J. Dairy Sci. 90(E. Suppl.):E87–E99
doi:10.3168/jds.2006-612

Design and Analysis of Pen Studies in the Animal Sciences[1,2]

N. R. St-Pierre[3]
Department of Animal Sciences, The Ohio State University, Columbus 43210

ABSTRACT

Increasingly, research is being performed in which animals subjected to a common treatment are also housed in a common pen. Issues have been raised regarding the proper planning of experiments and conduct of statistical analyses in these instances. This paper reviews the problems associated with ignoring animal grouping during data analyses, and examples are provided for appropriate methods to use when animals are grouped in pens. Using animals as the error term when treatments are applied to pens can result in biased estimates of treatment effects when pens are of unequal sizes and animals are moved in and out of the pens. It always results in biased probability statements regarding their significance. The pen effect includes systematic effects other than that of the treatment, which is why pens must be replicated and randomized. In essence, pen studies have an implicit split-plot design in which the main plots (pens) receive the treatments of interest, whereas the subplots (cows) receive all the same subplot treatment. Using the subplot error to test main-plot treatment effects creates inflated degrees of freedom and uses the wrong denominator mean square to test the effect; hence, severely biasing the test of significance for the treatment effects and resulting in an invalid causal inference base. The interactions of pens with the fixed-effect elements of the treatment design are the correct error terms for those fixed-effects factors applied to the pens. The same statistical designs used with animals as experimental units can be used with pens. The number of experimental units to achieve a given power can be, and generally is, considerably less with pens because the variance among pens is generally less than the variance of cows within pens. Pens must be replicated, randomized, and included in the statistical model to ensure valid statistical inference.

Received September 19, 2006.
Accepted April 3, 2007.
[1]Presented at the ADSA-ASAS Joint Annual Meeting, Minneapolis, MN, July 2006.
[2]Salaries and research support were provided by state and federal funds appropriated to the Ohio Agricultural Research and Development Center, the Ohio State University. Manuscript No. 02-07AS.
[3]E-mail: st-pierre.8@osu.edu

Key words: pen studies, experimental unit, randomization, causal inference

INTRODUCTION

Historically, most of the research in dairy production has been conducted with animals that were individually housed and fed. Just a few decades ago, tie-stall housing was the predominant type of housing used in commercial dairy operations in the United States, and the applicability of the research results to different housing environments was not questioned. Substantial changes in animal housing have occurred in US commercial operations. The USDA estimates that over 75% of US milk production occurred in herds of 100 and more cows in 2003, up from 33% in 1980 (USDA, 2006). The majority of cows in larger herds are housed either in free stalls or in dry lots, grouped in pens containing numerous cows. This has raised concerns regarding the applicability of results from studies using individually housed cows for the commercial operations in which cows are housed in groups. Consequently, some public and private dairy research organizations have modified, or are considering modifying, the housing used for their research herds to one where cows are housed in groups.

Simultaneously, there appear to be greater opportunities to conduct research in large commercial herds where cows are invariably grouped in medium to large size pens. The grouping of cows during the conduct of an experiment has a significant effect on how data should be analyzed to reach valid scientific conclusions (Gill, 1987, 1989). The debate regarding proper statistical models and methods of data analyses to be used with pen studies has been marred by a lack of understanding of a few fundamental statistical concepts such as experimental units, degrees of freedom, and randomization. The frequent but erroneous notion that the smallest unit upon which a measurement is made can serve as the identifier of the experimental unit has infiltrated the discussion, although there is no theoretical basis in the statistical literature to support this position. In this paper, the statistical concepts underlying pen studies are first introduced using an intuitive approach (i.e., using examples), followed by a discussion on the fundamental

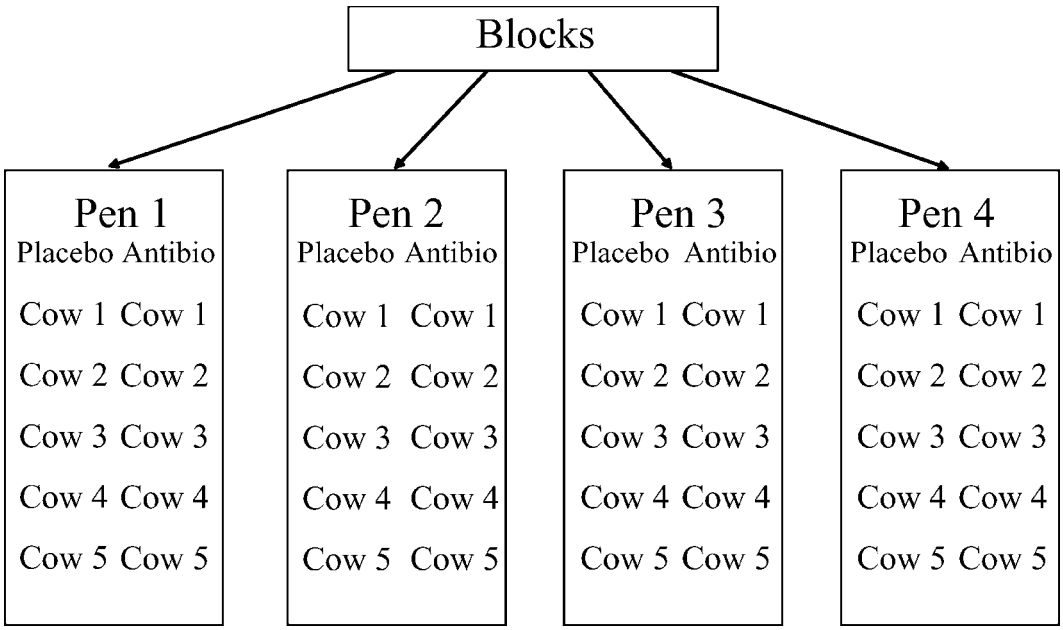

Figure 1. Schematic of the randomized complete block design used in experiment 1.

statistical issues related to the analyses of such studies, including causal inference range. A few useful statistical designs that can be used with pen studies are presented, including in each case the programming statements to be used with SAS software (SAS Institute, 2004), which is the predominant software used for data analyses in papers published by the *Journal of Dairy Science*.

INTUITIVE APPROACH

Series of Experiments

To understand the issues involved with pen studies, a series of virtual experiments is presented, defining in the process the correct models to be used for the statistical analyses of the data.

Experiment 1. The objective of the first experiment is to determine the efficacy of a new antibiotic in dairy cattle. The treatment structure is a simple one-way structure with 2 levels: a placebo and an antibiotic injection. Forty cows in 4 separate pens of 10 cows each are used to test the hypothesis. Cows are first assigned at random to each of the 4 pens. Within each pen, 5 cows are assigned at random to the placebo injection, and 5 cows are assigned to the antibiotic injection. The general structure of this experiment is shown in Figure 1. Cows within the same pen have something more in common than cows across different pens. For example, they share the same microclimate, they are milked at the same time, they are simultaneously restrained for pregnancy check, and so on. Theoretically, there is a near-infinite list of known and unknown factors that contribute to cows within pens having more in common than cows across pens. This commonality must be accounted for during the statistical analysis. One should recognize that the experimental design used is a randomized complete block design with subsampling (**RCBD**) with pens acting as blocking factors (Damon and Harvey, 1987). The statistical model for observed responses is

$$y_{ijk} = \mu + \alpha_i + p_j + \alpha p_{ij} + \varepsilon_{ijk}; \quad [1]$$
$$i = 1, 2; j = 1, 2, 3, 4; k = 1, 2, ..., 5$$

where y_{ijk} are the observed values, μ is the overall mean, α_i denotes the fixed effect of the ith injection treatment, p_j is the random block effect associated with the jth pen, where the block effects are assumed to be independently and identically distributed (**iid**) $N(0, \sigma^2_p)$, αp_{ij} is the random interaction effect associated with the jth pen and the ith injection treatment, assumed iid $N(0, \sigma^2_{ap})$, and ε_{ijk} denotes the random error, which is assumed to be iid $N(0, \sigma^2_\varepsilon)$.

