ANIMAL BEHAVIOUR, 2002, **63**, 815–822
doi:10.1006/anbe.2001.1950, available online at http://www.idealibrary.com on IDEAL®

COMMENTARIES

Incorrect analysis of crossover trials in animal behaviour research

RAMÓN DÍAZ-URIARTE
Departments of Zoology and Statistics, University of Wisconsin-Madison, U.S.A.

(Received 9 July 2001; initial acceptance 26 September 2001; final acceptance 30 October 2001; MS. number: SC-1213)

In crossover trials each experimental unit receives two or more treatments through time; in the simplest case of two treatments, the subject is first given one of the treatments and then crosses over to the other treatment (Jones & Kenward 1989; Ratkowsky et al. 1993; Senn 1993a; Vonesh & Chinchilli 1997). Thus, crossover studies differ from parallel studies where each subject is exposed to the same treatment for the duration of the experiment. In crossover trials at least one key covariate (treatment) changes within subject over time. As the comparison of treatments is made within subjects, each subject acts as its own control which increases statistical power to detect a direct treatment effect (e.g. Crowder & Hand 1990, page 101; Senn 1993a, pp. 201ff.). This is particularly important when repeated testing of one subject is simpler than recruitment of new subjects. For these reasons, crossover trials are frequently used in behavioural experiments. However, crossover trials are often analysed inappropriately, as if they were typical matched-pairs designs, which they are not. The main problems are, first, not accounting for period effects (which leads to the inappropriate use of paired *t* tests in the two-treatment, two-period case) and, second, failure to consider carryover effects. (A treatment effect is the effect of a treatment at the time of its application, whereas carryover effects are effects of a treatment that persist after the end of the period, and a period is each occasion on which a treatment is applied; see Terminology below.)

For instance, in the 12 issues of *Animal Behaviour* from July 1998 to June 1999, there are 22 articles that use crossover designs in at least one experiment. Eight of these papers use variants of the two-treatment, two-period design (generally the typical 2×2 design); 17 papers use designs for more than two treatments. Results are analysed with paired *t* tests or Wilcoxon signed-ranks tests for two treatment designs, or with linear models (usually referred to as repeated measures ANOVA), and on a few occasions with methods specific for categorical data. Only two studies explicitly consider period effects, and one mentions that there are no effects of order of presentation (although the test is not explained); however no paper explains how potential carryover effects are dealt with. Counterbalancing (each treatment appears in each period the same number of times) is used in 11 papers. When counterbalancing is not used, order of presentation is 'randomized'. Thus, it seems that the majority of authors believe that counterbalancing or randomization of order of presentation, per se, will take care of any other nuisances (periods and carryovers) but, as we will see, this is not true. Authors seem unaware that carryover effects can bias their conclusions. The practical consequences of the analyses used in these papers are that: (1) if there are carryover effects, all reported results could be biased; (2) even in the absence of carryover effects, in the studies that do not use counterbalancing the estimates of direct treatment effects are biased if there are period effects; and (3) in studies that use counterbalancing, the estimates of the variance of direct treatment effects are inflated (i.e. they are larger than they should be) if there are period effects, making it more unlikely to reject the null hypothesis when it is false, and thus increasing type II error rates (and even if the study shows significant differences, the true direct treatment effect will be underestimated). Therefore, the conclusions reached in the majority of these papers are questionable: the lack of effects reported in some studies could be the consequence of inflated variances, and the significant effects reported in others could be the result of either period or carryover effects.

Statistics textbooks used by behaviourists such as Lehner (1979), Campbell (1989), Bailey (1995), Sokal & Rohlf (1995) and Bart et al. (1998) do not mention crossover designs. Other texts provide potentially misleading advice: Martin & Bateson (1993, pp. 29–30) would apparently use a paired test to analyse a 2×2

Correspondence and present address: R. Díaz-Uriarte, Navacerrada 37, 28430 Alpedrete, Madrid, Spain (email: ramon-diaz@teleline.es).

0003–3472/02/$35.00/0

design; Zar (1996, pp. 259–263) analyses a crossover design, and refers to carryover, but he fails to mention that period should be incorporated in the analyses, and seems to imply that counterbalancing per se can eliminate problems from carryover effects; Edgington (1995, pp. 114–117) suggests counterbalancing to prevent undesired effects from order of presentation; and Zolman (1993, pp. 59–63), although explicitly mentioning crossover designs and discussing carryover effects, apparently suggests (page 160) that a paired *t* test is appropriate for a 2×2 design.

