstances, and of skill in formulating hypotheses, identifying proper controls, and correct experimental units, and picking efficient designs in different postulated experimental situations.

Figure 1 Average agreement rating in initial and final questionnaires to different statements with an attitudinal content. Each of the 36 respondents gave a rating on the five-point scale against the statement shown, both at the beginning and at the end of a two-day workshop run by the RSPCA International in Taiwan. The ratings for each statement were averaged to give the bars shown (D Fry, M Jennings, P Littlefair, unpublished data).
Case Studies

Three-Day Courses

This section principally considers the courses run by the Fund for the Replacement of Animals in Medical Experiments (FRAME), which despite its name has a Reduction Steering Committee and has taken considerable interest in experimental design. The courses cater for up to 50 participants and involve a number of tutors, so there is variety in style of presentation as well as modes of delivery. These courses have developed over the years in response to participant feedback and as tutors have adjusted the timing and content of presentations. Figure 2 shows the timetable of the most recent one (run in 2014). Although UK-based, the courses have also been run in The Netherlands, Portugal, and Denmark and attract researchers from across the world.

Topics and Time Allocations

Some idea of what time is needed for a researcher audience to appreciate various topics can be seen from the schedule in Figure 2. The three days do not allow much coverage of experiments where the data generated are nonparametric nor proper consideration of more complex designs or those specific to a particular field. Neither does it give long enough for participants to get much practice in formulating clear objectives or devising testable hypotheses or determining appropriate sample size. The analysis of each of the designs covered is included in the presentation, but there is little opportunity for participants to develop their capacity for analysis.

Delivery Methods

As the timetable in Figure 2 shows, there is a mix of presentations, group exercises and discussions, and computer sessions.

Oral presentations. These supplement each theoretical point with pertinent illustrations based on the wide and varied experience of the tutors, and several are interactive, involving questioning of the participants or responding to their questions. For each design presented in detail, an example with points covered in the course.

Group exercises. These reinforce the content of the previous presentation(s). Experimental scenarios based on the presented information, each with questions that need only short responses, are given to each participant. Groups of up to eight then discuss what would be the appropriate response for each scenario, with no tutor assistance apart from any clarification of a scenario needed and occasional steering of the discussion away from an unproductive line or cooling any debate becoming heated. This gives opportunity for those who have better understood the preceding presentations or for whom the information given was known and the presentation was mainly a reminder to help the learning of others in the group. About one-third of the time allocated for the session is given to plenary discussion in which the response from each group for each scenario in turn is taken with little comment and no discouragement, and then the scenario and responses are discussed by the tutor.

Computer sessions. These both reinforce basic points and give practice in simple analysis. They give an introduction to Minitab and/or R and show how these can be used to generate normal populations and randomly sample from them and perform two-group comparisons and ANOVA. Random samples taken by a class of 50 usually show the chance occurrence of sample means that could be taken as coming from another population if a 5% significance level were used. The session can also demonstrate the value of plotting the data and discuss the tests for nonnormality.

Assessments

In these courses, there is no formal assessment, but there are questions in some of the presentations, the group exercises include problem-solving, and a quiz-like test is given near the beginning of the course and another near the end. Each test has a series of 12 multiple choice questions given as question slides in a class format for individual response without collusion. From one to two minutes is allowed for each, depending on the length of the text in the question and its complexity. Some of the first test questions are repeated in the second, and others are replaced by questions of similar difficulty testing similar understanding. Although these are the pre- and post-tests generating the data shown in Figure 3, the primary intention is education not evaluation. The participants only identify themselves by pseudonyms. The objective of the first test is to stimulate thought and provide a baseline, and that of the second is to allow participants to realize that they have learnt something from the sessions. The discussion of each question following the second test reinforces some of the points covered in the course.

Effectiveness

The main indication of the effectiveness of these courses comes from the increased confidence in use of terms and understanding of concepts that the participants show in the later discussions. This impression is supported by the difference in numbers giving correct answers to the quiz-like pre- and post-tests just mentioned, which are shown for the most recent course in Figure 3. Some of the difference could be due to improved understanding of the quiz format, but that contribution is likely to be small. The marked post-test improvement is similar to that shown for an earlier course in Fry (2013b) and has been seen in all the courses (although development of both the course material and the question set means that the results from different courses are only qualitatively comparable). In the survey Howard and colleagues (2009) conducted of those attending the first courses, 26 of the 30 respondents considered the course had improved their research a medium or large amount, and there is anecdotal evidence that that is true for later courses.
Participant Comments and Evaluation

Participants’ comments and external assessments of the workshops rate them as informative, educational, and enjoyable. The feedback forms from the four FRAME courses running from 2009 to 2012 have been analyzed. Of the 99 general comments offered, 51 praised the course as a whole, 11 others picked out the quality of oral presentations, and 8 others valued the group discussions. Twelve specifically mentioned how much they appreciated interaction with the

---

**Figure 2** Timetable for a three-day course (courtesy of FRAME).
tutors and the ability to have questions on their own research answered. Only 4 asked for more on statistics and analysis. The other 13 mentioned various specific issues.