Certainly, the interest is in making inference beyond the narrow range of the 4 pens used in the experiment. Pens are thus considered as a random sample from a large population of pens (i.e., levels of pen effects come from a probability distribution). Even though the variance components in this model would generally be esti-

Table 1. Schematic of the ANOVA table for experiment 1

Source	df	Effect type	Error term
Pen	3	Random	Pen × Treatment
Treatment	1	Fixed	Pen × Treatment
Pen × Treatment	3	Random	Cow (Pen × Treatment)
Cow (Pen × Treatment)	32	Random	
Total	39		

mated using (restricted) maximum likelihood, inference on treatment effects is generally based on classical ANOVA. It is thus useful to sketch an ANOVA table, calculate the degrees of freedom (df), and identify the correct error terms for each effect to be tested (Table 1). In [1], the usual notation for the residual error was used. In all statistical analyses, this residual error term truly has an identity. In this first experiment, the residual error is really the nested effect of *Cow* (*Pen* × *Treat*) (Table1). Each column of cows in Figure 1 represents a level of this nested factor. There are 5 cows in each column; hence, each column has 4 df. There are 8 subclasses of *Pen* × *Treatment*. Thus, there are 8 × 4 = 32 df for *Cow* (*Pen* × *Treat*), a value that corresponds to the df for the error reported by statistical software (e.g., PROC MIXED; SAS Institute, 2004). The error term for *Treatment*, however, is not the residual error; rather, it is the interaction of *Pen* × *Treatment* (St-Pierre and Jones, 1999). The *Pen* × *Treatment* effect is not part of the design structure but is an essential part of the error structure.

Experiment 2. This experiment is identical to experiment 1 except that 2 of the 4 pens in experiment 2 were randomly selected to be cooled with fans and sprinklers (cooled), whereas the 2 other pens were selected to be uncooled (control). The objectives of the experiment are to determine the efficacy of a new antibiotic treatment, the effect of cooling, and to determine whether the efficacy of the new antibiotic treatment is dependent on whether the animals are cooled or not. The general structure of this experiment is shown in Figure 2. There are still 4 pens as in experiment 1, but the labeling of the pens must now be changed to reflect that they are no longer sampled from a simple population of pens; in fact, 2 pens were sampled from a population of uncooled pens, and 2 were sampled from a population of cooled pens.

All experimental designs consist of 3 structures: treatment structure, design structure, and error structure (Milliken and Johnson, 1992). The treatment structure of an experimental design consists of the sets of treatments, treatment combinations, or populations that the experimenter has selected to study. The design structure of an experimental design consists of the grouping of the experimental units into homogeneous groups or blocks. The interactions between elements of the design structure and the treatment structure form the error structure(s) (Milliken and Johnson, 1992). In experiment 2, the treatment design is a 2 × 2 factorial in which the main factors are the cooling treatments (control vs. cooled), and the injection treatments (placebo vs. antibiotic). In this experiment the design structure is associated with the 4 pens. It must be recognized that although the cows within each pen are randomly assigned to the injection treatments, it is the pens that were randomly assigned to the cooling treatments. Thus, the experimental units are different for the 2 factors. The experiment design is a classic split plot (Damon and Harvey, 1987) with the following statistical model:

$$y_{ijkl} = \mu + \delta_k + p_{j:k} + \alpha_i + \delta\alpha_{ik} + \alpha p_{ij:k} + \varepsilon_{ijkl}; \quad [2]$$
$$i = 1, 2; j = 1, 2; k = 1, 2; l = 1, 2, ..., 5$$

where δ_k denotes the fixed effect of the kth cooling treatment, $p_{j:k}$ is the random effect associated with the jth pen nested within the kth cooling treatment, assumed iid N(0, $\sigma^2_{p)}$, α_i denotes the fixed effect of the ith injection treatment, $\delta\alpha_{ik}$ is the fixed effect of the interaction between the kth cooling treatment and the ith injection treatment, and $\alpha p_{ij:k}$ is the random interaction effect associated with the jth pen within the kth cooling treatment and the ith injection treatment, assumed iid N(0, σ^2_{ap}).

The corresponding ANOVA is presented in Table 2. In essence, a split plot is a repeated measurement design in space (the pens are repeatedly measured in space) with a compound symmetry covariance of errors to account for the fact that cows within a pen have more in common than cows across pens (i.e., they are correlated). The split plot is the result of a merging of 2 experiments, each with its own design. In this instance, the main-plot experiment consists of 4 experimental units (the pens) assigned at random to 1 of 2 cooling treatments. The statistical design for the main plot is thus a completely randomized design. The subplot design, as explained previously, is an RCBD with subsampling, with cows nested within pens as the experimental units. As shown in Table 2, the error term for testing the cooling treatment is *Pen* (*Cooling*). The error term for testing the injection treatment and its interaction with the cooling treatment is *Injection* × *Pen* (*Cooling*). The terms *Pen* (*Cooling*) and *Injection* × *Pen* (*Cooling*) represent the interactions between elements of the design structure and the treatment structure, which is why they are the correct error terms to be used. They form what Milliken and Johnson (1992) called the error structure of the experimental design. Some statisticians would argue that elements of the error structure can be pooled with the residual error. Yandell (1997) has explained the