The 22 examples from 1 year of *Animal Behaviour* show that crossovers, a powerful and widespread type of design, are often analysed inappropriately; and the textbook examples indicate that information on the appropriate design and analysis of crossover trials is not accessible to animal behaviour researchers. Thus, my main objective in this paper is to make animal behaviour researchers (and referees) aware of the most important pitfalls in the design and analysis of crossover trials. I first explain why the usual analyses of crossover trials in animal behaviour research are inappropriate and then I briefly discuss the problems of carryover effects. Next, I emphasize that there are a variety of designs available for crossover trials, as well as different methods to analyse these experiments. I conclude with a discussion on when to use crossover designs in behavioural ecology experiments. Elsewhere (Díaz-Uriarte 2001) I review in detail the statistical methods available for the analysis of data from crossover experiments in animal behaviour research.

Terminology

A direct treatment effect is the effect of a treatment at the time of its application. A period is each occasion on which a treatment is applied. Carryover effects are effects of a treatment that persist after the end of the treatment period; in other words, the response to a current treatment is affected by what treatment was applied in a previous period. A sequence is the order in which the within-individual treatments are applied. Designs will be referred to using sequences, such as ABB,BAA, which means that animals assigned to the sequence ABB are first given treatment A (first period), then B (second period), then B (third period), and animals assigned to the BAA sequence are first given B, then A, then A (first, second, and third periods, respectively). Therefore, a sequence effect or a group main effect is any effect related to a particular sequence of treatments, such as an overall difference in the responses to the treatments in animals of sequence AB compared to those of sequence BA. A sequence effect can result if animals assigned to one sequence are different from animals assigned to the other sequence, but under a randomized design it is reasonable to assume that there are no sequence effects (Crowder & Hand 1990). In many designs, however, a sequence effect can be confounded with other effects (see below, and also Jones & Kenward 1989; Crowder & Hand 1990; Ratkowsky et al. 1993).

Table 1. Fixed effects for the 2×2 design

Sequence group	Period 1	Period 2
AB	$\mu+\pi_1+\tau_1$	$\mu+\pi_2+\tau_2$
BA	$\mu+\pi_1+\tau_2$	$\mu+\pi_2+\tau_1$

In this table, carryover effects have not been included; including them would result in the fixed effects for period 2 being $\mu+\pi_2+\tau_2+\lambda_1$ and $\mu+\pi_2+\tau_1+\lambda_2$, in sequences AB and BA, respectively, where λ_1 is the carryover effect of treatment A on treatment B and λ_2 is the carryover effect of treatment B on treatment A.

Example of the 'Usual' Analyses and Their Problems

The 2×2 crossover design (the design with sequences AB,BA) is frequently analysed with a paired *t* test; this is equivalent to subtracting the response value under treatment B from the response value under treatment A for each individual and testing whether the mean is significantly different from 0 with a one-sample *t* test. However, in many behavioural experiments period has an effect: whether a response is measured on the first or second occasion will affect the value of the response (e.g. because of growth or seasonal changes, or through habituation to the measurement itself, and regardless of previous treatment(s), such as in Díaz-Uriarte 1999). With period effects the above analysis is inappropriate for two reasons (Senn 1993a, page 38; also Schneider 1983). First, if there are unequal numbers of subjects in each sequence, the test and the estimate of direct treatment effects will be biased. (Bias means that the expected value of the estimator is not equal to the parameter we are trying to estimate; bias does not decrease with increasing sample size.) Second, even if there are equal numbers of subjects in each sequence, we will lose power: period is a systematic trend, but by lumping together animals from both sequences, we are ascribing this systematic variation to the random component (the error term) and the standard errors of our estimates will be inflated. This second problem is similar to ignoring the effects of blocking (a known source of variation).

To understand these problems better it is convenient to write down an explicit expression for the statistical model (e.g. Jones & Kenward 1989):

$$y_{ijk}=\mu+s_{ik}+\pi_j+\tau_d+e_{ijk} \quad (1)$$

where μ is the intercept, π_j is the period effect of period $j=1, 2$, τ_d is the direct treatment effect of the treatment $d=1, 2$, s_{ik} is the random subject effect of subject k in sequence i, and e_{ijk} is the random error for subject k in period j in sequence i (for the moment we ignore carryover effects). From that model, the fixed effects for each period and sequence for a 2×2 design are shown in Table 1.

The expected value of the difference A − B for animals from sequence AB (dAB_{AB}) is $(\tau_1-\tau_2)+(\pi_1-\pi_2)$, and the expected value of the difference A − B for animals in sequence BA (dAB_{BA}) is $(\tau_1-\tau_2)+(\pi_2-\pi_1)$. The paired *t* test is the same as testing if the set of all dAB_{AB} and