In response to specific questions on the feedback forms, 87.4% of the 151 respondents marked “the academic level” as “correct”, 80.8% agreed “the lectures were clear and easy to understand”, 97.4% agreed with “the course exposed you to new knowledge and practices,” and 95.4% agreed with “you would recommend this course to your colleagues.”

This all points to a successful format with the right content and variety of delivery for this audience made up mainly of postgraduate and postdoctorate researchers with some senior scientists and also veterinary and medical academics.

The Course as a Postgraduate Module

The course adapts well to being a module in the taught part of a postgraduate science degree. However, run as a block of 3 to 4 days, it is intensive and unforgiving of even short absences. With a residential course, participants have freed the days from other commitments, but for a junior researcher there may be calls from the administration or to activities in the research group that conflict with a total commitment. So it would be better run over a series of half days, which would allow other activities to be timed outside the course periods and for those having more difficulty with the topics to catch up by reading or discussion. A weakness is that it does not de-

![Figure 3](https://example.com/figure3.png)

**Figure 3** Percentage of 30 participants giving correct responses to questions testing understanding of basic principles and use of different types of design. The data comes from the three-day course whose timetable is shown in Figure 2, which researchers from several European Union countries attended. Quiz 1 formed the pre-test and Quiz 2 the post-test. The question set used was the same as for the results shown in Figures 5 and 6.

vote much time to giving participants practice in analysis, and it is likely to need complementing by a statistics course that does provide tutored practice in the various methods and the locally available statistics package.

Another consideration is logistics. The course requires easy movement between oral presentation, group discussion, and subsequent plenary at times that fit with its schedule. University facilities geared to lecturing and laboratories may have few rooms with projection facilities and scope for cabaret-style seating, and the alternative of booking both lecture theatres and break-out rooms may be difficult. A lack of nearby computer facilities that can be used at the times scheduled is less of a problem, as the exercises can run satisfactorily in a lecture theatre with participants using their laptops.

Half-Day Courses

A half day is a very short time to deal with a subject as broad as experimental design and means there are difficult choices as to what is included and how the course is conducted. Figure 4 shows a timetable based on the considerations below that has been successfully used for groups of academic or pharmaceutical researchers in both the United Kingdom and the United States.

![Figure 4](https://example.com/figure4.png)

**Figure 4** Timetable for a half-day course.

The anticipated audiences were researchers at graduate level or above with at least some experience of animal studies and some statistics knowledge or nongraduate scientists with considerable experience. The decision was to select topics to address the concerns raised in published surveys, namely failure to keep to basic principles and tendency to use only simple
group comparisons when there would be more efficient designs. As many of the researchers were expected to be involved in conducting a series of experiments around a main aim, a brief presentation on planning these was thought worthwhile.

Another decision was to indicate appropriate analysis but not go into much detail and either refer participants to suitable texts or run the course in liaison with local statisticians who could provide back-up and advice outside the half-day constraint.

Delivery

If the whole time were given to oral presentations, more topics could be covered, more illustrative examples discussed, and the analysis of several designs could be shown. However, this would be almost entirely knowledge transfer without skill development or opportunity to put that knowledge into practice in a tutored environment. Also without the advantage of encouraging learning by problem-solving or interaction with other researchers that the time-consuming group discussions and computer exercises provide. The compromise was to include group exercises only on basics and some readily understood designs and to leave computer-based practice to individual study.

Assessment

Pre- and post-tests like those in the three-day course would take a fifth of the time available. However, it was clear from the responses of participants in the three-day courses that the quiz-like format had an entertainment element, the pre-test (quiz) reminded people there were matters on which they were uncertain, and the post-test with discussion afterwards was both a confidence-enhancing and learning experience. For these reasons the pre- and post-tests were retained but shortened to the 10 key questions.

Evaluation

The sessions need to keep strictly to time, and keeping up the pace is demanding on the tutor. For the oral presentations, a judgment has to be made as to what gets only a mention and what gets more consideration; group discussions edging towards the unproductive have to be quickly corrected, and the subsequent reporting and discussion have to be efficiently run. There is little opportunity to deal with participants’ questions, and these have generally had to be discussed outside the timetabled session. Despite these challenges, the courses have been judged worthwhile and effective, with participant feedback showing a high approval rating and comments echoing those about the three-day course. The pre- and post-test comparisons from 10 of them that were fully comparable are shown in Figure 5. Although the post-test improvements are generally not quite as good as with the three-day course, they are most encouraging for such a short format.

Adaptation to a Two-Hour Time-Slot

This could be achieved by omitting the pre- and post-tests and the presentation on planning from the half-day timetable and

---

**Figure 5** Percentage of 281 participants giving correct responses to questions testing understanding of basic principles and use of different types of design. Aggregate data from 10 half-day courses, all with the half-day timetable shown in Figure 4, taught by the same tutor and with similar participants. “Pre-quiz” formed the pre-test and “Post-quiz” the post-test.