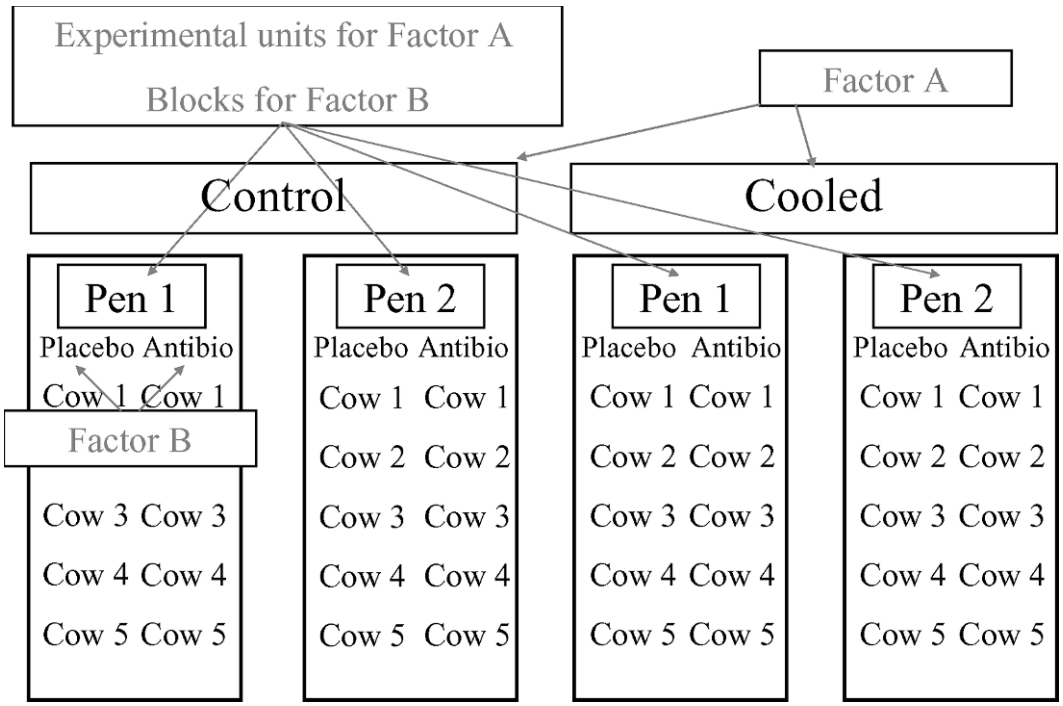

Figure 2. Schematic of the split-plot design used in experiment 2.

problems with pooling of effects. In essence, either there is too much risk in pooling (i.e., the chance of a type II error is large) or there is no benefit to it (i.e., it has little effect on the tests). Milliken and Johnson (2002) pointed out that blocking imposes a restriction on the randomization, which is something that cannot be ignored during the analysis. Because randomization was done between and within pens, the term *Pen* (*Cooling*) is an explicit component of the design structure. Thus, *Pen* (*Cooling*) should never be pooled with the residual error term. Most statisticians would also object to the pooling of *Injection* × *Pen* (*Cooling*) with the error term because of the potential for large type II errors in the pooling decision, and inflated (i.e., incorrect) type I error when testing the effect of *Injection* × *Cooling*.

Comparing the df in Tables 1 and 2, one can observe that the 3 df associated with the pen effect in experiment 1 are split into 2 components in experiment 2: 1 df for *Cooling* and 2 df for *Pen* (*Cooling*). Likewise, the 3 df for *Pen* × *Injection* in experiment 1 are now split into 1 df for *Injection* × *Cooling* and 2 df *for Injection* × *Pen* (*Cooling*) in experiment 2.

Experiment 3. This experiment is identical to experiment 2, and it is performed as such. At the conclusion of the experiment, though, the researchers realize that all the syringes contained the placebo; none contained the new antibiotic. The treatment design is thus a 2 × 1 factorial, with 2 levels of cooling and 1 level of injection. The experimental design is still a split plot (the randomization process has not been modified from experiment 2) with 2 levels of cooling as the first main factor applied to the main plot and 1 level of injection as the second factor applied to the subplot. In this situation, the design

Table 2. Schematic of the ANOVA table for experiment 2

Source	df	Effect type	Error term
Cooling	1	Fixed	Pen (Cooling)
Pen (Cooling)	2	Random	Cow (Cooling × Pen × Injection)
Injection	1	Fixed	Injection × Pen (Cooling)
Injection × Cooling	1	Fixed	Injection × Pen (Cooling)
Injection × Pen (Cooling)	2	Random	Cow (Cooling × Pen × Injection)
Cow (Cooling × Pen × Injection)	32	Random	
Total	39		

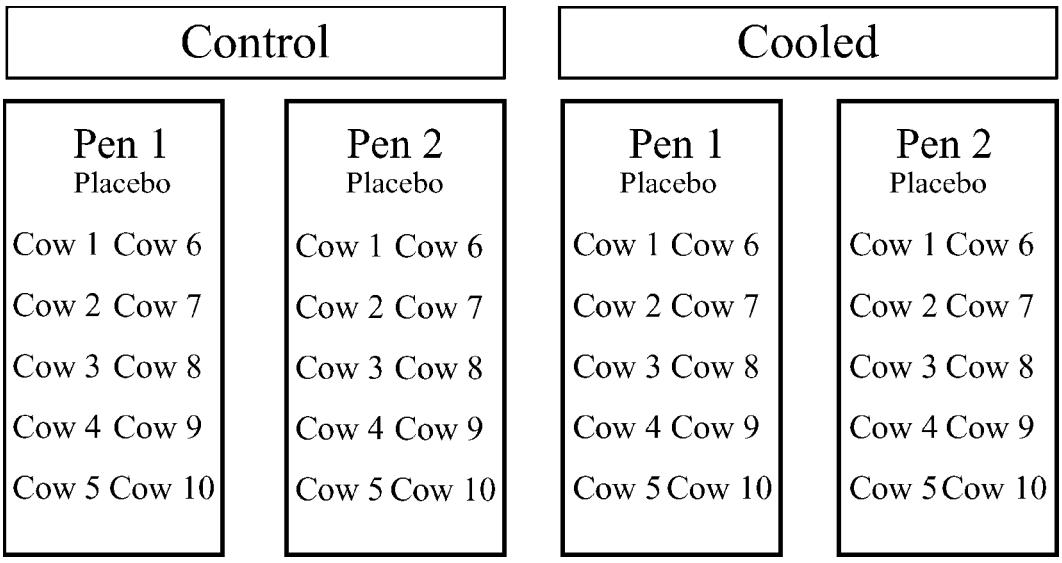

Figure 3. Schematic of the research design used in experiment 3, a split-plot design with only 1 level of the second factor treatment in the subplots.

is better known as a nested design in the statistical literature. The statistical model is thus

$$y_{ijkl} = \mu + \delta_k + p_{j:k} + \alpha_i + \delta\alpha_{ik} + \alpha p_{ij:k} + \varepsilon_{ijkl}; \quad [3]$$
$$i = 1; j = 1, 2; k = 1, 2; l = 1, 2, ..., 10.$$

This model is identical to [2] except for the levels in the subscripts. The structure of the experiment is shown in Figure 3. Because there is only 1 level of the injection treatment applied to the subplots, there are now 10 cows nested in each of the pens as opposed to the 5 cows nested within each pen and injection treatment in experiment 2. The resulting ANOVA table is presented in Table 3. Because there is only 1 level of the injection treatment, there are zero df associated with *Injection* and all interaction effects of which it is a component. Thus, the terms *Injection*, *Injection* × *Cooling*, and *Injection* × *Pen* (*Cooling*) are removed from the model because they are not identifiable (i.e., 0 df). The *Pen* (*Cooling*) is still the correct error term for testing the effect of *Cooling*. The significance of *Pen* (*Cooling*) can be assessed using the residual error, which is really *Cow* (*Cooling* × *Pen*). However, the *Pen* (*Cooling*) term cannot be pooled with the residual error because it is an explicit element of the design structure. The analysis must follow the randomization. Intuitively, it should be apparent that the test for the cooling effect (the main plot factor) should not be changed because the subplots contained only 1 level of injection as opposed to 2. Randomization has not changed.