Table 2. Simulated data (columns 3 and 4) for a 2×2 trial

Sequence	Subject	Period 1	Period 2	Period differences	Crossover differences
				$d12_{AB}$	dAB_{AB}
AB	1	16.5	11.1	5.4	5.4
AB	2	14.9	9.2	5.7	5.7
AB	3	14.2	6.9	7.3	7.3
AB	4	20.6	13.8	6.8	6.8
AB	5	18.2	12.8	5.4	5.4
				$d12_{BA}$	dAB_{BA}
BA	6	15.0	13.3	1.7	−1.7
BA	7	13.9	9.8	4.1	−4.1
BA	8	9.8	6.5	3.3	−3.3
BA	9	16.8	14.8	2.0	−2.0
BA	10	14.9	12.0	2.9	−2.9

A common (incorrect) analysis of direct treatment effects uses a paired t test, which is the same as testing if the crossover differences are centred around zero. The Hills–Armitage approach compares period differences between the two sequence groups. Period differences are subject differences between the first and the second period; crossover differences are subject differences between treatment A and treatment B; see text for explanation.

dAB_{BA} are centred around zero, using a one-sample t test. If there are more animals in AB than in BA, our estimate of direct treatment effects $(\tau_1 - \tau_2)$ will be biased by a factor proportional to $(\pi_1 - \pi_2)$; when the sample sizes of both sequences are the same, there will be no bias in the estimate of the direct treatment effect, but the error term will be inflated by a term proportional to $(\pi_1 - \pi_2)^2$. Thus, a paired test results in biased estimates of direct treatment effects and/or inflated variance estimates; counterbalancing per se does not result in a correct analysis, contrary to what is sometimes believed.

To prevent these problems, we should use the Hills–Armitage approach, illustrated in Table 2 and described in more detail in Jones & Kenward (1989, pp. 23–28), Crowder & Hand (1990, page 101) and Senn (1993a, pp. 42–44). We take period differences (subtract period 2 from period 1) for both sequences, yielding $d12_{AB}$ and $d12_{BA}$ for animals from sequences AB and BA, respectively. The expected values of these differences are: $E(d12_{AB})=(\tau_1 - \tau_2)+(\pi_1 - \pi_2)$, $E(d12_{BA})=(\tau_2 - \tau_1)+(\pi_1 - \pi_2)$. We can test for treatment differences by comparing the means of $d12_{AB}$ and $d12_{BA}$ ($\overline{d12_{AB}}$ and $\overline{d12_{BA}}$) between the two sequences (e.g. a two-sample t test). Define $\hat{\tau}=0.5(\overline{d12_{AB}} - \overline{d12_{BA}})$; its expected value is $(\tau_1 - \tau_2)$ (so there is no bias) and the variance contains only a term for the within-individual errors (see Jones & Kenward 1989, page 26). In other words, to test for treatment differences we compute the difference between the first and the second period for each individual, and then we use a two-sample t test to compare these values between the two sequences.

To test for period effects, we compute crossover differences (difference between periods 1 and 2 for subjects in AB, and difference between periods 2 and 1 for subjects in BA: equivalent to computing differences between A and B for all subjects), and use a two-sample t test comparing these differences between the two sequences. Finally, to test for inequality of carryover effects we compare the sum of the values in the two periods between the two sequences (see Jones & Kenward 1989, pp. 24–25). Note that we cannot test for absence of carryover effects, only inequality or differential carryover effects (see next section), and in the 2×2 design the test of carryover effects requires us to make the assumption that there are no sequence effects (were we to try to estimate both, we would not be able to, since in the design matrix the columns for sequence effect and differential carryover effects are identical; the 2×2 design yields only four cell means, and thus only a maximum of four parameters can be estimated, one of which is the overall grand mean, the other is treatment effect, and the third is period effect, so that we can only estimate a fourth parameter, be it either sequence or carryover, but not both; see Jones & Kenward 1989 and Ratkowsky et al. 1993; the assumption of no sequence effects, though, could be defended as a reasonable one when there has been randomization of subjects to sequences). A nonparametric version of these tests was first described by Koch (1972) and is explained in Jones & Kenward (1989, pp. 51ff.) (but see Taulbee 1982 for corrections of expressions to 4 and 6 in Koch 1972 and Jones & Kenward 1989, pp. 27 and 56).

As an example, Table 2 shows a set of data from an AB,BA trial (these are simulated data, from a model with main effects of period and treatment and normally distributed random subject effects and random errors). Using the paired t test approach to test for treatment differences we obtain $t_9=1.098$, $P=0.3$. Using the Hills–Armitage approach we obtain $t_8=5.666$, $P=0.0005$ (with the Hills–Armitage approach we have one less df as this is a two-sample t test). In this example the paired t test fails because there are period effects, whereas the Hills–Armitage approach has no problems with the period effects.