**Figure 6** Percentage of 13 animal technicians giving correct responses to questions testing understanding of basic principles and use of different types of design. Higher education module with five taught days, as described in the text.
taking some of the problems out of the group exercises so the discussions can be kept to 20 minutes with a 10-minute plenary. This has been done and could be judged successful on the basis of comments received, but it is very restrictive.

Continuous Professional Development for Animal Technicians

Animal technicians are a group worth considering for experimental design teaching, because some of them may come to be setting up or directing experiments themselves, and almost all will be dealing with experimental animals. It is important that they can appreciate that what they do can affect variability, that randomization is fundamental and departure from a random arrangement for convenience could make the results invalid, that housing may affect what can be taken as the experimental unit, and so on. However, if they are not graduates or used to academic approaches, providing an effective course that they appreciate may be quite a challenge. The topics selected for the half-day course detailed above form a good core, but a few technicians attending such a course have found the pace daunting.

Key features in adapting the half-day course material to be more suitable for an animal technician audience have been a much slower pace, a lot of relevant examples, many problem-solving group exercises with discussion, practice by means of computer exercises, and plenty of time for questioning the tutor. So the oral presentations shown in the timetable in Figure 4 were subdivided into short talks on each topic and interspersed with group exercises and computer sessions. The two group sessions shown in Figure 4 were each split and additional ones recruited. These were on what might contribute to variation, signal/noise ratio, finding design flaws, identifying suitable blocks in some experimental scenarios, setting out factorial arrangements, planning an animal husbandry study, and criticizing published studies. Computer exercises covered randomization, values with a normal distribution, means of samples of different sizes, distribution of means, two-group comparisons, ANOVA, and power analysis. To allow cross-comparison, the quiz-like pre- and post-tests were kept the same as for the other course formats, but there was also some interim assessment. Altogether, these changes expanded the material to five taught days. Both the formal assessment of the module and the pre- and post-test comparison shown in Figure 6 confirmed the impression gained from the discussion sessions that this nonacademic group advanced considerably in knowledge, understanding, and skill in the subject.

Summary and Conclusions

Education in experimental design for life scientists undertaking animal-based research needs to address the concerns about study quality of animal experiments expressed by many authors and the common failings found. In order to do so, the tuition should discuss formulating clear experimental objectives and the nature of hypothesis testing and give thorough consideration to the basics of good design. To encourage wider use of more efficient designs, fully randomized, randomized block, and factorial arrangements should be compared and explained, including when they are suitable and how they are analyzed.

These topics can form a core content that can be expanded to include a range of different designs and statistical tests according to the time available and the suitability of the audience. Teachers addressing those involved in animal work need to be aware that the audience may include several daunted by the mathematics on which the statistical tests are based, and adjust content, presentation, and pace accordingly. Feedback from the courses considered in the case studies indicates much can be achieved without losing the interest and comprehension of that part of the audience.

Although discussion of ethics of animal use and related legislative requirements may take place under other headings, consideration of housing and environment, and of severity issues should be integrated into the teaching of experimental design to those using animals. Experimenters should be aware that housing arrangements and local conditions may affect animal welfare and influence what can be taken as the experimental unit. They also need to recognize the risk of poor quality and variability in the results from distressed animals and may need to take possible loss of data points into account in the design.

In addition to good information provision, problem-solving and tutored practice are important elements in experimental design teaching, as they reinforce the information provided and participants gain in understanding and skill. Giving time to these activities also avoids the instruction received only being applied later and within the apprenticeship style learning of the research group, which carries the risk of perpetuating imperfect practice. Courses based on these ideas have been developed to suit different time constraints and somewhat different audiences. These courses have proved effective as judged by the level of appreciation expressed by those participating, the gain in knowledge, understanding, and skill indicated by some objective measures, and the impressions of the tutors both at the time and on later occasional contact. Participants have also stated that the courses improved their research.

All the formats discussed rely to some extent on there being local follow-up, additional self-directed learning, or opportunities to take matters further. Each could be improved by being longer. That could allow other topics to be covered, give more opportunity for development of skills and discussion of more experimental scenarios, and enable sessions in which the participants practised statistical analysis of different designs to be included.

Acknowledgments

The three-day courses reported on would not have developed without the active support of FRAME, and the others stem...
from those. I am greatly indebted for the development of my ideas to the other tutors on the early FRAME courses, in particular Dr. M. Festing, Dr. R. Gaines Das, both retired, Dr. D. Lovell of St George’s, University of London, and Professor R. Preziosi of the University of Manchester. I am also most grateful to J. Brien and others from the Education Department, University of Chester, for advice on the educational aspects of this article.

References


Fry D. 2013b. The need to improve experimental design. ALTA 41:61–64.