Experiment 4. This experiment is identical to experiment 3 except that the nonexistence of a subplot treatment (injection) was planned rather than being a postexperimental discovery. This experiment is a classic pen study. The randomization process is identical to that of experiment 3. Thus, whether the nonexistence of a subplot treatment is planned or not, the error term for testing the main-plot treatment (*Cooling*) remains the *Pen*

Table 3. Schematic of the ANOVA table for experiment 3

Source	df	Effect type	Error term
Cooling	1	Fixed	Pen (Cooling)
Pen (Cooling)	2	Random	Cow (Cooling × Pen × Injection)
Injection	0	Fixed	—
Injection × Cooling	0	Fixed	—
Injection × Pen (Cooling)	0	Random	—
Cow (Cooling × Pen × Injection)	36	Random	
Total	39		

(*Cooling*) term. Ignoring this term in the analysis (hence, using cows as the experimental units for the cooling treatment) violates a critical and important assumption used for the calculation of the probability of a type I error.

ISSUES WITH PEN STUDIES

Estimates of Treatment Effects

When pens are unreplicated (i.e., 1 pen per treatment), the pen and treatment effects are completely confounded, meaning that the estimates of treatment effects include any and all pen effects. When pens are replicated but not included in the model, the least squares means for treatments might or might not contain a bias depending on how the pens were assigned to the treatments. If the assignment was random and the pens are balanced (i.e., equal number of pens per treatment and equal number of cows per pen), then the least squares means for the treatments are unbiased. But if the pens are systematically unbalanced or if systematic factors are introduced during the pen assignment, whether this is done consciously or not, treatment means can be biased.

Randomization. Randomization plays a critical role in controlled experiments. As stated by Fisher (1935), randomization forms the "reasoned basis for inference" in experiments, and that the same inference would not be justified from identical data obtained in nonrandomized studies. Randomization is a concept that often has been misunderstood and, thus, deserves further clarification.

A process is random if it is without definite method or purpose; if it is unsystematic (van Belle, 2002). Randomization is important for 3 reasons (van Belle, 2002). First, randomization turns uncontrolled systematic effects into errors. Second, there is an expected balance (expected is taken in its mathematical sense) in the assignment of known and unknown factors that might influence the outcome. Perhaps more importantly, randomization provides the foundational basis for statistical procedures such as tests of significance. For example, in the following GLM

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad [4]$$

where $\mathbf{Y}$ is a vector of observations, $\mathbf{X}$ is the design matrix, β are population parameters to be estimated, and ε is a vector of errors, standard tests on the β estimates require either that the elements of ε are independently distributed, or that their dependency be accounted for in the error structure of the model through modelization of the covariance matrix of errors. In either case, randomization is essential for having accurate probability assessments (van Belle, 2002; Rubin, 2005). A simple example illustrating the critical importance of either independence or knowledge about the dependency can be easily constructed using 2 dice. The probability of getting a five on a single roll of a fair die is 1/6. If a second independent die is rolled, the probability of getting a five on both dice is simply $1/6 \times 1/6 = 1/36$. This calculation rests entirely on the assumption that the 2 dice are independent. If, however, the 2 dice are glued together, the probability of getting 2 fives is no longer 1/36, but depends on how the dice were glued. If 1 of the 4 rectangular faces made by the 2 glued dice has 2 fives, the probability of getting 2 fives on a roll is ¼. However, this probability is zero if none of the 4 faces has a double five. Thus, the structure of the dependence must be known to calculate the correct probability. It must also be noted that the probabilities under the lack of independence ($P = 0.25$ or $P = 0.0$) are both considerably different from the probability under independence ($P = 0.028$). Using the latter would be erroneous regardless of the true dependence structure.

Randomization requires considerably more than simply assigning units to treatments in a random fashion. This is a necessary condition to randomization, but it is not sufficient. All uncontrolled factors must be randomized. For example, simply assigning 20 cows to a control and 20 cows to a treatment does not ensure independent errors unless all other factors are randomized across all 40 cows. This would not be the case if the control cows were housed in one state and the treatment cows in another state, or if the control cows were all housed on the south side of the barn whereas the treatment cows were all housed on the north side, or if the control cows were all in the same pens whereas the treatment cows were all in other pens. The importance of randomization was well worded by Sir Ronald A. Fisher when he stated that "Designing an experiment is like gambling with the devil: only a random strategy can defeat all his gambling systems" (Box et al., 2005).

Inaccurate Probability of Treatment Effects

This problem is essentially one of inflated type I error, a consequence of pseudoreplication and incorrect df for the test statistic. The term pseudoreplication, first used by Hurlbert (1984), is defined as a lack of statistical independence among measurements. It results in an assignment of treatment effects with an error term inappropriate to the hypothesis being tested (van Belle, 2002). Pseudoreplication can be conceptualized intuitively. For example, one can easily understand that measuring hourly ozone levels for 24 h is not the same as measuring ozone levels 1 h per day for 24 d, or measuring 1-h ozone levels at 24 locations.

The concept of df is central to statistical inference theory, yet it is seldom defined in statistical textbooks. Standard statistical tests estimate the probability of all outcomes more extreme than that actually observed in an experiment to occur under the assumption that the null hypothesis of no treatment effect is true. Because many parameters are simultaneously estimated from a given set of data, the df represent the number of independent pieces of information available for a given estimate, or a given test. It expresses the ability of the system to wiggle in a multidimensional hyperspace. In a planar, 2-dimensional world, a 2-legged stool has no df because it can always be perfectly set on any line in the plane, regardless of the shape of the line. Likewise, a 3-legged stool has no df in a 3-dimensional world; the 3 legs can perfectly rest on any 3-dimensional surface. Expanding the analogy, counting df consists in counting the number of dimensions in the data and the number of legs in the model. Using cows as the experimental units in a pen study overestimates the number of dimensions, or the ability of the system to wiggle in the hyperspace. In such instances, the df of the error are grossly overestimated, with the consequence that the type I error is severely underestimated.

Problems of Causal Inferences

It is one thing to say that a group of cows had performance that differed from that of another group; however, it is an entirely different matter to assign a cause to that difference. The issue of causality has long been a controversial issue in statistics, leading at times to heated exchanges such as those among Fisher, Pearson, and Neyman (Rubin, 2006). Recently, causality was brought into a coherent theory for both experimental and observational studies (Cochran, 1968; Rubin, 1974; Holland, 1986; Reiter, 2000). The seminal work of Donald Rubin at Harvard University is particularly noteworthy and will be emphasized in this exposé.