We can also analyse these data using a split-plot ANOVA (Table 3; see Jones & Kenward 1989, pp. 30–33). The first stratum is individual; the second is within individual and is used for the tests of interest (direct treatment effects). In this ANOVA, we use as explanatory

Table 3. ANOVA table for the analysis of the data in Table 2 using split-plot (parameterization as in Jones & Kenward 1989, except no carryover included)

Source	*df*	Sums of squares	Mean squares	*F*	*P*
Between-subjects	9	130			
Within-subjects stratum					
Period (adjusted for treatment)	1	99.5	99.5	231.7	0.0001
Treatment (adjusted for period)	1	13.8	13.8	32.1	0.0001
Within-subjects residuals	8	3.4	0.4		
Total	19	246.7			

or independent variables treatment and period, and test for direct treatment effects after having entered period in the model (and for period after entering treatment); these are called marginal tests. In this ANOVA we have adjusted for the effects of period by incorporating period into the model, and thus we obtain the same results as the Hills–Armitage approach ($F_{1,8}=32.1=5.666^2=t_8^2$). (However, an ANOVA that did not include period would yield the same incorrect results as the paired *t* test.)

The problem of the paired comparison is the same regardless of whether we use a *t* test, a nonparametric test, or a randomization test. The cause of the problem is not the type of statistic but failure to account for the effect of period. Unless there is strong evidence to the contrary, in the majority of behavioural experiments we should assume that period can affect the results: there is little to lose from making an allowance for period effects even if there are none, but if there are period effects, and we do not account for them, all our inferences could be seriously affected. Furthermore, there are biological reasons why we could expect period effects (growth, repeated use of a specific measuring device), and period effects seem common in previous behavioural studies, suggesting that they might be widespread. In the presence of period effects, a paired test should not be used because it is inappropriate, regardless of whether or not counterbalancing is used and whether or not there are the same number of subjects in each sequence. Problems with period effects are not limited to two-treatment crossover designs, but affect all other designs as well (e.g. three-treatment designs).

Carryover Effects

In the presence of carryover effects, the response to a treatment is affected by what treatment was applied in previous period(s). In this situation, past treatments have effects that last, or carryover, to the following periods. Carryover effects can bias the estimates of direct treatment effects and affect designs with any number of periods and treatments. The cause of the problem is generally not carryover per se, but differential carryover effects, that is, the carryover from different treatments being different. (For example, in Table 1, if there are differential carryover effects, our estimate of direct treatment effects using the Hills–Armitage approach will be biased by $(\lambda_1 - \lambda_2)$, where (λ_1) is the carryover effect of treatment A on treatment B, and (λ_2) is the carryover effect of treatment B on treatment A; if there are equal carryover effects, they will be indistinguishable from period effects, and the Hills–Armitage approach will be unbiased). Contrary to what is sometimes believed, counterbalancing does not eliminate bias caused by carryover effects, regardless of the number of treatments (e.g. Abeyasekera & Curnow 1984).

There are two strategies for dealing with carryover effects: (1) minimize the chances that they can happen by allowing enough time (washout periods) between successive treatments; and (2) include them explicitly in the statistical model. Which of these approaches is taken will affect both the design of the experiment and the analysis of the data. There is considerable debate about which approach should be taken, and readers might want to read some of the discussions (for example, Abeyasekera & Curnow 1984; Senn 1993a, pp. 14–15, 52–54 and chapter 10; Grieve & Senn 1998; Jones & Wang 1998; Koch 1998). A practical solution might be as follows: first, design studies so that carryover effects are unlikely (i.e. use long enough washout periods). Second, design experiments so that carryover effects can be included in the statistical model. If carryover turns out to be present, a design that made a provision for carryover would make it possible to salvage the experiment, and would indicate that future experiments might need to increase the washout period.

Furthermore, from a behavioural perspective, in some studies the presence of carryover effects after what was considered a sufficiently long washout period could reveal a phenomenon of interest in its own right, since a carryover effect would indicate that a past experience is much longer lasting than expected (e.g. effects of prior defeats in aggressive encounters that affect fight performance more than 24 h after the defeat). Finally, in some instances we might combine crossover designs with between-subject designs: an interaction between carryover and between-subjects treatment might indicate a potentially interesting biological phenomenon. For instance, we might examine simultaneously the effect of hormonal treatment (a between-subject treatment) and effects of presentation of a female versus a control (using a crossover trial). In this study, an interaction between carryover and hormone treatment would suggest that hormonal treatment has affected how long lasting the presentation of a female is.

Table 4. Methods of analysis of crossover trials of special interest for animal behaviour research: a quick guide to the literature (for details, see Díaz-Uriarte 2001)

Type of data	Type of analysis	Methods and reference
Metric response	Nonparametric, robust, and randomization	Within-individual contrasts [which are essential for all nonparametric and many multivariate tests] (Hafner et al. 1988; Jones & Kenward 1989, pp. 23–28 and 60–65; Senn 1993a, pp. 42–44 and 238–248); *t* test, Wilcoxon and similar (Jones & Kenward 1989, pp. 51–60; Senn 1993a, page 93; Tudor & Koch 1994); randomization (Shen & Quade 1983; examples in Díaz-Uriarte 1999); blocking and among-subject treatments (Elswick & Uthoff 1989; Tudor & Koch 1994; for background on Mantel–Haenszel test: Koch & Edwards 1988, page 418; Agresti 1990, page 283; for background on randomization tests and blocking: Noreen 1989, page 28; Edgington 1995, page 131; Maritz 1995, page 191); more than two treatments (Shen & Quade 1983; Koch & Edwards 1988; Peace & Koch 1993; Senn 1993a; Bellavance & Tardif 1995; Ohrvick 1998)
	Linear mixed-effects models	Linear mixed-effects models (Lindsey 1993, pp. 136 ff.; Diggle et al. 1994, chapters 4 and 5; Littell et al. 1996, pp. 392 ff.; Vonesh & Chinchilli 1997, chapter 4; McCulloch & Searle 2000; Pinheiro & Bates 2000); parameterization (Jones & Kenward 1989, page 30; Crowder & Hand 1990, page 107; Lindsey 1993, pp. 15 and 135; Ratkowsky et al. 1993, chapter 3; Diggle et al. 1994, chapter 4 and page 156; Littell et al. 1996, page 392; Vonesh & Chinchilli 1997, chapter 4); degrees of freedom (Jones & Kenward 1989, page 141)
Categorical data	Nonparametric-like methods	2×2 trial, Mainland-Gart test and Prescott's test (Fidler 1984; Jones & Kenward 1989, pp. 89–105; Crowder & Hand 1990, pp. 109–110; Senn 1993a, pp. 106–109); three or more treatments (Senn 1993a, pp. 153–155); ordinal data (Ezzet & Whitehead 1991, 1993; Senn 1993a, pp. 109–113, 1993b; Tudor & Koch 1994, pp. 359–361; Jung & Koch 1999)
	Explicitly model-based methods	Generalized linear models, general (McCullagh & Nelder 1989; Agresti 1990; Dobson 1990; Crawley 1993; McCulloch & Searle 2000); generalized linear mixed models: subject-specific versus marginal or population averaged (Zeger et al. 1988; Liang et al. 1992; Lindsey 1993, chapter 2; Diggle et al. 1994, chapter 7; Kenward & Jones 1994; Albert 1999); generalized estimating equations (GEE: Albert 1999; Horton & Lipsitz 1999); examples with crossover trials (Agresti 1993; Diggle et al. 1994, pp. 154–159 and pp. 175–181; Lindsey 1993, pp. 201–204; Kenward & Jones 1994)
Censored observations	Nonparametric-like methods	Methods (Tudor & Koch 1994, page 365; Feingold & Gillespie 1996); examples (Díaz-Uriarte 1999)
	Survival analysis	General introductions (Kalbfleisch & Prentice 1980; Lawless 1982; Lee 1992; Collett 1994; Klein & Moeschberger 1997; Therneau & Grambsch 2000); repeated time to event data: marginal models (Wei et al. 1989; Lee et al. 1992; Lin 1993, 1994; Hougaard 2000; Therneau & Grambsch 2000); repeated time to event data: frailty models (Hougaard 2000; Therneau & Grambsch 2000); repeated time to event data: log-linear approach (Lindsey et al. 1996)
Multiple responses: multivariate data	Nonparametric methods for metric responses	Within-individual contrasts (Patel & Hearne 1980; O'Brien 1984; Johnson & Grender 1993; Johnson & Mercante 1996; examples in Díaz-Uriarte 1999)
	Metric responses	2×2 design (Rodríguez-Carvajal & Freeman 1999); other designs (Grender & Johnson 1993, pp. 71–74 and 84); multivariate linear mixed model (Galecki 1994)
	Categorical data	See GEE and generalized mixed models above
	Censored observations	See survival analysis above
Repeated measures within periods		Within-individual contrast approaches (Patel & Hearne 1980; Jones & Kenward 1989, chapter 6; Grender & Johnson 1993; Rodríguez-Carvajal & Freeman 1999, page 399); split-plot in time repeated measures ANOVA and linear mixed-effects models (Galecki 1994; Littell et al. 1996, pp. 388 ff.)

Design and Analysis of Crossover Trials

Details of the design of crossover trials and the assignment of subjects to sequences are provided in Jones & Kenward (1989), Ratkowsky et al. (1993), Senn (1993a) (additional issues are discussed in Donev 1998; Vonesh & Chinchilli 1997; Jones & Donev 1996). Here, I will only highlight a few major points. First, during the design phase, it is essential to understand how the data will be analysed. For example, some nonparametric methods for more than two treatments require that the designs be of a specific kind or that allocation of subjects be done in a particular way; some other methods work only with large sample sizes. Second, for two-treatment trials there are many designs available besides the AB,BA; in particular, designs such as [ABB,BAA], [ABBA,BAAB], or [ABBA, BAAB, AABB, BBAA] perform well under a variety of assumptions and also allow one to use simple and robust analyses based on within-individual comparisons (see Jones & Kenward 1989; Senn 1993a). Third, for experiments that involve more than two treatments, a variety of designs are available and some of these designs (Senn 1993a, pp. 122 and 123) have special properties that allow us to use some nonparametric and multivariate analyses. Finally, designs for more than three treatments will require sample sizes larger than those available in most behavioural studies.