The Rubin causal model (**RCM**) perspective for statistical inference for causal effects is founded on 5 primitives: unit, treatment, potential outcomes, causal effects, and fundamental problem of inference (Rubin, 2005). A unit is defined as a person, place, or thing upon which a treatment operates at a particular time. A treatment is defined as an intervention, the effects of which, on some particular measurements of the units, the investigator wishes to assess relative to no intervention (i.e., the "control"). For simplicity, our discussion will focus entirely on the simple case of a treatment vs. control situation, understanding that the theory and method do extend to cases of multiple treatments. Potential outcomes are then defined as the values of a unit's measurement after a) application of the treatment and b) nonapplication of the treatment (i.e., under control). The causal effect is then simply for each unit the comparison of the potential outcome under treatment and the potential outcome under control. This development leads to the fundamental problem of inference: we can observe at most one of the potential outcomes for each unit. Resolving this fundamental problem in a statistical sense requires a) replications, b) an assumption regarding the unit-treatment value, and c) an assignment mechanism. Replication implies that at least 1 unit receives the treatment and at least 1 unit receives the control (Rubin, 2005). The stable unit-treatment value assumption (**SUTVA**) has 2 parts: a) there is only 1 form of the treatment and 1 form of the control, and b) there is no interference among units. The assignment mechanism is simply defined as the process for deciding which units receive the treatment and which receive the control. Under SUTVA and known assignment mechanisms, the unit-level causal effect can never be observed, but it can be estimated because we have replication. The assignment mechanism determines which potential outcome we will observe for each unit. The assignment mechanism is critical even if SUTVA holds. It is essential to know or be able to infer a rule for how each unit received either the treatment or the control. A stochastic unconfounded assignment mechanism is one in which the assignment of treatment or control for all units is independent of all potential outcomes, observed and unobserved, and one in which the assignment is probabilistic. In essence, the assignment mechanism allows us to use the observations from the units assigned to the control as proxies for the unobservable potential outcomes under nonapplication of the treatment of the units assigned to the treatment. This is statistically tenable if the propensity score, defined as the probability of a unit to be assigned to treatment, is the same for the units in the treatment as well as those in the control. In a completely randomized design, randomization ensures that the propensity score is the same for all units. In a randomized block design, randomization ensures that the propensity score is the same for all units in the same block. It is the process of randomization; that is, the absence of a systematic allocation of units to treatment, that allows causal inference.

We are now in a position to understand the fundamental problem of causal inference in pen studies. We start with the simple example of 2 pens each of n cows, with the first pen assigned to the control and the second pen assigned to the treatment. The propensity score for all cows in the first pen is zero, whereas the propensity score of all cows in the second pen is 1. That is, if we know the pen, then we know the propensity score of the cows within that pen. In this situation, it is impossible for any of the cows in the second pen (the treatment),

to be matched with a cow in the first pen with the same propensity score. Thus, we have no way of estimating the potential outcome under the nonapplication of the treatment of the cows under the application of the treatment. In such instances, causal inference is impossible. All that can be said is that this set of cows out produced this other set of cows. One cannot assess the role that random chance could have played in this outcome, nor identify the cause for such difference.

We can apply the causal inference principles to a pen study with 4 pens, each of n animals. The pens are randomly assigned to the control and treatment. For illustration purpose, assume that one outcome of this randomization is to assign pens 1 and 3 to control and pens 2 and 4 to treatment. The propensity score of all cows in pens 1 and 3 is zero, whereas the propensity score of all cows in pens 2 and 4 is 1. It is impossible to match a cow in pen 2 (treatment) with a cow with the same propensity score in the control from either pen 1 or 3. Thus, we cannot make causal inference using cows as units. However, the propensity score for the 4 pens is the same, 0.5, because of the randomized assignment of pens to each of the 2 treatments. Thus, one can use either the measurements on pen 1 or pen 3 as a control match to pen 2 or pen 4. Pens are legitimate units for causal inference of the treatment. In fact, some nonparametric statistical tests are based on the computation of all pen permutations under the null hypothesis.

Optimal Pen Size

The issue of determining the optimal pen size is identical to that of determining an optimal plot size in agronomy. Plants within a plot are sampling units just like cows within a pen are also sampling units. Thus, not all cows in a pen need to be measured, although the relative cost of sampling (i.e., measuring the dependent variable on each cow) is generally so much lower than that of a cow, that the economically optimal design involves the sampling of all sampling units (i.e., all cows).

The optimum pen size is dependent, among other factors, on the size of intraexperimental and interexperimental unit competition (Federer, 1955). Intraexperimental unit competition exists when cows within the same pen affect each other either advantageously or detrimentally, something that we know happens based on animal behavior studies. This is a fundamental characteristic of cows housed in the same pen. Suppose that we have h cows per pen, and that we measure (sample) k cows per pen. In the absence of intraexperimental unit competition, the variance of the mean of the k sampling units is simply $V_p + (V_s/k)$, where V_p represents the variation among pens treated similarly, and V_s is the variance among sampling units (the cows) within the pen. But in the presence of competition between cows within pens, the variance due to this competition (V_c) must be accounted for, and the variance among experimental units then becomes (Federer, 1955)

$$V(Y) = V_p + \frac{V_s}{k} + \left(\frac{1}{k} - \frac{1}{h}\right)\mathrm{V_c} = V_p' + \frac{V_s'}{k} \qquad [5]$$

where $V_p' = V_p - \frac{V_c}{h}$ and $V_s' = V_s + V_c$.

If all the variation within the pen is due to competition, then [5] reduces to $V_p' + V_c/k$. With no competition or if all cows in the pens are sampled and the sum of the competition effects, c_i, adds to zero within pen (a weak assumption), then [5] reduces to $V_p + (V_s/k)$. Thus, as long as all cows in a pen are participating in the experiment and are being sampled, the variance due to competition is not a factor in determining the optimal pen size. In this instance, V_c is totally confounded within V_s, so the variance of cows within pen includes all of the variance due to competition. The values of V_p and V_s associated with a particular experiment are completely situation-dependent. Increased uniformity of pens decreases V_p, and increased uniformity of cows within pens decreases V_s. Prior or estimated values of these variances in combination with the cost per cow and the cost per pen can then be used to estimate an optimal pen size, using, for example, the Fairfield Smith's variance law (Federer, 1955).

From Eq. [5], it should also be apparent that, unless the variation between pens is relatively large compared with the variation between cows or if the variance due to competition is relatively large, the number of replicates to achieve a given power with pen studies can be substantially less than the number of replicates required when cows are the experimental units. The number of experimental units (pens) required for a given power is function of the variation among experimental units. In Eq. [5], V_p represents the variance due to pen after accounting for the variance due to cows forming the pens. The term "pen" implies a grouping of cows but does not necessarily imply a conventional pen found on farms. In fact, "pens" could actually be "farms". That is, one could use farms as experimental units and randomly assign them to the various treatments. In this instance, V_p would likely be large, and a relatively large number of replicates would be required for a given power. Likewise, pens on commercial farms may not be very uniform (i.e., animals are penned according to parity, production levels, pregnancy status, health status, etc.), resulting in large variance between pens. In such instances, a greater number of replications than with uniform pens are required, unless some form of switchback design is used, as explained in the next section.

Table 4. Schematic ANOVA table for a completely randomized design with pen as the experimental unit

Source	df[1]	Effect type	Error term
Treatment	i – 1	Fixed	Pen (Treatment)
Pen (Treatment)	(j – 1)i	Random	Cow (Pen × Treatment)
Cow (Pen × Treatment)	(k – 1)ij	Random	—
Total	(ijk – 1)		

[1]i = number of treatments, j = number of pens per treatment, k = numbers of cows per pen.

STATISTICAL DESIGNS FOR PEN STUDIES

All statistical designs using cows as the experimental units can also be used with pens. Fundamentally, a pen study has an implied split-plot design with only one treatment level in the subplot, a design more commonly known as a nested design. Therefore, any design can be applied to the main plots (the pens). The design selection must consider the research hypothesis, the desired power, as well as the inevitable assumptions associated with each design. The following description of a few designs is not meant to be exhaustive, but rather to serve as examples of how the statistical model and the resulting analysis are implemented with pen studies while accounting for the cow effect in the subplot. Littell et al. (2006) can be consulted for additional designs and explanations.