General references on the analysis of crossover trials are Jones & Kenward (1989), Ratkowsky et al. (1989), Senn (1993a), Vonesh & Chinchilli (1997); Tudor & Koch (1994) emphasize nonparametric methods. However, in many behavioural experiments researchers often record data (such as categorical data or censored time to event data) that might not allow us to use standard parametric analyses, and frequently measure several response variables that ought to be analysed with multivariate techniques. These cases require the use of very specific techniques. A recent review of many of these methods, oriented towards behavioural research, is Díaz-Uriarte (2001); Table 4 presents a quick guide to the literature.

Conclusions

Crossover designs can result in an increase in statistical power and reduce the number of animals needed in a study, which is particularly important if there are ethical concerns or we are working with small or threatened populations. However, the analysis of crossover trials tends to be more complicated than the analysis of parallel trials, and the potential for aliasing of effects (which prevents estimation of all the parameters of the model) in crossover designs is larger; in addition, crossover trials require that subjects be used repeatedly. Thus, election of crossover designs versus parallel trials will have to consider how costly it is to obtain new subjects versus how costly it is to obtain repeated measures of the same subject. Additional (but rarely available) information on within- versus between-individual variance would allow more informed choices between crossover and parallel group designs (see details in Senn 1993a, chapter 9; for a short explanation of why crossover designs are often more powerful than parallel group designs, see Crowder & Hand 1990, page 101).

In many studies conducted in the laboratory or in field enclosures that require lengthy training or habituation of animals, crossover trials are good choices (if not the only option). In some field studies, relocating subjects might be too time consuming compared to finding new ones. Even when subjects are individually marked and easy to relocate, crossover designs might be difficult to use in field conditions: the assignment of subjects to sequences will have been done before the animals are actually found on a particular day, and for period to have the same meaning across subjects, the time interval between periods should be comparable among animals. These conditions might impose too many constraints on which particular animals need to be found on a particular day, and could make crossover designs less attractive in field conditions.

The type of response must be considered when choosing an experimental design, because certain types of analyses are very difficult or impossible with some designs, and some types of data might require special analyses. These problems can be detected during the design stages (i.e. before any data have been gathered) and could prompt a change in the design.

In summary, I have argued that: (1) appropriate analysis of crossover trials is crucial to obtain valid conclusions from many behavioural experiments; (2) we will (virtually) always have to include period in our statistical analyses, because of the serious consequences of incorrectly assuming there are no period effects; (3) a large number of crossover designs are available for behavioural studies, and when using two treatments we might not want to limit ourselves to the 2×2 design; (4) we need to think about carryover effects and what constitutes an appropriate washout period; how we are dealing with period and carryover effects should be made explicit.

C. Lázaro-Perea provided advice, discussion and comments on the manuscript. A. R. Ives, B. Jones, J. K. Lindsey, C. A. Marler, E. V. Nordheim, C. T. Snowdon, B. C. Trainor and two anonymous referees provided comments that have significantly improved the manuscript. E. V. Nordheim suggested that this paper should be written. The book by B. Jones & M. G. Kenward is a delightful and amazing book that opened my eyes to crossover trials (and taught me a lot of what I know about them). I wrote this manuscript using LyX, an open-source (and free) LaTeX based document processor released under the GNU GPL licence. Manuscript preparation was partially supported by a John and Virginia Emlen Graduate Student Fellowship, Department of Zoology, University of Wisconsin-Madison.

References

Abeyasekera, S. & Curnow, R. N. 1984. The desirability of adjusting for residual effects in a crossover design. *Biometrics*, **40,** 1071–1078.

Agresti, A. 1990. *Categorical Data Analysis*. New York: J. Wiley.

Agresti, A. 1993. Distribution-free fitting of logit models with random effects for repeated categorical responses. *Statistics in Medicine*, **12**, 1969–1987.

Albert, P. S. 1999. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine*, **18**, 1707–1732.

Bailey, N. T. J. 1995. *Statistical Methods in Biology*. 3rd edn. Cambridge: Cambridge University Press.

Bart, J., Flinger, M. A. & Notz, W. I. 1998. *Sampling and Statistical Methods for Behavioral Ecologists*. Cambridge: Cambridge University Press.

Becker, M. P. & Balagtas, C. C. 1993. Marginal modeling of binary cross-over data. *Biometrics*, **49**, 997–1009.

Bellavance, F. & Tardif, S. 1995. A nonparametric approach to the analysis of three-treatment three-period crossover designs. *Biometrika*, **82**, 865–875.

Campbell, R. C. 1989. *Statistics for Biologists*. 3rd edn. Cambridge: Cambridge University Press.

Collett, D. 1994. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

Crawley, M. J. 1993. *GLIM for Ecologists*. Oxford: Blackwell.

Crowder, M. J. & Hand, D. J. 1990. *Analysis of Repeated Measures*. New York: Chapman & Hall.

Díaz-Uriarte, R. 1999. Anti-predator behaviour changes following an aggressive encounter in the lizard *Tropidurus hispidus*. *Proceedings of the Royal Society of London, Series B*, **266**, 2457–2464.

Díaz-Uriarte, R. 2001. The analysis of cross-over trials in animal behavior experiments: review and guide to the statistical literature. Samizdat Press. http://samizdat.mines.edu/.

Diggle, P. J., Liang, K.-Y. & Zeger, S. L. 1994. *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Dobson, A. J. 1990. *An Introduction to Generalized Linear Models*. London: Chapman & Hall.

Donev, A. N. 1998. Crossover designs with correlated observations. *Journal of Biopharmaceutical Statistics*, **8**, 249–262.

Edgington, E. S. 1995. *Randomization Tests*. 3rd edn. New York: Marcel Dekker.

Elswick, R. K. & Uthoff, V. A. 1989. A nonparametric approach to the analysis of the 2-treatment, 2-period, 4-sequence crossover model. *Biometrics*, **45**, 663–667.

Ezzet, F. & Whitehead, J. 1991. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*, **10**, 901–906.

Ezzet, F. & Whitehead, J. 1993. A random effects model for ordinal responses from a crossover trial: Reply. *Statistics in Medicine*, **12**, 2150–2151.

Feingold, M. & Gillespie, B. W. 1996. Cross-over trials with censored data. *Statistics in Medicine*, **15**, 953–967.

Fidler, V. 1984. Change-over clinical trial with binary data: mixed-model-based comparison of tests. *Biometrics*, **40**, 1063–1070.

Galecki, A. T. 1994. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics. Theory and Methods*, **23**, 3105–3119.

Grender, J. M. & Johnson, W. D. 1993. Analysis of crossover designs with multivariate response. *Statistics in Medicine*, **12**, 69–89.

Grieve, A. & Senn, S. 1998. Estimating treatment effects in clinical crossover trials. *Journal of Biopharmaceutical Statistics*, **8**, 191–233; discussion 235–247.

Hafner, K. B., Koch, G. G. & Canada, A. T. 1988. Some analysis strategies for three-period changeover designs with two treatments. *Statistics in Medicine*, **7**, 471–481.

Horton, N. J. & Lipsitz, S. R. 1999. Review of software to fit Generalized Estimating Equation regression models. *American Statistician*, **53**, 160–169.

Hougaard, P. 2000. *Analysis of Multivariate Survival Data*. New York: Springer-Verlag.

Johnson, W. D. & Grender, J. M. 1993. Multivariate nonparametric analysis for the two-period crossover design with application in clinical trials. *Journal of Biopharmaceutical Statistics*, **3**, 1–12.

Johnson, W. D. & Mercante, D. E. 1996. Analyzing multivariate data in crossover designs using permutation tests. *Journal of Biopharmaceutical Statistics*, **6**, 327–342.

Jones, B. & Donev, A. N. 1996. Modelling and design of cross-over trials. *Statistics in Medicine*, **15**, 1435–1446.

Jones, B. & Kenward, M. G. 1989. *Design and Analysis of Cross-Over Trials*. New York: Chapman & Hall.

Jones, B. & Wang, J. 1998. Comments on 'Estimating treatment effects in clinical crossover trials'. *Journal of Biopharmaceutical Statistics*, **8**, 235–238.

Jung, J. W. & Koch, G. G. 1999. Multivariate non-parametric methods for Mann-Whitney statistics to analyse cross-over studies with two treatment sequences. *Statistics in Medicine*, **18**, 989–1017.

Kalbfleisch, J. D. & Prentice, R. L. 1980. *The Statistical Analysis of Failure Time Data*. New York: J. Wiley.

Kenward, M. G. & Jones, B. 1994. The analysis of binary and categorical data from crossover trials. *Statistical Methods in Medical Research*, **3**, 325–344.

Klein, J. P. & Moeschberger, M. L. 1997. *Survival Analysis*. New York: Springer-Verlag.

Koch, G. G. 1972. The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics*, **28**, 577–584.

Koch, G. G. 1998. Comments on 'Estimating treatment effects in clinical crossover trials'. *Journal of Biopharmaceutical Statistics*, **8**, 239–242.