Completely Randomized Design

This is the simplest and least powerful design, but it also has the fewest assumptions. Cows are randomized across pens, and pens are randomized across treatments. The completely randomized design (**CRD**) requires a minimum of 1 pen per treatment, plus 1 additional pen for 1 of the treatments, a situation that results in very low power because the error has only 1 df. The number of pens required for a desired power is dependent on the significance level of the tests (i.e., the desired size of type I error), the size of the expected treatment differences, and the variance among pens. With this design, the researcher should form pens that are as much alike as possible. The variation among cows within each pen is not very relevant to the analysis unless one suspects an interaction between the initial level of production and the response to treatments. Initial (covariate) measurements on the cows as well as repeated measures can easily be incorporated in the model, as explained later. For now, we shall only consider the case of a single measurement (or a mean of measurements) for each cow.

The statistical model underlying the analysis is

$$y_{ijk} = \mu + \alpha_i + p_{j:i} + \varepsilon_{ijk} \quad [6]$$

where y_{ijk} are the observed values, μ is the overall population mean, α_i is the effect of the ith treatment, $p_{j:i}$ is the random effect of the jth pen within the ith treatment, and ε_{ijk} is the random error, assumed iid N(0, σ^2_ε).

In [6], the ε_{ijk} is actually the effect of cow nested within treatment × pen. The schematic ANOVA table for this model is shown in Table 4. A data record consists of 1 observation for each cow used in the experiment. This model is easily fitted with the MIXED procedure of SAS (SAS Institute, 2004) using the following statements:

```
PROC MIXED;
CLASSES treat pen;
MODEL Y = treat;                    [7]
RANDOM pen(treat);
LSMEANS treat;
RUN;
```

RCBD

Frequently, the number of pens on a given farm is insufficient to reach a satisfactory power using a CRD. An easy solution in such instances is to use a multisite design (i.e., conduct the experiment across many farms) in what is known as a RCBD. In such case, farms act as blocking factors whose effects can be considered either fixed or random, depending on the process used for selecting the farms, the inference range, and the number of farms used in the experiment. Theoretical considerations and guidelines for selecting the type of effects for blocks are found in McCulloch and Searle (2001). If the effect of farm is considered fixed, the statistical model that underlies the analysis is

$$y_{ijk} = \mu + \lambda_i + \alpha_j + \lambda\alpha_{ij} + p_{k:ij} + \varepsilon_{ijkl} \quad [8]$$

where y_{ijk} are the observed values, λ_i is the fixed effect of the ith farm, α_j is the fixed effect of the jth treatment, $\lambda\alpha_{ij}$ is the interaction effect between the ith farm and the jth treatment, $p_{k:ij}$ is the random effect of the kth pen within the ith farm and jth treatment, assumed iid N(0, σ^2_p), and ε_{ijkl} is the random error, assumed iid N(0, σ^2_ε).

In [8], the ε_{ijkl} is actually the effect of cow nested within farm, treatment, and pen. The schematic ANOVA for this model is shown in Table 5. In this instance, the *Pen* (*Farm* × *Treatment*) is the correct error term for testing

Table 5. Schematic ANOVA table for a randomized block design with pen as the experimental unit

		Fixed blocks		Random blocks	
Source	df[1]	Effect type	Error term	Effect type	Error Term
Farm	i – 1	Fixed	Pen (Farm × Treatment)	Random	Farm × Treatment
Treatment	j – 1	Fixed	Pen (Farm × Treatment)	Fixed	Farm × Treatment
Farm × Treatment	(i – 1)(j – 1)	Fixed	Pen (Farm × Treatment)	Random	Pen (Farm × Treatment)
Pen (Farm × Treatment)	(k – 1)ij	Random	Cow (Pen × Farm × Treatment)	Random	Cow (Pen × Farm × Treatment)
Cow (Pen × Farm × Treatment)	(l – 1)ijk	Random	—	Random	—

[1]i = number of farms; j = number of treatments; k = number of pen per treatment, and per farm; l = number of cows per pen.

the treatment effect. Model [8] can be fitted using the MIXED procedure of SAS with the following statements:

```
PROC MIXED;
CLASSES farm treat pen;
MODEL Y = farm treat farm*treat;                    [9]
RANDOM pen(farm treat);
LSMEANS farm treat farm*treat;
RUN;
```

If the effect of farm is considered random, the model that underlies the analysis is

$$y_{ijk} = \mu + f_i + \alpha_j + f\alpha_{ij} + p_{k:ij} + \varepsilon_{ijkl} \quad [10]$$

where y_{ijk} are the observed values, f_i is the random effect of the ith farm, iid N(0, σ^2_f), α_j is the fixed effect of the jth treatment, $f\alpha_{ij}$ is the random interaction effect between the ith farm and the jth treatment, iid N(0, $\sigma^2_{f\alpha}$), $p_{k:ij}$ is the random effect of the kth pen within the ith farm and jth treatment, iid N(0, σ^2_p), and ε_{ijkl} is the random error, assumed iid N(0, σ^2_ε).

Model [10] is solved using the following statements with PROC MIXED of SAS:

```
PROC MIXED;
CLASSES farm treat pen;
MODEL Y = treat;                                    [11]
RANDOM farm farm*treat pen(farm treat);
RUN;
```

In [11] the proper error term for testing the effect of treatment is *Farm* × *Treatment* (Table 5). In this design as well as in the previous one, it is in the researcher's interest to minimize the variance across pens. Thus, one would want to have pens within farms as uniform as possible to maximize the power of the test.

Switchback Designs

In instances where it is difficult to assemble a large number of uniform pens, designs in which the pen acts as its own control, such as Latin squares and switchback designs, can prove to be considerably more powerful than the first 2 classes of designs previously outlined. This gain in power is achieved through additional assumptions, such as the implicit pooling of interaction terms with the error (Cochran and Cox, 1957), which can be a possible drawback.

Figure 4 shows an example of an experiment involving 10 pens, 2 treatments, and 3 periods in a switchback design. Notice that the order in which the 2 treatments are applied can follow 2 distinctive sequences. The statistical model underlying the analysis is

$$y_{ijklm} = \mu + \lambda_i + p_{j:i} + \pi_k + \alpha_l + \lambda\alpha_{il} + \pi\alpha_{kl} + p\pi\alpha_{jkl:i} + c_{m:ij} + \varepsilon_{ijklm} \quad [12]$$

where y_{ijklm} are the observed values, λ_i is the fixed effect of the ith sequence, $p_{j:i}$ is the random effect of the jth pen with the ith sequence, iid N(0, σ^2_p), π_k is the fixed effect of the kth period, α_l is the fixed effect of the lth treatment, $\lambda\alpha_{il}$ is the fixed interaction effect of the ith sequence with the lth treatment, $\pi\alpha_{kl}$ is the fixed interaction effect of the kth period with the lth treatment, $p\pi\alpha_{jkl:i}$ is the random interaction effect between the jth pen with the kth period and the lth treatment within the ith sequence, iid N(0, $\sigma^2_{p\pi\alpha}$), and ε_{ijklm} is the random error, assumed iid N(0, σ^2_ε).