Koch, G. G. & Edwards, S. 1988. Clinical efficiency trials with categorical data. In: *Biopharmaceutical statistics for drug development* (Ed. by K. E. Peace), pp. 403–457. New York: Marcel Dekker.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data*. New York: J. Wiley.

Lee, E. T. 1992. *Statistical Methods for Survival Data Analysis*. New York: J. Wiley.

Lee, E. W., Wei, L. J. & Amato, D. A. 1992. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: *Survival Analysis: State of the Art* (Ed. by J. P. Klein & P. K. Goel), pp. 237–247. Dordrecht: Kluwer Academic.

Lehner, P. N. 1979. *Handbook of Ethological Methods*. New York: Garland Press.

Liang, K.-Y., Zeger, S. L. & Qaqish, B. 1992. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society B*, **54**, 3–40.

Lin, D. Y. 1993. MULCOX2: a general computer program for the Cox regression analysis of multivariate failure time data. *Computer Methods and Programs in Biomedicine*, **40**, 279–293.

Lin, D. Y. 1994. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, **13**, 2233–2247.

Lindsey, J. K. 1993. *Models for Repeated Measurements*. Oxford: Clarendon Press.

Lindsey, J. K., Jones, B. & Lewis, J. A. 1996. Analysis of cross-over trials for duration data. *Statistics in Medicine*, **15**, 527–535.

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. 1996. *SAS System for Mixed Models*. Cary, North Carolina: SAS Institute.

McCullagh, P. & Nelder, J. A. 1989. *Generalized Linear Models*. 2nd edn. New York: Chapman & Hall.

McCulloch, C. E. & Searle, S. R. 2000. *Generalized, Linear, and Mixed Models*. New York: J. Wiley.

Maritz, J. S. 1995. *Distribution-Free Statistical Methods*. 2nd edn. London: Chapman & Hall.

Martin, P. & Bateson, P. 1993. *Measuring Behaviour*. 2nd edn. Cambridge: Cambridge University Press.

Noreen, E. W. 1989. *Computer-intensive Methods for Testing Hypotheses: an Introduction*. New York: J. Wiley.

O'Brien, P. C. 1984. Procedures for comparing samples with multiple endpoints. *Biometrics*, **40,** 1079–1087.

Ohrvik, J. 1998. Nonparametric methods in crossover trials. *Biometrical Journal*, **40,** 771–789.

Patel, H. I. & Hearne III, E. M. 1980. Multivariate analysis for the two-period repeated measures crossover design with application to clinical trials. *Communications in Statistics. Theory and Methods*, **A9,** 1919–1929.

Peace, K. E. & Koch, G. G. 1993. Statistical methods for a three-period crossover design in which high dose cannot be used first. *Journal of Biopharmaceutical Statistics*, **3,** 103–116.

Pinheiro, J. C. & Bates, D. M. 2000. *Mixed-Effects Models in S and S-Plus*. New York: Springer-Verlag.

Ratkowsky, D. A., Evans, M. A. & Alldredge, J. R. 1993. *Cross-Over Experiments: Design, Analysis, and Application*. New York: Marcel Dekker.

Rodriguez-Carvajal, L. A. & Freeman, G. H. 1999. Multivariate AB-BA crossover design. *Journal of Applied Statistics*, **26,** 393–403.

Schneider, B. 1983. Crossover designs and repeated measurements. *Neuropsychobiology*, **10,** 49–55.

Senn, S. 1993a. *Cross-Over Trials in Clinical Research*. New York: J. Wiley.

Senn, S. 1993b. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*, **12,** 2147–2151.

Shen, C. D. & Quade, D. 1983. A randomization test for a three-period three-treatment crossover experiment. *Communications in Statistics. Simulation and Computation*, **12,** 183–199.

Sokal, R. R. & Rohlf, F. J. 1995. *Biometry*. 3rd edn. New York: W. H. Freeman.

Taulbee, J. D. 1982. A note on the use of nonparametric methods in the statistical analysis of the two-period changeover design. *Biometrics*, **38,** 1053–1055.

Therneau, T. & Grambsch, P. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Tudor, G. & Koch, G. G. 1994. Review of nonparametric methods for the analysis of crossover studies. *Statistical Methods in Medical Research*, **3,** 345–381.

Vonesh, E. F. & Chinchilli, V. M. 1997. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

Wei, L. J., Lin, D. Y. & Weissfeld, L. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84,** 1065–1073.

Zar, J. H. 1996. *Biostatistical Analysis*. 3rd edn. Englewood Cliffs, New Jersey: Prentice Hall.

Zeger, S. L., Liang, K.-Y. & Albert, P. S. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44,** 1049–1060.

Zolman, J. F. 1993. *Biostatistics*. New York: Oxford University Press.