In [12], the ε_{ijklm} are in fact the *Period* × *Cow* (*Pen Sequence*). This model would be solved using PROC MIXED of SAS with the following statements:

```
PROC MIXED;
CLASSES sequence pen period treat cow;
MODEL Y = sequence period treat
  sequence*treat period*treat;                      [13]
RANDOM pen(sequence) period*treat*pen(sequence)
  cow(pen sequence);
RUN;
```

In [13], the error term used for testing the effect of *treatment* is *period***treat***pen*(*sequence*) and not the *pen*(*sequence*). Consequently, the experimental unit for

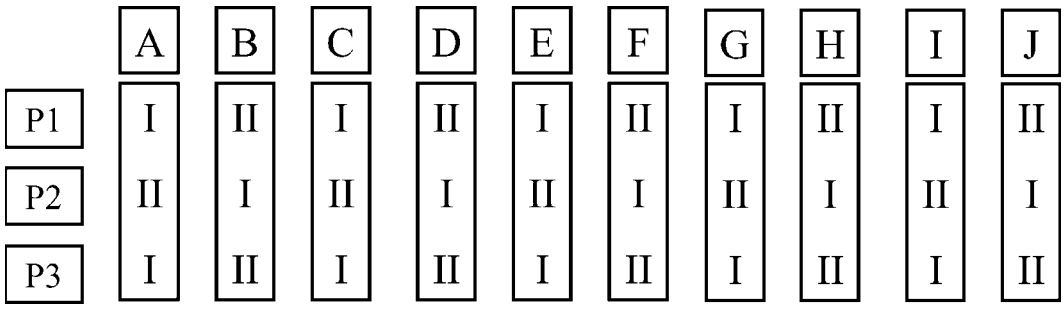

Figure 4. Schematic of the research design for a switchback design with 10 pens (A to J), 2 treatments (I and II), and 3 periods (P1, P2, P3).

the treatment is no longer simply a pen, but a pen-period. Each pen serves as its own control. This is important because this implies that having uniformity across pens is not an attribute important for testing the effect of treatments, a feature that was important in the CRD and RCBD.

Covariate and Repeated Measures

The designs and their associated SAS statements that were previously described were based on data records for individual cows. If pens are approximately all of the same size, and if there is no change in animal numbers during the study (i.e., no cow leaves or is brought into an experimental pen), the observations could be first averaged within pens and these averages used as records for statistical analyses. In which case, the pen is explicitly the experimental unit. In general, however, it is preferable to use individual cow records because of the ease of incorporation of covariates in the analysis, and the correct use of partial records. Likewise, repeated measurements on the cows are easily incorporated in the analysis.

To understand how covariates and repeated measures are incorporated in the statistical analysis of pen studies, we reuse the example from our virtual experiment 2, but this time we make use of production measurements done immediately before the assignment of cows to pens (i.e., a covariate measurement). In addition, we are now told that production was measured weekly for 4 wk. The model used for the statistical analysis is

$$\begin{aligned} y_{ijkl} = \mu + \delta_k + p_{j:k} + \alpha_i + \delta\alpha_{ik} + \alpha p_{j:ik} + \beta X_{ijkl} \\ + c_{l:ijk} + \omega_m + \omega\delta_{mk} + \omega\alpha_{mi} + \omega\delta\alpha_{mki} \\ + \omega p_{mj:k} + \omega\alpha p_{mij:k} + \varepsilon_{ijklm}; \end{aligned} \quad [14]$$

$$i = 1, 2;\ j = 1, 2;\ k = 1, 2;\ l = 1, 2, ..., 5;\ m = 1, 2, 3, 4;$$

where δ_k denotes the fixed effect of the kth cooling treatment; $p_{j:k}$ is the random effect associated with the jth pen nested within the kth cooling treatment, assumed iid $N(0, \sigma^2_p)$; α_i denotes the fixed effect of the ith injection treatment; $\delta\alpha_{ik}$ is the fixed effect of the interaction between the kth cooling treatment and the ith injection treatment; $\alpha p_{j:ik}$ is the random interaction effect associated with the jth pen within the kth cooling treatment and the ith injection treatment, assumed iid $N(0, \sigma^2_{ap})$; βX_{ijkl} is the covariate adjustment for each cow; $c_{l:ijk}$ is the random effect of the lth cow within the kth cooling treatment, ith injection treatment, and jth pen, assumed iid $N(0, \sigma^2_c)$; ω_m is the fixed effect of the mth week; $\omega\delta_{mk}$, $\omega\alpha_{mi}$, and $\omega\delta\alpha_{mki}$ are the fixed effects of interaction terms; $\omega p_{mj:k}$ is the random interaction effect associated with the mth week and the jth pen within the kth cooling treatment, assumed iid $N(0, \sigma^2_{p\omega})$; $\omega\alpha p_{mij:k}$ is the random interaction effect associated with the mth week, the ith injection, and the jth pen within the kth cooling treatment, assumed iid $N(0, \sigma^2_{p\omega})$; and ε_{ijklm} denotes the random error, which are no longer assumed independent because of the repeated measurements, but are assumed identically distributed $N(0, \Sigma)$.

In essence, model [14] is a split-split plot in time, with the second split taken loosely because unlike a true split-plot where the subplots are fully randomized, the subplots of a repeated measures design cannot be randomized (i.e., wk 3 always follows wk 2). The ANOVA table and the corresponding df are shown in Table 6, along with the 3 levels in the design structure (plots) and their associated experimental units (error terms).

Remembering that a pen study is in fact a split-plot with only one level of the subplot treatment, it is very easy to identify the factors to be included in a pen study with covariate and repeated measurements (Table 7). Because of the absence of subplot treatments, model [14] simplifies to

$$\begin{aligned} y_{ijkl} = \mu + \delta_k + p_{j:k} + \beta X_{jkl} + c_{l:jk} + \omega_m \\ + \omega\delta_{mk} + \omega p_{mj:k} + \varepsilon_{jklm}; \end{aligned} \quad [15]$$

$$j = 1, 2;\ k = 1, 2;\ l = 1, 2, ..., 5;\ m = 1, 2, 3, 4;$$

Table 6. Schematic ANOVA table for a split-plot design (experiment 2) with covariate and repeated measurements on the subplots (cows)

Design group	Source	df	Effect type	Error term
Main plot	Cooling	1	Fixed	Pen (Cooling)
	Pen (Cooling)	2	Random	
Subplot	Injection	1	Fixed	Injection × Pen (Cooling)
	Injection × Cooling	1	Fixed	Injection × Pen (Cooling)
	Injection × Pen (Cooling)	2	Random	
	Covariate	1	Fixed	Cow (Cooling × Pen × Injection)
	Cow (Cooling × Pen)	31	Random	
Sub-sub-plot	Week	3	Fixed	Week × Pen (Cooling)
	Week × Cooling	3	Fixed	Week × Pen (Cooling)
	Week × Injection	3	Fixed	Week × Pen (Cooling)
	Week × Cooling × Injection	3	Fixed	Week × Injection × Pen (Cooling)
	Week × Pen (Cooling)	6	Random	
	Week × Injection × Pen (Cooling)	6	Random	
	Cow (Cooling × Pen × Week)	96	Random	
	Total	159		

where all the terms are defined as in [14] with obvious changes in the subscripts. This model can be fitted using the following SAS statements:

```
PROC MIXED;
CLASSES cooling pen cow week;
MODEL Y = cooling preY week week*cooling;
RANDOM pen(cooling) cow(cooling*pen)                [16]
  week*pen(cooling);
REPEATED week / TYPE=AR(1) SUBJECT =
  cow(cooling*pen);
RUN;
```

Note that the term *cow*(*cooling*∗*pen*) appears in both the RANDOM and the REPEATED statements. In fact, this is dependent on the type of covariance structure chosen. In some structures (e.g., unstructured, UN, or compound symmetry, CS) the 2 are completely redundant, leading to a failure of PROC MIXED to identify a solution. This is not a problem per se, as it reflects a set of statements that, in essence, over-parameterize the model (Littell et al., 2006). Although it is theoretically possible to have the term *cow*(*cooling*∗*pen*) in both the RANDOM and REPEATED statements with an autoregressive structure [AR(1)], the algorithm frequently experiences convergence problems, and the term must be removed from the RANDOM statement without any changes to the test of the fixed effects of interest, but a different interpretation to the estimate of the *cow*(*cooling*∗*pen*) component of variance. Additionally, it should be noted that in all sets of SAS statements presented in this article, we did not include any option to correct the df to account for the uncertainty in estimating the **G** and **R** matrices of the mixed models equations (SAS Institute, 2004). Corrections to the df are generally very small for data sets that are relatively well balanced. In cases of markedly unbalanced data (i.e., different number of cows per pen or differing number of pens per treatment), it is generally preferable to compute inflation factors along with Satterthwaite-based degrees of free-

Table 7. Schematic ANOVA table for a pen study with covariate and repeated measurements on the subplots (cows)

Design group	Source	df	Effect type	Error term
Main plot	Cooling	1	Fixed	Pen (Cooling)
	Pen (Cooling)	2	Random	
Subplot	Injection	0		
	Injection × Cooling	0		
	Injection × Pen (Cooling)	0		
	Covariate	1	Fixed	Cow (Cooling × Pen)
	Cow (Cooling × Pen)	35	Random	
Sub-sub-plot	Week	3	Fixed	Week × Pen (Cooling)
	Week × Cooling	3	Fixed	Week × Pen (Cooling)
	Week × Injection	0		
	Week × Cooling × Injection	0		
	Week × Pen (Cooling)	6	Random	Cow (Cooling × Pen × Week)
	Week × Injection × Pen (Cooling)	0		
	Cow (Cooling × Pen × Week)	108	Random	
	Total	159		

dom (SAS Institute, 2004). In such instances, the MODEL statement in [15] would include the following option:

```
MODEL Y = cooling preY week week*cooling /          [17]
DDFM=KENWARDROGER;
```

The advantages of using individual records for each cow in the analysis should now be apparent. As long as cows do not leave the experiment for reasons related to the treatments, removing or moving a cow into an experimental pen during the conduct of the experiment only leads to imbalance in the data (i.e., missing observations on some cows), without any consequence on the tests of interest. Additionally, covariance corrections can be applied directly at the cow level. Other correction factors, such as parity, can be introduced in the subplot as long as cows of different parities were not completely separated into different pens. These corrections to the cow records lead to reduced between-pen variance, something that is desirable with longitudinal design in which the pen does not serve as its own control.

CONCLUSIONS

Ignoring animal grouping during data analysis of pen studies can result in biased estimates of treatment effects under some condition of imbalance, biased tests of significance for the treatment effects, and produces an incoherent causal inference framework. These attributes are highly inconsistent with the scientific process, where absence of bias and coherence of methods are sought. Thus, pen studies that are improperly designed or analyzed should not be misrepresented as scientific studies, and our scientific journals should refrain from publishing such studies. Properly designed and analyzed, pen studies can yield production and management information that may be more applicable to commercial dairy operations than the traditional studies where cows are individually housed and fed. The vast array of experimental designs that have served so well research conducted on individual cows can be successfully used with pen studies.

ACKNOWLEDGMENTS

The author thanks Jeff Firkins and Joanne Knapp for helpful comments and suggestions on an earlier version of this manuscript.

REFERENCES

Box, G. E. P., J. S. Hunter, and W. G. Hunter. 2005. Statistics for Experimenters. 2nd ed. John Wiley & Sons Inc., New York, NY.
Cochran, W. G. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24:295–313.
Cochran, W. G., and G. M. Cox. 1957. Experimental Designs. Wiley, New York, NY.
Damon, R. A., and W. R. Harvey. 1987. Experimental Design, ANOVA, and Regression. Harper & Row, Cambridge, UK.
Federer, W. T. 1955. Experimental Design: Theory and Application. The Macmillan Company, New York, NY.
Fisher, R. A. 1935. Design of Experiments. Oliver & Boyd, Edinburgh, UK.
Gill, J. L. 1987. Biased statistical analysis when the animal is not the experimental unit. J. Am. Vet. Med. Assoc. 190:5–6.
Gill, J. L. 1989. Statistical aspects of design and analysis of experiments with animals in pens. J. Anim. Breed. Genet. 106:321–334.
Holland, P. W. 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81:945–960.
Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:198–211.
Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. 2006. SAS System for Mixed Models. 2nd ed. SAS Institute, Inc., Cary, NC.
McCulloch, C. E., and S. R. Searle. 2001. Generalized, Linear, and Mixed Models. John Wiley & Sons Inc., New York, NY.
Milliken, G. A., and D. A. Johnson. 1992. Analysis of Messy Data. Volume I: Designed Experiments. Chapman & Hall/CRC, New York, NY.
Milliken, G. A., and D. A. Johnson. 2002. Analyses of Messy Data. Volume III: Analysis of Covariance. Chapman & Hall/CRC, New York, NY.
Reiter, J. 2000. Using statistics to determine causal relationships. Am. Math. Monogr. 107:24–32.
Rubin, D. B. 1974. Estimating causal effect of treatments in randomized and nonrandomized studies. J. Educ. Psych. 66:688–701.
Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modelling, decisions. J. Am. Stat. Assoc. 100:322–331.
Rubin, D. B. 2006. Matched Sampling for Causal Effects. Cambridge Univ. Press, New York, NY.
SAS Institute. 2004. SAS/STAT 9.1 User's Guide, Vol. 4. The Mixed Procedure. SAS Inst. Inc., Cary, NC.
St-Pierre, N. R., and L. R. Jones. 1999. Interpretation and design of nonregulatory on-farm feeding trials. J. Dairy Sci. 82(Suppl. 2):177–182.
USDA. 2006. Farms, Land in Farms, and Livestock Operations Report. http://usda.mannlib.cornell.edu/usda/nass/FarmLandIn/2000s/2006/FarmLandIn-01-31-2006.pdf Accessed Jul. 7, 2006.
van Belle, G. 2002. Statistical Rules of Thumb. John Wiley & Sons Inc., New York, NY.
Yandell, B. S. 1997. Practical Data Analysis for Designed Experiments. Chapman & Hall, New York, NY.